Submitted to *INFORMS Journal on Optimization*
manuscript (Please, provide the manuscript number!)

Novel Target Discovery of Existing Therapies: Path to Personalized Cancer Therapy

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, dbertsim@mit.edu

Ying Daisy Zhuo
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, zhuo@mit.edu

Discovering new drugs involves tremendous effort and financial resources, often at a significant risk of failed trials. Identifying new targets of existing drugs provides a promising direction, especially for molecular targeted cancer therapies. This paper presents a novel, machine learning and optimization-based method that identifies potential targets of existing drugs to expand the treatable patient population. The method has the following advantages: 1) it is based on clinical and genomic data from a large national cancer hospital; 2) it incorporates state-of-the-art knowledge of cancer molecular biology and signaling pathways; 3) it models patient heterogeneity explicitly outside genomics. The output is an ordered list of therapy-target pairs that our algorithm identifies as highly promising to be further tested. The results are highly accurate when validated against known mechanisms of action for existing drugs, where relationships such as pertuzumab-ERBB2, cetuximab-EGFR, and erlotinib-EGFR were independently identified. We found similar results in the external The Cancer Genome Atlas data set. The findings suggest that a data-driven optimization approach to precision cancer medicine may lead to breakthroughs in the drug discovery process and recommend effective personalized cancer treatments given patient-specific genomic and phenotypic information.

Key words: mixed integer optimization, targeted therapy, cancer genomics

1. Introduction

Drug discovery is a costly and slow process. The average cost of developing a new pharmaceutical drug has risen to 2.5 billion in the United States (DiMasi et al. 2016), and the average time from discovery to market is estimated to be 13.5 years (Paul et al. 2010). The high cost does not necessarily translate to the clinical efficacy of drug candidates identified; in fact, only 11% of drugs in clinical trials are eventually approved.

This increasing burden, risk, and low return-on-investment from *de novo* drug discovery have forced drug developers to look at alternatives. Repurposing existing drugs for novel uses is an attractive option to improve the research and development productivity, so long as the drugs turn out to be clinically effective on the novel uses identified. As the toxicity profile and pharmacokinetics of existing drugs has been extensively tested in prior trials, effective drug repurposing can significantly reduce the cost and regulatory approval time frame.

Cancer therapeutic development presents a particular opportunity for alternatives to *de novo* drug discovery. The advances in understanding of cancer biology at the molecular level and the wide availability of genetic biomarker testing expanded cancer drugs from traditional cytotoxic chemotherapies to more focused targeted therapies in the past decade (Hanahan and Weinberg 2011). These drugs interfere with specific molecular targets that are involved in the growth, progression, and spread of cancer, and have demonstrated promising efficacy and tolerability for cancer patients in clinical settings.

Yet as cancer therapies are becoming more precision-based and individualized, the size of patient cohorts with the specific molecular biomarker for a new drug continues to decline. In 2016, more than 40% of newly approved drugs have the orphan designation (indicated for fewer than 200,000 Americans), among which half are cancer drugs (U.S. Food and Drug Administration 2016). Between 2009 and 2015, FDA approved 12 orphan cancer drugs indicated for biomarker-defined subsets (Kesselheim et al. 2017). If additional targets are identified for these highly specialized drugs, the increased market size can justify the high research and development cost.

Drug repurposing in cancer has seen some success stories. A notable example is crizotinib, first developed for anaplastic large-cell lymphoma as a MET inhibitor. Its ALK inhibiting properties was later discovered and therefore repositioned to treat a subpopulation of non-small cell lung cancer (NSCLC). The approval process for NSCLC took only 4 years, significantly shorter than the average time (Shaw et al. 2011, Li and Jones 2012). Imatinib, originally approved in 2001 for chronic myelogenous leukemia (CML), now has many indications including for patients with KIT positive gastrointstinal stormal tumors (GIST) (Novartis 2017, Druker 2004). Many such hidden interactions remain to be discovered; it is estimated from known databases that on average a targeted cancer drug interacts with six molecular targets (Mestres et al. 2009).

Identifying such novel drug-target relationship accurately is, unfortunately, a difficult problem. The current approach to find new targets of existing drugs in practice typically relies on pure serendipity or anecdotal evidence. The combinatorial nature of the drug-target network prohibits the conventional trial-and-error process to be effective, as the number of possibilities explodes with respect to number of potential targets. With next-generation genome sequencing becoming more widely available, additional potential targets are identified, creating further opportunities as well as challenges for drug repurposing.

As a result, systematic, statistical, and computational approaches are needed for such discovery. In the past decade, progress has been made leveraging various data sources to suggest new targets and drug indications (Chen and Butte 2016, Li et al. 2015). In one strategy, the relationship is inferred based on the structural and chemical similarity between drugs (Li and Lu 2012), and/or based on existing indications and documented side effects (Dudley et al. 2011, Chiang and Butte 2009, Campillos et al. 2008). These studies incorporate the rich information from expert curated biological knowledge databases, which represent the current understanding but may often be incomplete and inaccurate. In addition, the validation for these studies are typically based on internal data only. As a result, most of the identified drug repositioning opportunities have not been successfully translated into clinical practice.

Recent advances in genomic technology opened up opportunities for large scale association studies on drugs and biomarkers. Garrett et al. screened several hundred cancer cell lines under 130 drugs, identified some mutated genes that were associated with sensitivity to certain drugs (Garnett et al. 2012). The Connectivity Map (Lamb et al. 2006) offers an excellent data source on the effect of compounds on the gene expression level of cancer cell lines, and has been used by researchers for drug repositioning (Dudley et al. 2011, Zerbini et al. 2014). However, the data are from a few isolated cell lines *in vitro*, which may not reflect the true effect *in vivo*. Consequently, most studies typically focus on the gene signature changes rather than long-term clinical outcomes. Further, the effects of phenotypic information are often ignored in these studies (Spainhour and Qiu 2016, Hanahan and Weinberg 2011, Rubio-Perez et al. 2015).

The end goal for the drug repositioning research would be to inform the selection of personalized drug combinations given a patient's genomic and phenotypic data, completing the bench-to-bedside translation. In clinical settings, treatment selection is still based on one or two signatures. To extend to multiple targets and combination treatments, some network-based approaches were taken to infer combinatorial drug effects (Tang et al. 2013). Alternatively, gene expression-based methods were also considered - for instance, Pritchard et al. used a signature-based approach and identified cytotoxic chemotherapy combination mechanisms (Pritchard et al. 2013); Zhao et al. further incorporated drug efficacy and side effects to derive optimal combinations (Zhao et al. 2014). On the phenotypic side, some recent work by Bertsimas et al. built a large database of clinical trials and used machine learning/optimization to design combination chemotherapy regimens for cancer (Bertsimas et al. 2016). It is an important step towards personalized cancer therapy, although the nature of the aggregate data prevents it from fully capturing individual heterogeneity.

This study offers a data-driven approach to repurpose existing drugs by identifying novel drug-target relations. By integrating clinical and genomic data from thousands of patients at a large national cancer hospital with knowledge from biological processes, we are able to infer the interactions between drugs and targets from the binary coefficients of a non-linear regression analysis,

learned via mixed-integer optimization. Our model outputs a priority list of candidate targets for each drug to be further investigated. The inferred relations are validated against some of the known mechanisms of certain drugs.

Without randomized clinical trials, it is impossible to establish causal relationships from observational data only. We make a number of assumptions for our inferred drug and target interactions to be valid. We assumed a Bayesian Network model structure with hidden variables to be the underlying ground-truth causal model. The model is assumed to be have enough capacity to represent the relationships (linear effects of the demographic and medical information, for example). We also assume that the data is sufficiently large for the inference. Under these assumptions, the mixed-integer optimization approach can then correctly recover with high precision the ground-truth drug-target relationships. We show this in a synthetic experiment where the ground-truth is known.

Because one cannot verify the validity of the assumptions given the observational nature of the data, the causal effects we inferred will still need to be confirmed by direct randomized experimentation. As a proxy, we test whether our assumed network model represents ground-truth causal model closer than simple ones by evaluating the generalization performance against simpler linear models in a hold-out test set and an external data set using the Cancer Genome Atlas.

The findings provide a natural pathway towards personalized, mechanism-guided cancer therapy. Quantifying the impact of each agent in the signaling network allows for more precise treatment selection that target the relevant pathways therapeutically, preventing the relapses often seen in targeted therapies. Given a patient with demographic, medical, and genomic information, the model allows practitioners to select combination therapies that holistically maximizes the long-term effectiveness by targeting the right processes while maintaining low level of off-target toxicity.

2. Data and Methods

2.1. Data

We retrospectively obtained the electronic health records (EHR) data for patients at the Dana-Farber/Brigham and Women's Cancer Center (DFCI) from 2004-2014. There are

- $i = 1, \ldots, n$ patients,
- $j = 1, \ldots, m$ treatments T_j,
- $k = 1, \ldots, K$ gene mutations G_k,
- $\ell = 1, \ldots, L$ sets S_ℓ that consists of genes (please see the significance in Section 2.2).

For each patient i we have the following information:

$$G_{ik} = \begin{cases} 1, \text{ if gene mutation } G_k \text{ is present in patient } i, \\ 0, \text{ otherwise,} \end{cases}$$

and

$$T_{ij} = \begin{cases} 1, \text{ if patient } i \text{ received treatment } T_j, \\ 0, \text{ otherwise.} \end{cases}$$

In addition, $\mathbf{x}_i$ is the D-dimensional vector of phenotype covariates for patient i, including demographics, cancer diagnoses, comorbidities, prior treatments, resource utilization, and vital signs/laboratory tests results. y_i is the observed survival time in days for patient i from initiation of anti-cancer regimen, including chemotherapy, immunotherapy, and targeted therapy.

To be eligible for the study, patients must have initiated at least one anti-cancer regimen at the cancer center. Furthermore, patients are required to have measurements from *Onco-Panel/OncoMap*, a sequencing platform that detects genetic mutations and other cancer-related DNA alterations.

We further obtained auxiliary data from curated biological and pharmacological databases. The sets $S_\ell,\ \ell = 1, \ldots, L$ are obtained from Molecular Signatures Database, a curated database representing canonical pathways (Subramanian et al. 2005, Kanehisa and Goto 2000), which is widely used as a reference knowledge base for the interpretation and analysis of datasets generated from genome sequencing.

Missing values were imputed using unconditional mean; the results based on other imputation algorithms such as knn-impute (Troyanskaya et al. 2001) and optimal-impute (Bertsimas et al. 2017) are included in sensitivity analyses. We used regimens initiated from 2004-11 as the training set, and used regimens initiated from 2012-14 as the validation set. The institutional review boards of the associated institutions (anonymized here) approved this study.

2.2. Model

It is widely believed that mutations in tumor suppressor genes or oncogenes are linked to the development of cancer. Such genes are involved in critical signaling pathways that regulate cell cycles. When a mutation is present, the protein expression for that gene can be abnormal and adversely affect cell functions controlled by the pathway, subsequently affecting patient survival. In this section, we build the mathematical model of cancer treatments based on this biological mechanism, the structure of which are based on well-known literature in cancer including Hanahan and Weinberg (2011), Weinberg (2013), Sawyers (2004).

Targeted therapies, as an effective treatment of cancer, are known to interfere with the action of some proteins expressed by the mutated genes (G_k) by inhibiting their functions. However, it is unknown whether a treatment T_j targets a specific mutation G_k. Correspondingly, our key decision variable is to infer from data:

$$w_{jk} = \begin{cases} 1, \text{ if treatment } T_j \text{ targets gene mutation } G_k, \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

We postulate that time of survival y_i of patient i is affected by the vector of phenotype covariates $\mathbf{X}_i$, and specifically decreases when there are abnormal gene expressions that are not targeted by appropriate therapies. In particular, groups of gene expressions S_ℓ function together to affect survival. Mathematically we assume the following form (where we denote observed data y_i, $\mathbf{x}_i$ with lower-case letters, but unobserved variables with capital letters Y_i, $\mathbf{X}_i$):

$$Y_i = \exp(\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} - \sum_{\ell=1}^{L} r_\ell s_{i\ell}), \tag{2}$$

where β_0 is the bias term, $\boldsymbol{\beta}$ is a D-dimensional vector to be estimated that models the influence of the corresponding phenotype $\mathbf{X}_i$, the variable r_ℓ to be estimated is the influence of having an abnormal set S_ℓ, and the decision variable $s_{i\ell}$ is defined as follows:

$$s_{i\ell} = \begin{cases} 1, \text{ if some gene expression for patient } i \text{ in the set } S_\ell \text{ is abnormal,} \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

Table 1 Data and variable descriptions for the network model

Data/Variable	Definition
T_{ij}	Patient i received treatment T_j
G_{ik}	Gene mutation G_k is present in patient i
$\mathbf{x}_i$	Phenotype covariates of patient i
s_ℓ	Gene set ℓ
y_i	Survival in days for patient i
w_{jk}	Treatment T_j targets gene mutation G_k
g_{ik}	Gene mutation G_k is present in patient i but not targeted
$\boldsymbol{\beta}$	Linear effect of covariate $\mathbf{x}$
β_0	Bias term
s_{il}	Whether some gene expression for patient i in set S_ℓ is abnormal
r_ℓ	Effect of having an abnormal set S_ℓ

We obtain from Eq. (2) that

$$\log(Y_i) = \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta} - \sum_{\ell=1}^{L} r_\ell s_{i\ell}, \tag{4}$$

i.e., presence of an abnormal set S_ℓ of gene expressions decreases life span by r_ℓ log days.

In summary, therapies target gene expressions. A gene expression is abnormal if the corresponding mutation is present and a blocking therapy is not applied to the patient. If a gene expression is abnormal, then all sets S_ℓ that contain it are also abnormal, and thus survival time is negatively affected from Eq. (4). Our objective is to infer the principal variable w_{jk}, defined in Eq. (1), the vector $\boldsymbol{\beta}$, and the coefficients r_ℓ. In the process of inferring the above principal variables, we also infer auxiliary variables $s_{i\ell}$ defined in Eq. (3), gene expression variable

$$g_{ik} = \begin{cases} 1, \text{ if gene mutation } G_k \text{ is present in patient } i \text{ but not targeted,} \\ 0, \text{ otherwise;} \end{cases} \tag{5}$$

as well as the product between r_ℓ and $s_{i\ell}$, denoted by:

$$v_{i\ell} = r_\ell s_{i\ell} = \begin{cases} r_\ell, \text{ if } s_{i\ell} = 1, \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

Figure 1 depicts the inference problem we are solving, with the data and variables summarized in Table 1.

Because of the logical relations among the variables, the model can be naturally formulated using Mixed Integer Optimization (MIO) (Bertsimas and Weismantel 2005). MIO is a particularly flexible

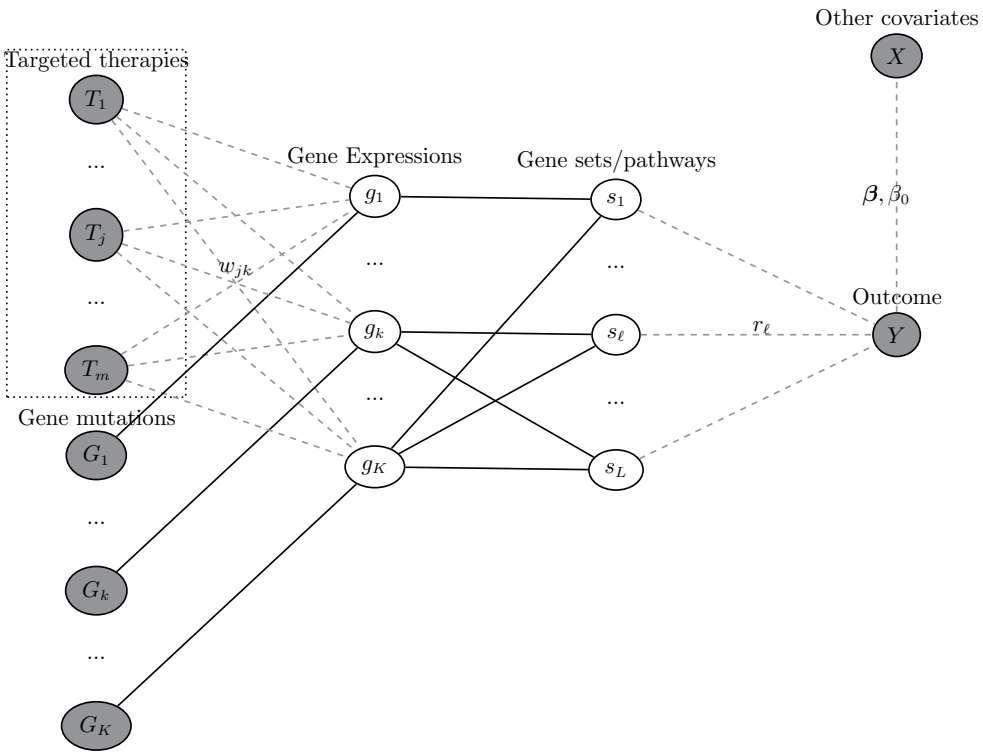

Figure 1 Model network incorporating the treatments, gene mutations, gene expressions, signaling pathway gene sets, phenotype covariates, and the survival outcomes. The dashed edges w_{jk} and the coefficients r_ℓ and $\boldsymbol{\beta}$ need to be determined by the mixed integer optimization problem. The solid edges and shaded variables are known and represent inputs to our model. The unshaded variables are derived.

tool for mathematical modeling, especially for incorporating logical constraints and relationships among variables and data. In recent years, advances in algorithms and computational hardware allowed a speedup of over several trillion times for solving MIO models, making even complex, large-scale problems tractable. We therefore selected it as our modeling technique.

The full MIO model formulation is as follows, with $i \in [n]$ et al. as shorthand notation for the enumeration $i \in 1, \ldots, n$, et al.:

$$\min_{\mathbf{w},\mathbf{g},\boldsymbol{\beta},\beta_0,\mathbf{s},\mathbf{r}} \quad \sum_{i=1}^{n} |\log(y_i) - (\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} - \sum_{\ell=1}^{L} v_{i\ell})| \tag{7a}$$

$$+\lambda \sum_{j,k} w_{jk} + \gamma|\boldsymbol{\beta}|_1 \tag{7b}$$

$$\text{s.t.} \quad g_{ik} \le G_{ik}, \qquad \forall i \in [n],\ k \in [K], \tag{7c}$$

$$\sum_{j=1}^{m} T_{ij} w_{jk} \le (\sum_{j=1}^{m} T_{ij})(1 - g_{ik}), \qquad \forall i \in [n],\ k \in [K], \tag{7d}$$

$$G_{ik} - \sum_{j=1}^{m} T_{ij} w_{jk} \le g_{ik}, \qquad \forall i \in [n],\ k \in [K], \tag{7e}$$

$$
\begin{aligned}
& g_{ik} \le s_{i\ell}, && \forall i \in [n],\ \ell \in [L],\ k \in S_\ell, && \text{(7f)}\\
& \sum_{k \in \mathcal{S}_\ell} g_{ik} \ge s_{i\ell}, && \forall i \in [n],\ \ell \in [L], && \text{(7g)}\\
& v_{i\ell} \le M s_{i\ell}, && \forall i \in [n],\ \ell \in [L], && \text{(7h)}\\
& v_{i\ell} \le r_\ell, && \forall i \in [n],\ \ell \in [L], && \text{(7i)}\\
& v_{i\ell} \ge r_\ell - M(1 - s_{i\ell}), && \forall i \in [n],\ \ell \in [L], && \text{(7j)}\\
& v_{i\ell} \ge 0, && \forall i \in [n],\ \ell \in [L], && \text{(7k)}\\
& r_\ell \ge 0, && \forall \ell \in [L], && \text{(7l)}\\
& s_{i\ell} \in \{0,1\}, && \forall i \in [n],\ \ell \in [L], && \text{(7m)}\\
& w_{jk} \in \{0,1\}, && \forall j \in [m],\ k \in [K], && \text{(7n)}\\
& g_{ik} \in \{0,1\}, && \forall i \in [n],\ k \in [K], && \text{(7o)}\\
& \boldsymbol{\beta} \in \mathcal{R}^D, \beta_0 \in \mathcal{R}. && && \text{(7p)}
\end{aligned}
$$

The objective function Eq. (7b) finds parameters to achieve high goodness-of-fit on empirical data (first term), as well as giving preference to simpler and sparser models (second and third terms). The penalty λ encourages fewer drug-target interactions identified, and γ is the coefficient for standard lasso-type ℓ_1 regularization (Tibshirani 1996). The values are selected via cross validation, a procedure where an unseen dataset is used to evaluate the best parameter choice for out-of-sample performance.

The constraints describe the biological mechanism and logical relationships in our model. Eqs. (7c), (7d), and (7e) jointly impose the logical definition in Eq. (5). Specifically, from Eq. (7c) if $G_{ik} = 0$ then $g_{ik} = 0$, i.e., if mutation G_k is not present in patient i, then gene expression g_k is normal for this patient. From Eq. (7d), if there is a therapy T_j that patient i received and T_j targets mutation G_k, then $g_{ik} = 0$. From Eq. (7e), if $G_{ik} = 1$ and $\sum_{j=1}^{m} T_{ij} w_{jk} = 0$, i.e., mutation is present but no targeting therapy is given, then $g_{ik} = 1$.

Eqs. (7f) and (7g) jointly impose the logical defined in Eq. (3). From Eq. (7f), if for some k in S_ℓ we have $g_{ik} = 1$, then $s_{i\ell} = 1$, i.e., if a gene expression g_k in patient i is abnormal, then all sets S_ℓ containing g_k are abnormal. From Eq. (7g), if $\sum_{k \in S_\ell} g_{ik} = 0$, then $s_{i\ell} = 0$; that is, if all of the gene expressions in S_ℓ are normal, then the set S_ℓ is normal.

Finally Eqs. (7h), (7i), and (7j) jointly define the nonlinear relationship between the s_ℓ and the coefficients r_ℓ as defined in Eq. (6).

The key output of the MIO (7) are the binary variables w_{jk} that represent our prediction of whether treatment T_j targets mutation G_k. We also produce confidence estimates for w_{jk} by using the bootstrap method via sampling the data with replacement (Efron 1979), solving the MIO (7) multiple times, and aggregating the results. To ensure the model generalizes well to different population, in addition to bootstrapping with random resampling, we also include subsamples based on some simple propensity model on the patient demographic information.

2.3. Evaluations

2.3.1. Model Fit Comparisons To validate that the MIO (7) model fits and generalizes better to real clinical outcomes than existing methods, we compare it against predictions made from linear models. We consider the following three models, each fitted on $\log(y)$ to minimize the total sum of absolute error, to be consistent with the MIO objective for a fair comparison:

- LM1: with phenotype covariates alone;
- LM2: with phenotype covariates plus treatments and gene mutations;
- LM3: with phenotype covariates, treatments, gene mutations, and the interactions between treatments and gene mutations. The interactions terms measure the incremental effects from a given treatment when a patient has a specific mutation. It is a common method for estimating individual treatment effects (Imai et al. 2013).

Regularization is used when the variables are in high dimensions. Each model is cross-validated to select best regularization parameter choice. Similar to bootstrapping with special subpopulation, the cross-validation procedure also includes subpopulation in addition to random sampled population splits. The in-sample and out-of-sample performance measured as mean absolute error (MAE) of each model is reported.

2.3.2. Inferred Therapy-Target Accuracy The key model output is the inferred values for variable w_{jk} on whether therapy T_j targets mutation G_k. It is, however, unknown what the ground truth is. Therefore, evaluating the accuracy of our inferred therapy-target pairs is not feasible.

To gain a rough sense on the quality of w_{jk} predictions, we use the current biomedical knowledge on existing drugs as a proxy for the ground truth. To do so, we obtain data from the Drugbank

database that has current knowledge on drug data and target information (Law et al. 2013). The current understanding of therapy-target relations from this database is considered as ground truth, where our findings are evaluated against it. The accuracy, true positive, and false positive rates are reported. Note that since the current knowledge has likely not discovered some true relationships, the estimated true positive and false positive rates from our model are quite conservative.

2.3.3. Further Validations We further validate the model externally on an additional dataset from the Cancer Genome Atlas (TCGA). TCGA is a program launched by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), with a goal to accelerate cancer research with comprehensive data on tumor samples. We extracted a similar set of genomic and clinical information from breast cancer, lung cancer, and skin cancer patients and ran the MIO (7) model to estimate the w_{jk}. The same set of model evaluations with model fit and inferred therapy-target accuracy were performed.

Finally, we perform an experiment on synthetically-generated data where the ground truth of the therapy-target relations variable w_{jk} is known. The goal is to validate that solving MIO (7) can recover the relationships assuming the model is an accurate description of the biological process. We first generate values for the principal variables w_{jk}, $\boldsymbol{\beta}$, and r_ℓ; next, we simulate the patient phenotype covariates $\mathbf{x}_i$, gene mutation G_{ik}, and treatments T_{ij}. Finally, we calculate the survival outcome y_i based on the model structure described in Section 2.2 with varying levels of additive noise. Because the ground truth for w_{jk} is known, we can evaluate the accuracy, true positive, and false positive rates of the inferred values.

The software used includes R (3.3.3), Julia (0.6.0), and Gurobi Optimizer (7.0).

3. Results

From the Dana Farber Cancer Institute (DFCI) data set, we obtained the genetic and clinical data of 3,118 eligible patients. Common cancer sites include breast, ovary, lung, and colon/rectum. In total, these patients initiated $n = 6,928$ anti-cancer regimens. The median survival from the initiation of therapy is 550 days. Phenotype covariates (133 variables) include race, age, gender,

tumor site, stage, metastatic status, comorbidities, laboratory test results, and major medications, with age at treatment (correlation coefficient of -0.1790, same measures below), age at diagnosis (-0.1626), and clinical stage IV (-0.1299) most negatively correlated with survival. The most frequently mutated genes are TP53, PIK3CA, and MYC (Figure 2, top) among $K = 176$ total measured gene mutations. The presence of most mutations are negatively correlated with the survival outcome, with TCF7L2 (-0.0615), CDKN2B (-0.0606), and PTEN (-0.0578) being the most negatively correlated. Among all prescribed chemotherapy drugs, the most commonly ones include carboplatin, paclitaxel, and cisplatin. Targeted therapies are less frequently prescribed, but we included the following $m = 10$ therapies that have sufficient observations: bevacizumab, trastuzumab, cetuximab, rituximab, ipilimumab, pertuzumab, pembrolizumab, panitumumab, erlotinib, and bortezomib. Among these, the use of trastuzumab (0.1351) and rituximab (0.0811) is most positively correlated with survival.

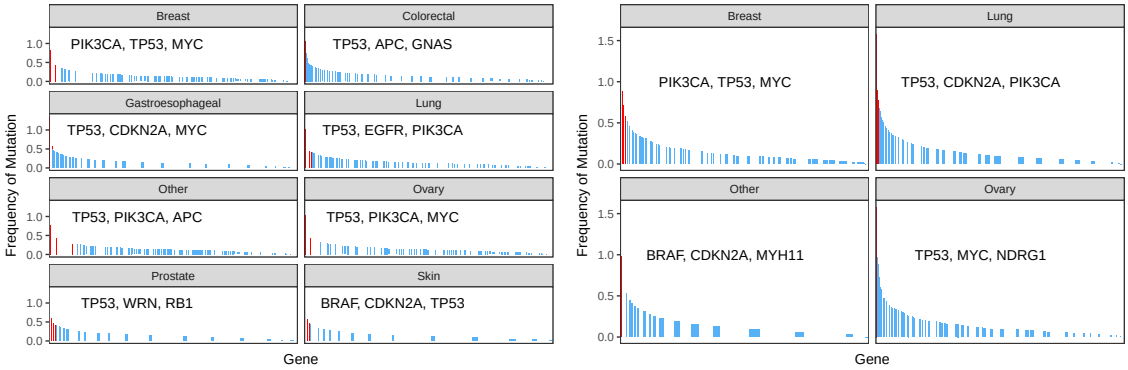

Figure 2 Most frequently mutated genes among all observations, stratified by the tumor site for DFCI on the top and TCGA on the bottom. The top three mutations for each site is highlighted in red and the gene is labeled.

From the Cancer Genome Atlas data set, we obtained the genetic and clinical data of 1,098 eligible patients from the following projects: breast invasive carcinoma, ovarian serous cystadenocarcinoma, lung squamous cell carcinoma, lung adenocarcinoma, and skin cutaneous melanoma. In total, these patients initiated 6,929 regimens. The median survival from the initiation of therapy is 529 days. Phenotype covariates used include race, gender, tumor site, age at diagnosis, history of neoadjuvant treatment, history of radiation therapy, and some breast cancer specific

ones (menopause status, estrogen receptor status, etc.). The most frequently mutated genes are TP53, PIK3CA, MYC, NDRG1, EXT1, RAD21 (Figure 2, bottom) out of a total of 299 common gene mutations under consideration. Among all prescribed chemotherapy drugs, the most commonly ones include paclitaxel, carboplatin, cyclophosphamide, doxorubicin, docetaxel, cisplatin, and tamoxifen. Targeted therapies are less frequently prescribed, but we included the following that have sufficient observations: bevacizumab, trastuzumab, erlotinib, ipilimumab, vemurafenib, gefitinib, denosumab, aldesleukin, and dabrafenib.

3.1. Model Fit Comparisons

Comparing the proposed mixed integer optimization (MIO) (7) model against linear regression models, we achieved better model generalization performance, as reflected in the lower mean absolute error (MAE) (Table 2) for both DFCI and TCGA data. Adding to LM1 (phenotypic information only) genomic data and treatment data as additional terms (LM2 and LM3) typically does not improve the generalization performance, even with appropriate level of regularization. In contrast, because the MIO (7) model directly optimizes the sum of absolute error while maintaining sparsity, it improves upon any linear models significantly under MAE as the goodness-of-fit metric.

Between the two data sources, the DFCI one has very rich phenotype data that are likely to be predictive of the survival outcomes (Bertsimas et al. 2018), and therefore was able to achieve lower MAE both in sample and out of sample. The TCGA data has abundant information on genomic data; however, even with regularization there is still some discrepancy between the in-sample and out-of-sample performance metrics. We attribute this discrepancy to the inherent high-dimensional nature of the data.

3.2. Inferred Therapy-Target Accuracy

The model outputs therapy-target pairs from solving MIO (7) with the DFCI data. Because the interactions are probabilities averaged from bootstrapped samples, at any chosen probability threshold cutoff, we can obtain a set of inferred interactions. We choose the threshold δ such that if the calculated confidence probability is greater than δ for a particular therapy-target combination,

Table 2 Model performance comparisons for DFCI and TCGA data. The in-sample and out-of-sample mean absolute errors are reported for each of the four models being compared.

Data	Model	In-Sample MAE	Out-of-Sample MAE
DFCI	LM1	0.65	0.73
	LM2	0.64	0.71
	LM3	0.62	0.71
	MIO	0.66	0.69
TCGA	LM1	0.89	0.92
	LM2	0.77	0.97
	LM3	0.73	0.99
	MIO	0.88	0.91

Table 3 Confusion table for estimated w_{jk} based on DFCI and TCGA data, comparing predicted therapy-target relationships against ones known from literature.

DFCI	Actual 0	Actual 1
Predicted 0	1684	3
Predicted 1	70	3

TCGA	Actual 0	Actual 1
Predicted 0	3524	11
Predicted 1	49	4

then we include this combination in the output we report, in a way to limit the number of distinct therapy-target combinations. When validating against known relationships, we aim at keeping around 60 predicted pairs, which correspond to a threshold of 0.34. The prediction achieved an accuracy of 95.8%, a true positive rate of 50.0%, and a false positive rate of 4.0% (Table 3, top section).

Table 4 Potential drug-target relationships identified by model, ordered by descending probabilities. The relationships known in literature are highlighted.

Drug	Target	Probability	Drug	Target	Probability
pembrolizumab	TP53	0.92	panitumumab	BCL2	0.48
trastuzumab	AKT2	0.74	cetuximab	PIK3CA	0.48
pertuzumab	PTK2	0.74	panitumumab	FGFR1	0.46
rituximab	SRC	0.74	pembrolizumab	MET	0.46
pembrolizumab	CDK4	0.70	cetuximab	MTOR	0.46
bevacizumab	MAPK1	0.70	ipilimumab	FLT4	0.44
pembrolizumab	MAPK1	0.70	cetuximab	PTK2	0.44
panitumumab	SRC	0.70	cetuximab	BRAF	0.42
panitumumab	IGF1R	0.68	panitumumab	JAK2	0.42
rituximab	PIK3CA	0.68	panitumumab	NTRK3	0.42
erlotinib	TP53	0.68	ipilimumab	RAF1	0.42
pembrolizumab	MTOR	0.66	rituximab	AKT1	0.40
rituximab	TP53	0.66	bevacizumab	AKT3	0.40
cetuximab	AKT3	0.64	pertuzumab	ERBB2	0.40
panitumumab	PTK2	0.64	cetuximab	IGF1R	0.40
pertuzumab	MAP3K1	0.62	erlotinib	MET	0.40
bevacizumab	PIK3CA	0.62	pertuzumab	PRKDC	0.40
panitumumab	TP53	0.62	trastuzumab	SRC	0.40
trastuzumab	AKT3	0.60	pertuzumab	SRC	0.40
trastuzumab	RAF1	0.60	trastuzumab	SYK	0.40
pembrolizumab	EGFR	0.58	pertuzumab	SYK	0.40
rituximab	PTEN	0.58	trastuzumab	AKT1	0.38
erlotinib	EGFR	0.56	pertuzumab	AKT1	0.38
pembrolizumab	FGFR3	0.56	ipilimumab	EGFR	0.38
bevacizumab	AKT1	0.54	bevacizumab	IGF1R	0.38
bevacizumab	BRAF	0.54	pertuzumab	TP53	0.38
trastuzumab	MAPK1	0.54	cetuximab	AKT2	0.36
bevacizumab	SRC	0.54	panitumumab	AURKA	0.36
bevacizumab	TP53	0.54	panitumumab	BCL2L1	0.36
panitumumab	MAPK1	0.52	panitumumab	PTEN	0.36
pertuzumab	PIK3CA	0.52	pertuzumab	PTPN11	0.36
cetuximab	RAF1	0.52	cetuximab	EGFR	0.34
cetuximab	SRC	0.52	ipilimumab	MAP3K1	0.34
pembrolizumab	BRAF	0.50	trastuzumab	PIK3CA	0.34
pembrolizumab	JAK2	0.50	bevacizumab	PRKDC	0.34

Sorting the predicted probabilities of therapy-target interactions gives a list to be further investigated in experiments (Table 4). We correctly recovered the well-known therapy-target pairs including erlotinib on EGFR, pertuzumab on ERBB2, and cetuximab on EGFR at this threshold. At a lower threshold of 0.20, we also discover temsirolimus on MTOR and panitumumab on EGFR. The relationship trastuzumab on ERBB2 was not identified, likely because data do not suggest pronounced incremental treatment effects from tratuzumab on patients with ERBB2 mutation.

We conducted similar analysis for the TCGA data. To obtain about 60 therapy-target pairs, we selected the threshold of 0.38. The prediction achieved an accuracy of 98.3%, a true positive rate of 26.7%, and a false positive rate of 1.4% (Table 3, bottom section). The prediction provides the list of most promising therapy-target to be further investigated in experiments (Table 5). We correctly

Table 5 Potential drug-target relationships identified by model, ordered by descending probabilities. The relationships known in literature are highlighted.

Drug	Target	Probability	Drug	Target	Probability
dabrafenib	BRAF	0.98	bevacizumab	PIK3CA	0.64
denosumab	IKBKB	0.98	pembrolizumab	LCK	0.62
vemurafenib	BRAF	0.90	vemurafenib	MET	0.62
ipilimumab	CACNA1D	0.86	trastuzumab	TP53	0.62
trastuzumab	MAPK1	0.86	pembrolizumab	GRIN2A	0.60
cetuximab	PRKACA	0.86	trastuzumab	IKBKB	0.58
dabrafenib	FGFR2	0.84	lapatinib	MAP2K2	0.58
denosumab	PIK3CA	0.84	bevacizumab	IKBKB	0.56
cetuximab	TP53	0.84	pembrolizumab	MLH1	0.56
erlotinib	NTRK1	0.82	lapatinib	PIK3CA	0.56
cetuximab	PIK3CA	0.80	pembrolizumab	PMS2	0.56
ipilimumab	AKT2	0.78	erlotinib	TSHR	0.56
pembrolizumab	BRAF	0.78	gefitinib	EGFR	0.54
gefitinib	TP53	0.76	lapatinib	ERBB2	0.54
vemurafenib	AKT2	0.74	trastuzumab	JAK1	0.54
dabrafenib	BCR	0.74	erlotinib	MDM2	0.54
cetuximab	MET	0.74	erlotinib	PDGFRA	0.54
dabrafenib	NFKB2	0.74	aldesleukin	AKT2	0.52
erlotinib	TP53	0.74	aldesleukin	BRAF	0.52
cetuximab	BRAF	0.72	bevacizumab	CACNA1D	0.52
dabrafenib	PTEN	0.72	dabrafenib	KDR	0.52
ipilimumab	TP53	0.72	cetuximab	MAP2K4	0.52
dabrafenib	PDGFRA	0.70	lapatinib	TP53	0.52
vemurafenib	PTEN	0.70	lapatinib	BCR	0.50
trastuzumab	PIK3CA	0.68	erlotinib	IKBKB	0.50
denosumab	FGFR1	0.66	lapatinib	MALT1	0.50
bevacizumab	GRIN2A	0.66	erlotinib	PIK3CA	0.50
cetuximab	JAK1	0.66	bevacizumab	POLE	0.50
ipilimumab	ATP1A1	0.64	trastuzumab	PRKACA	0.50
ipilimumab	BRAF	0.64	lapatinib	ATM	0.48

recovered the well-known relationships of dabrafenib on BRAF, vemurafenib on BRAF, gefitinib on EGFR, and lapatinib on ERBB2.

While interpreting the results, note that again since the known pairs used to validate our model against are not exactly the ground truth, as many pairs remain to be discovered, the reported model performance metrics of true positive and false positive rates may be overly conservative.

3.3. Synthetic Data Validation

In the validation studies with synthetic data, we confirmed that when the causal assumptions are met, the inferred values of w_{jk} are accurate. We ran the MIO (7) to assess the quality of recovery of w_{jk} based on synthetically generated data where the underlying parameters are known and the data follow the model structure. Figure 3 presents the model performance at varying levels of noise: the accuracy, true positive, and false positive rates are plotted against noise. At low levels of noise,

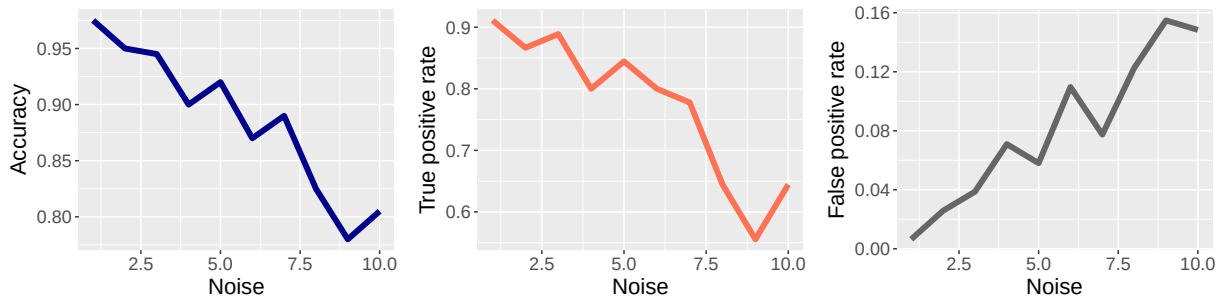

Figure 3 **Model performance on synthetic data for recovering the therapy-target relationship w_{jk}. The accuracy, true positive, and false positive rates against noise levels are presented.**

the model achieves almost perfect recovery of w_{jk}; as noise increases, the performance generally drops, but still remains reasonably high given the noise level.

4. Discussion

To our knowledge, this is the first study to leverage rich genomic and clinical information from observational *in vivo* data to inform drug discovery. It uses modern mathematical modeling tools to integrate data and biological knowledge base to identify the most likely therapy-target interactions. The network-based, MIO model is able to capture the non-linear interaction between therapies and targets. Compared to black-box machine learning models, it outputs interpretable chemical and biological processes while still allowing for the flexibility to represent the complex interactions across the patient inputs.

This novel approach to drug discovery has the following advantages: 1) it is based on large scale real-world evidence to infer potential biochemical mechanisms, overcoming the limitation of many *in vitro* experiments; 2) it incorporates the state-of-the-art understanding of cancer molecular biology and signaling pathways organically into a data-driven method; 3) it models patient heterogeneity explicitly outside genomic information, controlling for and capturing the rich information from patients' electronic medical records. As a result, the output of potential therapy-target interactions from this model was able to achieve high accuracy in external validation compared to other approaches.

The study has the following limitations: it is based on observational data, which unavoidably contain confounding factors. As a result, the learned relationships cannot be proven causal unless

the assumptions we made hold true. The set of gene mutations measured is relatively small (296 measured mutations) in the DFCI dataset; in particular, this dataset did not have information on immune cells, therefore restricting our inference on the potential targets for immunotherapies. In addition, the available human knowledge of gene sets is not complete. For future work, drug dosing information could be incorporated. As some of the target relationships are validated against findings in previous literature in this study, further validation should be conducted in prospective *in vitro* experiments, animal models, and clinical trials.

Extensions to the model are straight-forward. Because the MIO has a very flexible structure, it allows researchers to readily incorporate novel biological discoveries and updated data. In closer collaboration with cancer biologists, geneticists, and pharmacologists, we can further fine-tune the network model to reflect more comprehensive and new knowledge in the targeted therapy mechanism of action. The scalability of the model encourages further computational discoveries using data from other institutions and potentially data consortium from multiple institutions. As the data availability and quality improves, we believe the model will recommend new targets with even higher confidence and accuracy.

Matching patients to the right set of therapies, especially when the genetic signatures and clinical presentations are complex, is the ultimate goal for personalized cancer therapy. In this work, we present a general method with validated findings on novel therapy-target interactions, which provide a firm foundation for an evidence-based design of targeted treatment recommendations. Looking ahead, this approach will present not only significant opportunities for discoveries and breakthroughs in cancer medicine, but also higher quality, more individualized care for patients.

References

Bertsimas D, Dunn J, Pawlowski C, Silberholz J, Weinstein A, Zhuo YD, Chen E, Elfiky AA (2018) Applied informatics decision support tool for mortality predictions in patients with cancer. *JCO Clinical Cancer Informatics* 2(2):1–11, URL `http://dx.doi.org/10.1200/CCI.18.00003`.

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016) An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* 62(5):1511–1531.

Bertsimas D, Pawlowski C, Zhuo YD (2017) From predictive methods to missing data imputation: An optimization approach. *The Journal of Machine Learning Research* 18(1):7133–7171.

Bertsimas D, Weismantel R (2005) *Optimization over integers* (Dynamic Ideas Belmont).

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. *Science* 321(5886):263–266.

Chen B, Butte A (2016) Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology & Therapeutics* 99(3):285–297.

Chiang AP, Butte AJ (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics* 86(5):507–510.

DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics* 47:20–33.

Druker BJ (2004) Imatinib as a paradigm of targeted therapies. *Advances in cancer research* 91:1–30.

Dudley JT, Deshpande T, Butte AJ (2011) Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics* 12(4):303–311.

Efron B (1979) Bootstrap methods: another look at the jackknife. *The annals of Statistics* 1–26.

Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483(7391):570.

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *cell* 144(5):646–674.

Imai K, Ratkovic M, et al. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28(1):27–30.

Kesselheim AS, Treasure CL, Joffe S (2017) Biomarker-defined subsets of common diseases: policy and economic implications of orphan drug act coverage. *PLoS medicine* 14(1):e1002190.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science* 313(5795):1929–1935.

Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, et al. (2013) Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research* 42(D1):D1091–D1097.

Li J, Lu Z (2012) A new method for computational drug repositioning using drug pairwise similarity. *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference On*, 1–4 (IEEE).

Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z (2015) A survey of current trends in computational drug repositioning. *Briefings in bioinformatics* 17(1):2–12.

Li YY, Jones SJ (2012) Drug repositioning for personalized medicine. *Genome medicine* 4(3):27.

Mestres J, Gregori-Puigjané E, Valverde S, Solé RV (2009) The topology of drug–target interaction networks: implicit dependence on drug properties and target families. *Molecular BioSystems* 5(9):1051–1057.

Novartis (2017) Gleevec - prescribing information. URL `https://www.accessdata.fda.gov/drugsatfda_docs/label/2006/021588s009lbl.pdf`.

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery* 9(3):203–214.

Pritchard JR, Bruno PM, Gilbert LA, Capron KL, Lauffenburger DA, Hemann MT (2013) Defining principles of combination drug mechanisms of action. *Proceedings of the National Academy of Sciences* 110(2):E170–E179.

Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, Lopez-Bigas N (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer cell* 27(3):382–396.

Sawyers C (2004) Targeted cancer therapy. *Nature* 432(7015):294.

Shaw AT, Yasothan U, Kirkpatrick P (2011) Crizotinib. *Nature Reviews Drug Discovery* 10(12):897–898.

Spainhour JCG, Qiu P (2016) Identification of gene-drug interactions that impact patient survival in tcga. *BMC bioinformatics* 17(1):409.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43):15545–15550.

Tang J, Karhinen L, Xu T, Szwajda A, Yadav B, Wennerberg K, Aittokallio T (2013) Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. *PLoS computational biology* 9(9):e1003226.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for dna microarrays. *Bioinformatics* 17(6):520–525.

US Food and Drug Administration (2016) Drug innovation - novel drugs summary 2016. URL `https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/ucm534863.htm`.

Weinberg R (2013) *The biology of cancer* (Garland science).

Zerbini LF, Bhasin MK, de Vasconcellos JF, Paccez JD, Gu X, Kung AL, Libermann TA (2014) Computational repositioning and preclinical validation of pentamidine for renal cell cancer. *Molecular cancer therapeutics* 13(7):1929–1941.

Zhao B, Hemann MT, Lauffenburger DA (2014) Intratumor heterogeneity alters most effective drugs in designed combinations. *Proceedings of the National Academy of Sciences* 111(29):10773–10778.