Using deep bayesian neural networks for optimal treatment assignment in precision oncology. A contextual bandit problem.

Niklas Rindtorff, Nisarg Patel, Ming Yu Lu, HuaHua Zheng
Harvard Medical School

1 Introduction

The goal of precision medicine is providing the right treatment to the right patient at the right time. Oncology has been one of the central fields in modern precision medicine [1], with an increasing number of targeted compounds available, that are hypothesized to show activity in a very specific subset of patients (among the most popular examples being Imatinib for BCR-ABL positive CML).

Despite these early successes, assigning patients to adequate treatments remains a challenge until today. The current best practice in precision oncology is to base the treatment decision on published and frequently used therapeutic protocols that consider the patient's clinical characteristics and cancer genetics. For example, based on the status of a single biomarker, such as a BRAF V600E mutation, a treatment decision can be made [2].

These decision rules, or therapeutic protocols, are constantly evolving and used to link patients to optimal potential outcomes. Most therapeutic protocols in precision medicine are the result of biological reasoning that was validated in prospective [2] or retrospective [3] analysis of clinical trials. However, most clinical trials evaluate one biomarker and one targeted therapeutic at a time. This limits the ability to make high-confidence clinical decisions in a real world scenario with a large number of biomarkers to measure and potential treatments to choose from.

The consequences of strict therapeutic "if..when" protocols in precision medicine can be (I) high selectiveness - where only a small number of patients are assigned to a treatment with robust clinical evidence, (II) high compassionate use - where a majority of patients are left with limited treatment options that, if treated outside of a therapeutic protocol, do not contribute systematically to the development of new clinical evidence and (III) limited predictive ability - where due to a reduced number of new observations, therapeutic protocols can not improve beyond their initial version. For example, in recent precision medicine trials less than 50% of all screened patients were assigned to a treatment [6] and only a third of treated patients showed signs of increased progression-free-survival relative to their prior treatment [7].

Bandit problems, a class of problems in the field of policy learning, have their nominal roots in the rows of "one-armed bandit" slot machines seen in casinos. Each machine has a different probability of a payout and your goal is to maximize the total payout. You are limited by both the total number of bandits you can pull in a fixed period of time and uncertainty regarding which machine will deliver the best payout. The bandit problem here involves an exploration/exploitation tradeoff, i.e. the balance between trying different bandits to learn more about the potential payout of each slot machine and exploiting the best known bandit more to maximize the reward function.

The 'multi-armed bandit' algorithm outputs an action; however, it does not use information about the state of the environment, i.e. the context, to inform the output. The 'contextual bandit' extends the multi-armed bandit model by making the decision conditional on the state of the environment. This both optimizes the decision, choosing one action from a number of possible actions, based on prior observations and personalizes decisions for each situation.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

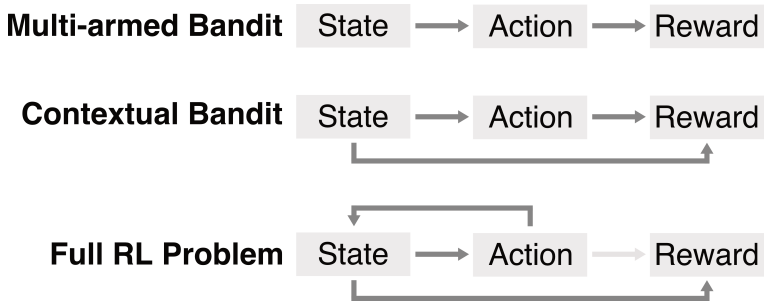

Figure 1: Overview of policy learning problems. Multi-armed bandits, contextual bandits and reinforcement learning problems have different causal relationships and delays between state, action and reward.

The contextual bandits algorithm observes a context, makes a decision, choosing one action from a number of alternative actions, and observes an outcome of that decision. However, in contrast to more complex reinforcement learning problems, the agent does not have to develop a strategy that spans multiple action-state pairs. However, in contrast to more complex reinforcement learning problems, the agent does not have to develop a strategy that spans multiple action-state pairs.

In this class project we aimed to formalize the administration of targeted therapeutics in oncology as a contextual bandit problem [10]. We curated a public dataset of 1000 cancer cell lines and their response to 7 different therapeutics. For every cell line we have aggregated mutation, CNV and gene expression data. We used a variety of algorithms, inlcuding Bayesian Neural Networks, to subsequently choose the best treatment for a randomly selected cell line after observing the genomic data. Based on the treatment decision, the agent will be rewarded with the in-vitro drug response the cancer cell line showed to the agent. Thus, the agent will try to learn the genomic predictors of drug response in the most efficient way possible. In parallel, we have collected a set of therapeutic protocols for each compound which reflect current clinical evidence. During this project, we integrated this prior knowledge into the available state information to facilitate the agent's learning.

The presented project has potential real-world implications for precision oncology. While in-vitro drug sensitivity screening of cancer cells can test every possible treatment response and link it to genomic predictors, this is not the case in precision oncology. Here, an individual patient is treated with a single drug and thus, only one potential outcome is realized and can be observed. Identifying strategies in which patients can be allocated to treatment options so that the overall patient benefit is maximized is a fundamental goal of precision oncology.

2 Related Work

Recently the number of targeted therapeutics tested in clinical trials has grown rapidly, prompting the introduction of modern master trial designs, such as baskets, umbrellas and platforms, thereby improving the throughput of efficacy testing [8]. Of note, first master protocols such as the I-SPY2 trial for neoadjuvant breast cancer therapy have employed an adaptive design based on Bayesian hierarchical models to guide treatment assignment and treatment arm discontinuation [4, 9]. Despite these changes in clinical trial design, most studies still stratify patients by a limited number of biomarkers and strictly adhere to pre-defined treatment arms.

Deep learning based methods have been used for the optimization of contextual bandit problems in the past and reach competitive performance to linear methods in most benchmarks in a Thompson sampling framework. As state information becomes more complex, the expressivity of linear methods shows clear limitations. Here we are used the author's public code base to evaluate a variety of algorithms, including Bayesian neural networks, for policy learning. We will compare the algorithms performance against other benchmarks, such as random allocation, Gaussian processes and, tailored to our scenario, current clinical evidence.

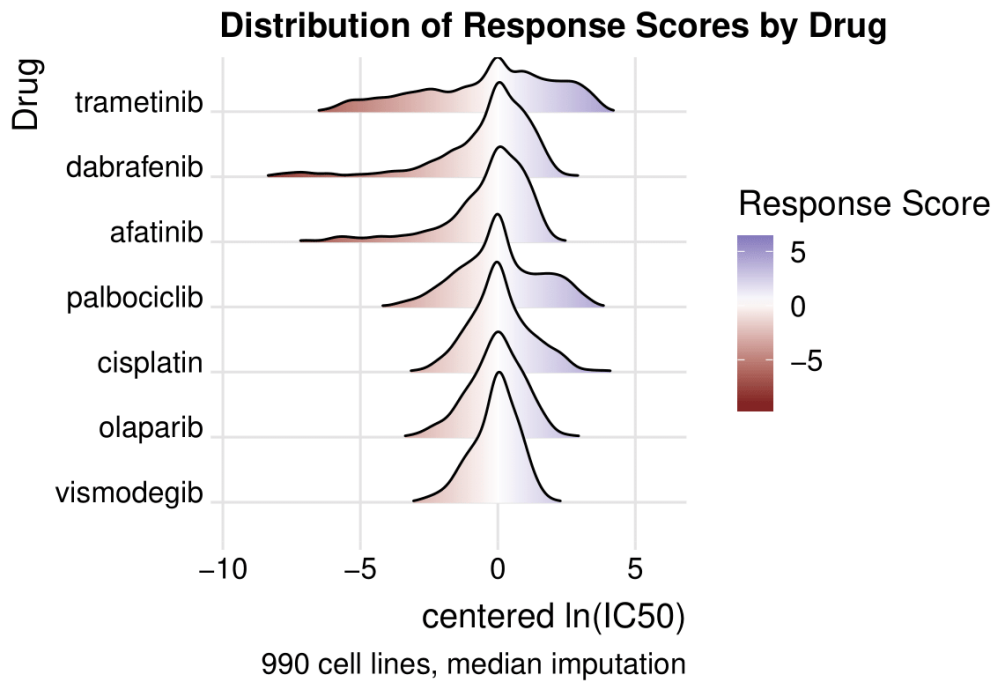

Figure 2: Distribution of Drug Response Scores after log-transformation and median centering of IC50 values.

2.1 Dataset

We synthesized a public dataset of in-vitro cancer cell-line drug sensitivity from [5]. In this dataset a complete matrix of treatment effects for >1000 cell-lines and >50 drugs has been measured and the IC50 values were recorded. Moreover, the dataset contains gene expression data, copy-number variants as well as mutation status for most cell-lines.

In a first pre-processing step we focused on a set of 7 drugs that are currently used in clinical practice. We then log transformed the $IC50$ values and normalized them relative to the median $ln(IC50)$ across cell-lines for each drug.

As we are comparing the effectiveness of potential outcomes on a cell-line level, the latter step might seem counter-intuitive. However, the pharmacokinetics and pharmacodynamics of therapeutic substances ex-vivo does, in general, not match in-vivo conditions. Thus, it is safer to estimate a treatment's in-vitro effect by comparing the transformed $IC50$ of a drug in a given cell-line relative to the median transformed $IC50$ of the same drug across the whole population of cell-lines.

Next we manually curated therapeutic protocols based on current clinical evidence and trial protocols with selected simplifications: (I) we excluded any protocols involving combination treatments, (II) we excluded any protocols that are based on the presence of oncogenic gene-fusions, (III) we did not include tissue type restrictions into any protocols.

We reduced 18523 cell-line specific features including scaled gene expression data, and binarized mutation and copy-number variant information into 20 dimensions by uniform manifold approximation and projection (UMAP).

UMAP projected features recovered tissue types 6 while not directly recovering overall drug sensitivities 7.

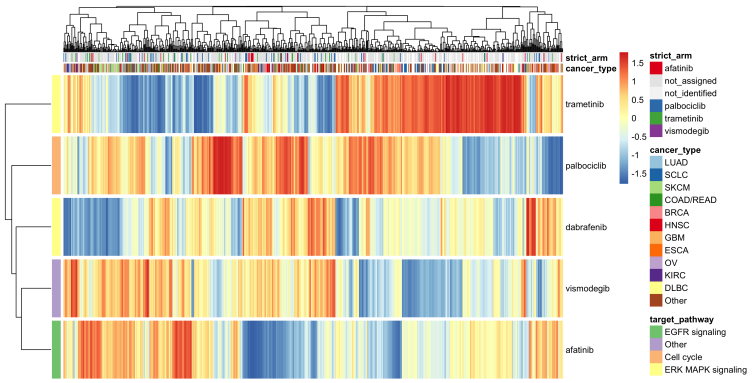

Figure 3: Drug vulnerability data for 1000 cell lines and 5 drugs. Blue color to drug sensitivity. The "strict arm" represents a cell line's treatment allocation based on current clinical evidence if it was a patient in a precision oncology program.

Therapeutic Protocols:

Trametinib - GNA11, NF1, BRAF non-V600E

Dabrafenib - BRAF V600E

Vismodegib - PTCH1

Afatinib - EGFR, ERBB2

Palbociclib - Rb expression & CCND1/ CDK4 amplification

Olaparib - BRCA1, BRCA2

Figure 4: Therapeutic protocols for 6 targeted agents. Every unit that did not match any of the above inclusion criteria was treated with the chemotherapeutic Cisplatin.

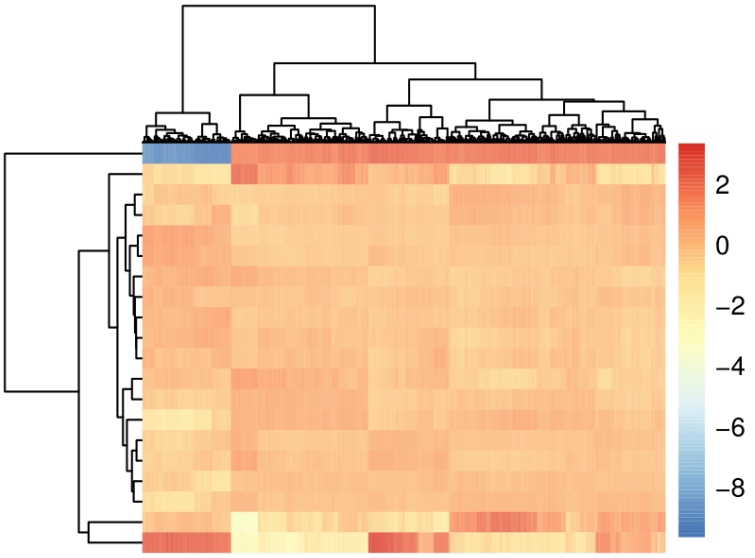

Figure 5: 20-dimensional embedding of genomic information for all cancer cell lines included in the study. Treatment covariates were summarized using uniform manifold approximation and projection (UMAP).

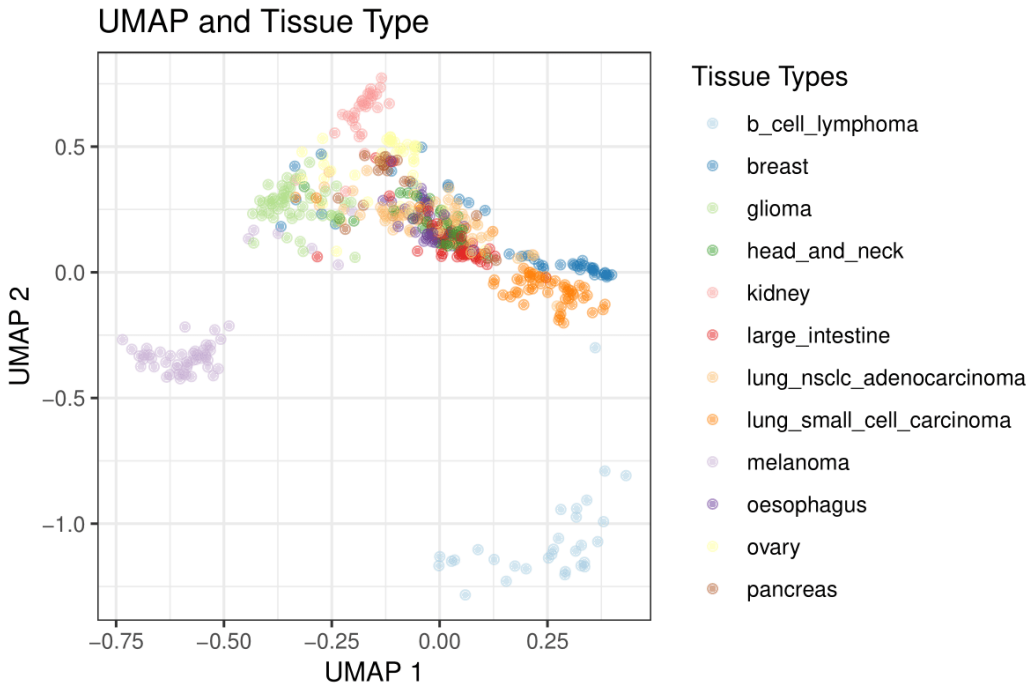

Figure 6: The first two dimensions of the UMAP embedding recovered tissue type differences between cancer cell lines

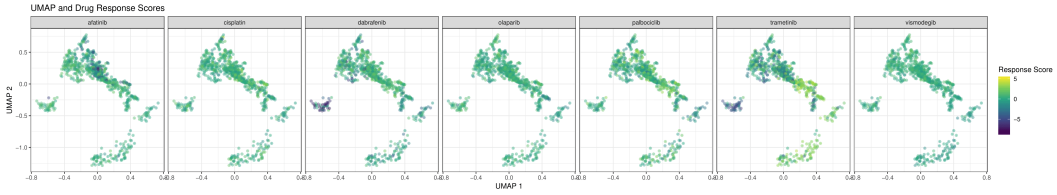

Figure 7: The first two dimensions of the UMAP embedding did not completely recover differences in drug sensitivity for most targeted agents across cancer cell lines

3 Methods

3.1 Contextual Bandit Models

The contextual bandit problem works as follows. At time $t = 1, ..., n$ a new context $X_t \in R$ arrives. The algorithm —based on its internal model and X_t selects one of the k available actions, a_t. Some reward $r_t = r_t(X_t, a_t)$ is then generated and returned to the algorithm, that may update its internal model with the new data. At the end of the process, the reward for the algorithm is given by $\sum_{t=1} r^t$,and cumulative regret is defined as R_A , where is the cumulative reward of the optimal $E[r * r]r*$ policy (i.e., the policy that always selects the action with highest expected reward given the context). The goal is to minimize R_A.

The tested algorithms included:

- **Linear Algorithm**: A baseline Bayesian regression implementation.
- **Neural Linear**: A neural network that predicts rewards for each context. The last layer was replaced with a Bayesian linear regression model.
- **Stochastic variational inference**
- **Expectation-Propagation**

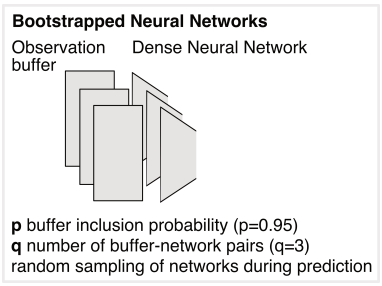

Figure 8: Overview of bootstrapped neural networks

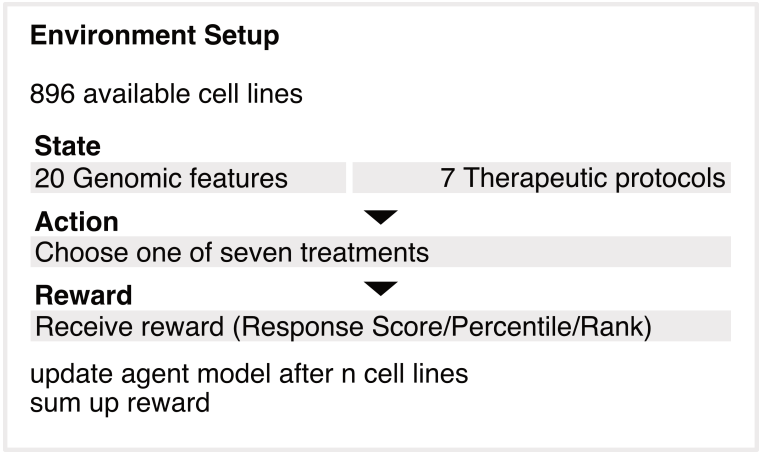

Figure 9: Environment Setup for the contextual bandit

- **Bootstrap**: A training technique where the output of each neuron is independently removed with a probability p during each evaluation. $D_1, ..., D_q$. These datasets are constructed by adding each new collected observation to every dataset independently with probability p. During evaluation, one of the q networks is selected with uniform probability. 8.

- **Dropout** - Dropout can be used to obtain a distribution of predictions for a specific input. Importantly, this network uses dropout during both training and evaluation. By using dropout during evaluation, the algorithm does not act completely greedy.

3.2 Implementation of Contextual Bandit Model

3.2.1 State X_t

To guide treatment decision making, the agent is provided with 20 representative genomic features and prior knowledge, formalized in one-hot encoded treatment recommendations based on current therapeutic protocols. Thus, in total, the state is represented using 27 features including unbiased genetic information and clinical evidence.

Throughout the study we evaluated different state representations, including both genomic data and clinical evidence, or clinical evidence alone.

3.2.2 Action A_t

The agent's actions are defined by choosing one of 7 medications 2. Actions are represented as one-hot encoded vectors with a corresponding index as the final output action.

	afatinib	cisplatin	dabrafenib	olaparib	palbociclib	trametinib	vismodegib
	0	1	2	3	4	5	6

Table 1: Treatment & Index

3.2.3 Regret R_t

After the agent's action was submitted, the reward was determined based on in-vitro drug response data of the cell-line that was represented in in the state information. Every state was only evaluated once during an experiment.

We used three ways to calculate the regret score for each cell line:

- The first method is to calculate the regrets for each drug by subtracting the lowest drug response score (the strongest response) from all other drug response scores. Thus the drug with the regret score 0 will be the best assignment 2.
- Second is the rank-based method, we ranked the drugs by drug response score in ascending order. The best drug will be ranked 7, while the least active drug will be ranked 1.
- Third is the percentile-based method. For each drug, we map its response score to its distribution over cell lines and use the percentile as reward.

	afatinib	cisplatin	dabrafenib	olaparib	palbociclib	trametinib	vismodegib
Response	0.0133365	-0.0499805	-1.014733	1.599829	-1.214424	-0.045493	0.5100345
Rank	5	4	2	7	1	3	6
Percentile	0.50	0.85	0.50	0.51	0.70	0.50	0.09

Table 2: Example Rank, Regret & Percentile for a single cell line

3.2.4 Experiment & Environment Settings

All experiments were run in python3.6 using a modified codebase from **Deep Bayesian Bandits Library**, including an additional therapeutic protocol based agent and a logging function to export agent's actions over all experiment steps.

For each subsets of features, We ran 5 trials with 100 training epochs for every sampling method with different random seed.

4 Results

Strict adherence to current clinical guidelines, as codified in the therapeutic protocols, consistently outperformed random allocation of treatments. However, most models were able to outperform agents that adhered to therapeutic protocols only. 10 Three Neural Network algorithms, bootstrapped-, greedy and Dropout based networks, consistently scored higher rewards compared to linear methods or Gaussian Processes, independent of the reward function. In addition, we evaluated alternative reward functions, such as the rank of each treatment option (from 1 to 7) or the percentile of the Response Score (from 0 to 1). 12, 13

Next, we evaluated the agent performance in a scenario with only therapeutic protocol assignments as state information. Most agents performed systematically worse in environments without genomic information, independent of reward function.

5 Discussion

In this work, our goal was to formalize the administration of targeted therapeutics in oncology as a contextual bandit problem using a public dataset of 1000 cancer cell lines, their genomic information, and their response to 7 different therapeutics. We tested a variety of algorithms, including Bayesian Neural Networks, to choose the best treatment for a randomly selected cell line, the 'action', using the genomic data as the 'state' of the agent and the drug response as the 'reward'. We compared the

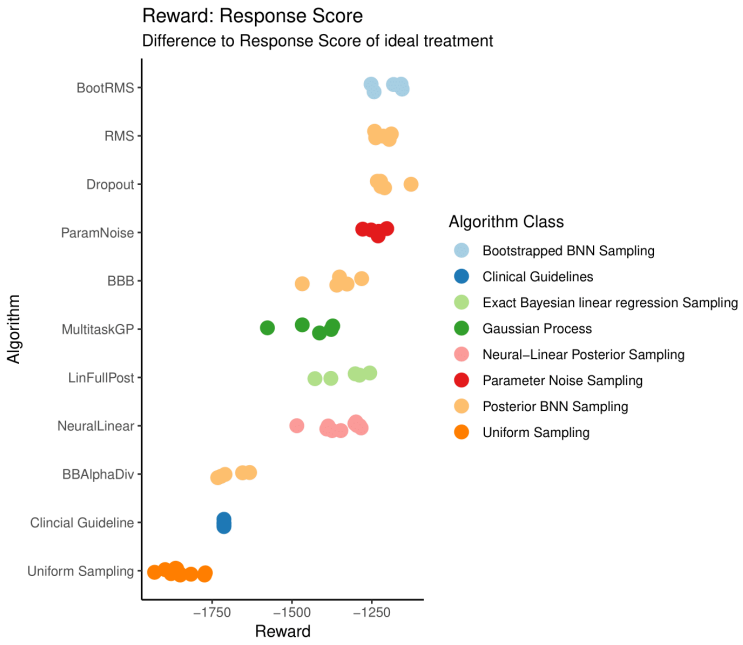

Figure 10: Reward function of Response Score (Difference of response score of chosen and ideal treatment) for each sampling method compared to clinical guidelines

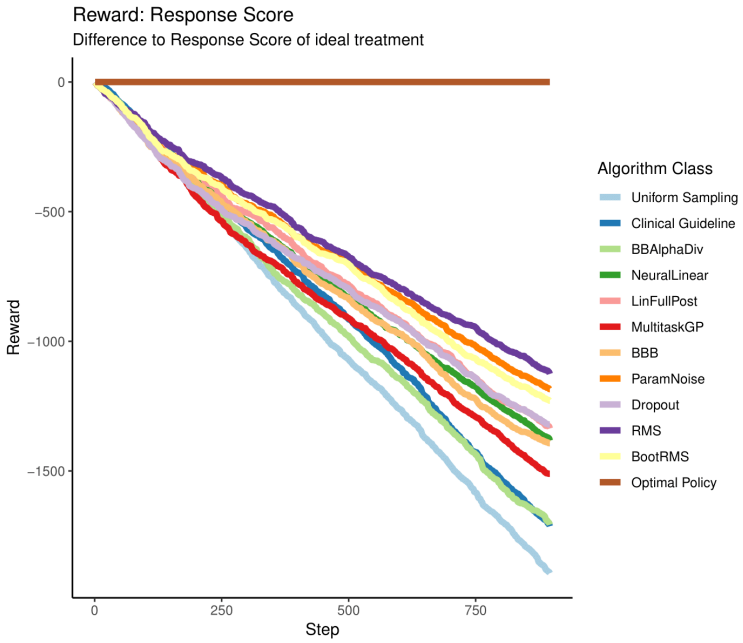

Figure 11: Change in treatment regret over each step for each algorithm tested, compared to clinical guidelines

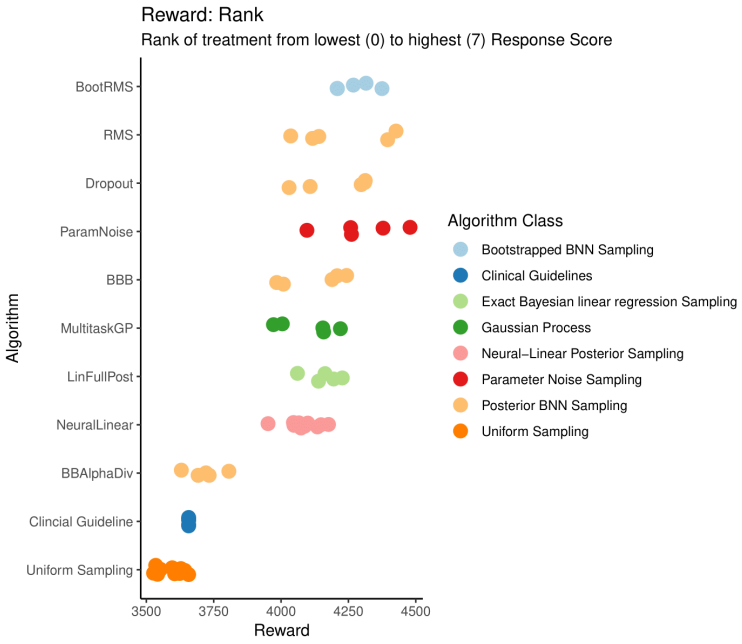

Figure 12: Reward function of Response Rank (Difference in rank of chosen and ideal treatment) for each sampling method compared to clinical guidelines

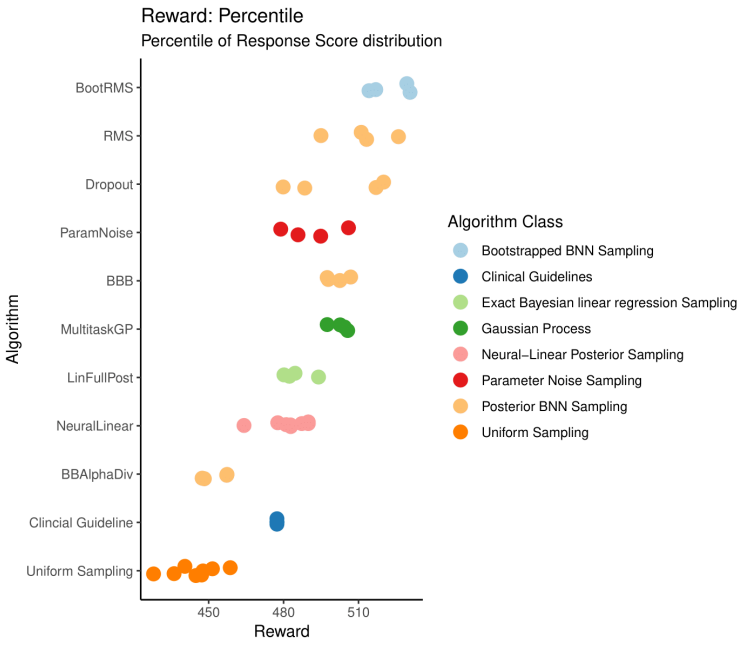

Figure 13: Reward function of Response Percentile (Difference in percentile of a sample among the distribution of response scores of chosen and ideal treatment) for each sampling method compared to clinical guidelines

Figure 14: Reward function of Response Percentile (Difference in percentile of a sample among the distribution of response scores of chosen and ideal treatment) of genomic and treatment information compared to treatment information alone

Figure 15: Reward function of Response Rank (Difference in rank of chosen and ideal treatment) of genomic and treatment information compared to treatment information alone

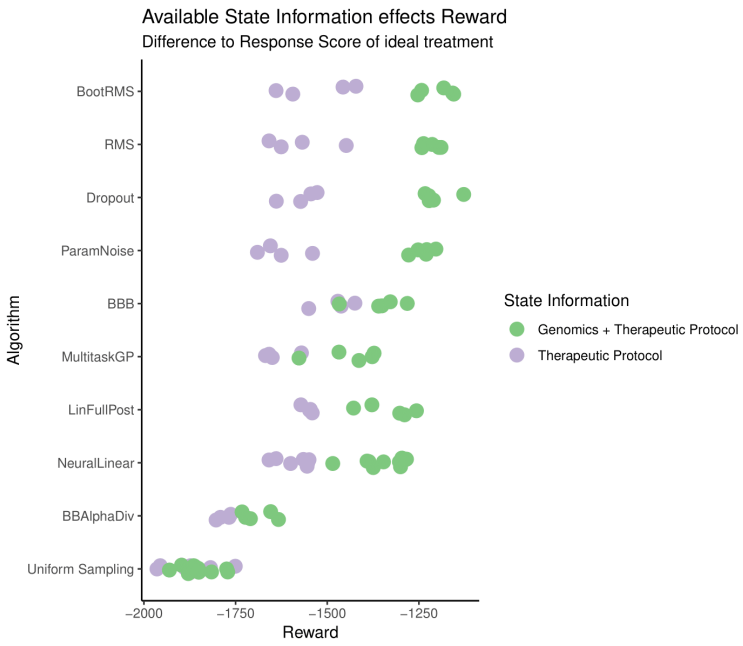

Figure 16: Reward function of Response Percentile (Difference in percentile of a sample among the distribution of response scores of chosen and ideal treatment) of genomic and treatment information compared to treatment information alone

performance of these models to the drugs chosen by clinical guidelines for each cancer cell line based on the presence of certain mutations.

In summary, we find that assignment mechanisms in precision oncology programs can be framed as a contextual bandit problem. When provided with genomic information and expert knowledge, contextual bandit agents outperform current clinical standards in an in-vitro cancer drug response dataset, in scenarios with three different reward functions, in-silico. Furthermore, the availability of genomic information increases the performance of most agents compared to clinical guidelines alone. Among the most successful agents were bootstrapped or simple dense neural networks that acted greedily. In principle, both bootstrapped and dropout networks add uncertainty information by sampling from multiple related models during prediction.

This study has several limitations including: (I) In-vitro drug response data of cancer models has limited transferability into a clinical context, (II) The response scores are on average lower in treatments vs. controls, (III) Cisplatin is a limited reference treatment for all considered cancer types.

A challenge we faced when designing this model was the high number of dimensions of the raw data relative to the sample size of the dataset. We addressed this challenge by performing UMAP for dimensionality reduction to reduce the number of features from over 18,000 to 20. The UMAP plot shows that the samples clustered by tissue type, revealing that either gene expression signatures or mutational signatures were preserved and represented in the final feature set.

In the future, we plan to validate our findings in alternative in-vitro drug response datasets, PDX experiments and pre-clinical Organoid model data. In addition, we plan to subsample the available genomic information and measure the impact on model performance. We would like to stimulate an open discussion about the limitations and potential benefits of AI guided treatment assignments in precision oncology program to minimize collective treatment regret. We acknowledge that further analysis needs to focus on avoidable regret on a per-patient level.

6 Disclosure

Niklas Rindtorff is working with the same curated dataset on a class project in CS282R, a Harvard run course in collaboration with Google Brain. Niklas Rindtorff has manually pre-processed the data and prepared it for use in causal inference of conditional treatment effects.

7 Code and Data availability

All code and data can be accessed in this repository or the following directory.

References

[1] David Blumenthal and Marilyn Tavenner. "A New Initiative on Precision Medicine". In: *The New England journal of medicine* 363.1 (2010), pp. 1–3. ISSN: 15334406. DOI: 10.1056/NEJMp1002530. arXiv: arXiv:1011.1669v3. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:New+engla+nd+journal%7B%5C#%7D0.

[2] Jeannie Hou et al. "Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation". In: *New England Journal of Medicine* 364.26 (2011), pp. 2507–2516. ISSN: 0028-4793. DOI: 10.1056/nejmoa1103782.

[3] Jonathan Ledermann et al. "Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: A preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial". In: *The Lancet Oncology* 15.8 (2014), pp. 852–861. ISSN: 14745488. DOI: 10.1016/S1470-2045(14)70228-1. URL: http://dx.doi.org/10.1016/S1470-2045(14)70228-1.

[4] Rajeshwari Sridhara et al. "Current Statistical Challenges in Oncology Clinical Trials in the Era of Targeted Therapy". In: *Statistics in Biopharmaceutical Research* 7.4 (2015), pp. 348–356. ISSN: 19466315. DOI: 10.1080/19466315.2015.1094673.

[5] Francesco Iorio et al. "A Landscape of Pharmacogenomic Interactions in Cancer". In: *Cell* 166.3 (2016), pp. 740–754. ISSN: 10974172. DOI: 10.1016/j.cell.2016.06.017. arXiv: NIHMS150003.

[6] Anthony Letai. *Functional precision cancer medicine-moving beyond pure genomics*. 2017. DOI: 10.1038/nm.4389. URL: http://dx.doi.org/10.1038/nm.4389.

[7] Christophe Massard et al. "High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: Results of the MOSCATO 01 trial". In: *Cancer Discovery* 7.6 (2017), pp. 586–595. ISSN: 21598290. DOI: 10.1158/2159-8290.CD-16-1396.

[8] Janet Woodcock and Lisa M LaVange. "Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both." In: *The New England journal of medicine* 377.1 (2017), pp. 62–70. ISSN: 1533-4406. DOI: 10.1056/NEJMra1510062. URL: http://www.ncbi.nlm.nih.gov/pubmed/28679092.

[9] Shinjo Yada and Chikuma Hamada. "Application of Bayesian hierarchical models for phase I/II clinical trials in oncology". In: *Pharmaceutical Statistics* 16.2 (2017), pp. 114–121. ISSN: 15391612. DOI: 10.1002/pst.1793.

[10] Carlos Riquelme, George Tucker, and Jasper Snoek. "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". In: (2018). arXiv: 1802.09127. URL: http://arxiv.org/abs/1802.09127.