CSTS Healthcare

Natalia Accomazzo[1], Bo Pan[2], Eric Rozon[1], Ellie Thieu[3], and Yiyu Yang[2]

[1]University of British Columbia
[2]University of Alberta
[3]Amherst College

September 9, 2021

Abstract

Cancer treatments have been developed to give the 'precision medicine' paradigm of cancer therapy for years. While we have had some success with specific cancers, many other patients are put on a roller coaster of emotion through cycles of remission and relapse. At this point, there is abundant scientific evidence that tells us cancer is driven by multiple genes working in concert. Thus, due to the complexity and heterogeneity of tumor, the design of personalized therapies that are unique to every individual are suggested. To date, precision oncology trails have been performed and, unfortunately, several of these trials have been hindered by very low matching rates. The reason of it is commonly because of the use of limited gene panels, restrictive matching algorithm, lack of drug availability and clinician's knowledge. Based on our previous work, we have developed a computational system that can help to identify a personalized cancer therapy for every cancer patient, given their unique set of DNA and RNA. For each patient, our system identifies the best of set of target genes and active hallmarks, each of which have sets of available corresponding therapies associated with them. In a clinical setting, however, a clinician prescribes the therapy they believe is most appropriate for a given patient. In this paper, we would like to construct a set of similarity measures that allow us to compare our Aiomic therapies with those actually given by oncologists and we are interested in the evaluation of the association between the adoption rate of Aiomic therapies and clinical measurement, which provide us the evidence about the performance of the system and help to make the improvement accordingly.

1 Introduction and Background

Cancer is a disease that affects 14M people each year. While we have had some success with specific cancers, many other patients are put on a rollercoaster of emotion through cycles of remission and relapse. At this point, there is abundant scientific evidence that tells us cancer is driven by multiple genes working in concert.

Traditionally in medicine, drugs are designed and approved by testing on large populations. This type of results in only a portion of the population actually responding to a given therapy. However, different patients respond differently to the same drugs. Typically, on a large double-blind clinical trial with thousands of patients and two arms, everyone on one arm receives the same therapy A, while everyone else on the other arm receives exactly the same therapy B.

In the past several decades, we have learned that cancer is a highly heterogenous disease with multiple genes involved in cancer progression, and therefore requires combination therapies tailored to each individual's life history and genomic profile. Based on multiomic profiles one can model the cancer biology: which genes are driving the cancer, and which of the 10 hallmarks of cancer are active.

This allows the design of personalized therapies that are unique to every individual. However, this also presents a novel statistical problem, called the N-of-One problem, where it is difficult to achieve statistical power. Because each patient in a trial is receiving a differing, unique therapy, these therapies are not obviously comparable. One recent study, the I-PREDICT trial attempted to provide a comparison of personalized combination therapies when the therapy was only partially adopted. They did not directly evaluate personalized therapies, but simply determined that when a given therapy targeted more mutations, it was correlated with better outcomes.

2 Knowledge Gap and Problem Statement

We have developed a computational system that identifies a personalized cancer therapy for every cancer patient, given their unique set of DNA and RNA. For each patient, our system identifies the best of set of target genes and active hallmarks, each of which have sets of available therapies associated with them. In a clinical setting however, a clinician prescribes the therapy they believe is most appropriate for a given patient. In this context, if there are 100 patients, there are 100 unique therapies. Moreover, the therapy a patient receives may not match what our system identified as the best therapy. This problem then breaks down into the following sub-problems:

- We would like to construct a set of similarity measures that allow us to compare our Aiomic therapies with those actually given by oncologists.
- Given the entire set of patients, can we quantify adoption rates of Aiomic therapies
- Can we measure outcomes for Aiomic therapies as to whether they succeeded, partially succeeded or failed?

In contrast to the I-PREDICT approach, for our problem, we are interested in not just matching gene targets, but evaluating the success of Aiomic therapies when the given therapies are not an exact match. The problem can be stated as follows:

- How can we compare Aiomic-Therapies to Given-Therapies? For example, would it simply be the % intersection between the set of targets and hallmarks that are covered by the drugs in each therapy?
- Across all patients, given the set of Aiomic-Therapies and Given-Therapies, can we say how often Aiomic therapies were adopted? To what degree?
- Across all patients, given outcomes, when a Given-Therapy is not exactly the same as a Aiomic- Therapy, and has non-zero overlap, can we assign an outcome to the corresponding Aiomic- Therapy?
- How much of a Given-Therapy outcome can we allocate to the Aiomic-Therapy in cases where there is partial overlap between the recommended therapy and the given therapy?
- If the adoption rate was 30%, is 30% of the outcome due to the recommendation? That is to say, if only one of the drugs from the recommendation were used in the given therapy, what % is attributable to the recommendation.
- Can we create a predictive model for partial adoption of our therapies?

3 Methods

3.1 Matching score using graphs

We know the aionic therapy identifies a subnetwork of genes that is the most active for each patient. This in turn we can think as a subgraph and derive our analysis from graph theory. An important measure that comes into place is the *first Betti number* or *circuit rank*, defined as $|E| - |V| + |C|$, where E is the set of edges, V the set of vertices and C the set of connected components. In a given graph, we define a connected component as a maximal subgraph that satisfies that any two vertices are connected by a path.

In our problem at hand, by the aiomic algorithm we have identified a certain subgraph of the gene map, let's call it G. From there, the proposed therapy would try to minimize the circuit rank of the subgraph generated by subtracting the vertices and edges from G that the proposed drugs target. For a given therapy T we denote this number by B_T. In this context, it seems natural to try to quantify the change of the circut rank of the subgraph generated by subtracting the vertices and edges that correspond to the targeted genes of the actual therapy T' that the patient receives. We could propose as matching score the quantity $B_{T'}/B_T$.

3.2 Similarity measure: Jaccard similarity coefficient

Analysis of similarity of personalized cancer therapy identified by the system and the therapy clinician prescribed help us understand the gap between the "precision medicine" paradigm of cancer therapy and the therapy that patients actually received. Essentially, the presence and absence of drugs are surveyed clinical research using sequencing, imaging and other technique. Then, the Jaccard coefficients is one of the most fundamental and population similarity measures to compare such presence-absence data [?]. In section 3.1, we give a brief introduction of the Jaccard coefficient and we present a hypothesis test for similarity for presence-and absence data, using Jaccard coefficient based on the boottrap procedure [?] in section 3.2. The boottrap procedure is considered in order to overcome the computational burden due to the high-dimensionality. And, we claim that, if the asymptotic distribution of similarity exist and shown to be normal distribution, the present bootstrap hypothesis test can be used for any such similarity. Finial, we close the section by introduction the population hypothesis test consider using extreme value distribution [?].

3.2.1 Jaccard Similarity Coefficient for Presence-Absence Data

Due to the complexity of data set and lack of clinical knowledge, we skip the data process procedure. Consider that, each patients are given two potential therapies, one is recommended by our system and the other is the one prescribed by the clinician. After processing the data, we image that the two therapies are in the format of presence-absence vectors [?], which could be the list of the targeted gene or of the drugs recommended. Ideally, any meaningful similarity to match two therapies should display the following desirable properties:

- **Quantification:** Different similarity measures force on different types of association, such as Pearson's correlation measuring the linear relationship between variables. It is important to select the one that satisfied our request and can be used to quantify the aspect we are interested in.
- **Interpretations:** Thanks to the machine learning technique, we are able to design a specific algorithms that can be used to calculate the similarity. Unfortunately, most of them are the 'black box' computing, which is hard to do the interpretation and lack of physical meanings. Thus, the similarity measure that could be explainable is the most fit one, especially for the clinical research.
- **Statistical guarantees:** Most of similarity measures lack probability interpretations or statistical error control. And, its statistical properties, hypothesis testing, and estimation methods for p-values have been inadequately studies. Thus, a rigorous statistical test evaluating the similarity is necessary.

The one we consider here is Jaccard coefficient. Given two presence-absence vectors A and B of length m that represent two two different therapies, the Jaccard similarity coefficient is the ratio of their interaction to their union. The set A and B can be viewed as the targeted genes or the recommended drugs for the unique individuals. This quantification of overlaps allows us to quantify matched genes / drug. To explain the basic idea of Jaccard coefficient, as a toy example, suppose we two set of drugs $A = \{1, 1, 1, 1\}$ and $B = \{0, 1, 0, 0\}$. Then, the union is $A \cup B = \{0, 1, 0, 0\}$ and the intersection between the sets is $A \cap B = \{1, 1, 1, 1\}$. Jaccard coefficient can be computed based on the number of elements in the intersection set divided by the number of elements in the union set.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{1}{4}$$

Note that $0 \leq J(A, B) \leq 1$. The higher the percentage, the more similar the two sets. The formula to find the Index is:

Jaccard Index = (the number in both sets) / (the number in either set) * 100

The formula in notation is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

In Steps, that's:

- Count the number of members which are shared between both sets.
- Count the total number of members in both sets (shared and un-shared).
- Divide the number of shared members (1) by the total number of members (2).
- Multiply the number you found in (3) by 100 .

This percentage tells you how similar the two sets are.

- Two sets that share all members would be 100% similar. the closer to 100%, the more similarity (e.g. 90% is more similar than 89%).
- If they share no members, they are 0% similar.
- The midway point −50%− means that the two sets share half of the members.

3.2.2 Bootstrap Procedure for N-to-1 Clinical Trail: Patient level

From the statistic perspective, statistical hypothesis testing using this similarity coefficient provides the confidence of the result. To evaluate whether A and B are independent, a following statistical hypothesis testing is performed in the patients level:

$$H_0 : J^c(A, B) = 0, \quad H_1 : J^c(A, B) \neq 0$$

The null hypothesis H_0 is that the centered Jaccard coefficient equals zero. Note that this is equivalent to that the conventional Jaccard coefficient equals an expected value under independence. However, like most similarity coefficient, the Jaccard coefficient lacks probabilistic interpretations and statistical error control. Such problems could make results lack of generalization and confidence. In additional, another challenging presents a novel statistical problem, called the N-of-1 problem [**?**], where it is difficult to achieve statistical power. Because each patient in a trial is receiving a differing, unique therapy, these therapies are not obviously comparable.

In order to utilize the Jaccard similarity coefficient, Chung, N. C. (2019) [**?**] propose a family of methods and algorithms. As indicated in Chung's paper, an unbiased estimation of expectation and a centered Jaccard coefficient has been proposed and an exact distribution of Jaccard similarity coefficients under independence that is shown to provide accurate p-values.

Proposition 1 *(Asymptotic property)[**?**] Given A and B are independent then,*

$$\sqrt{m} J^c(A, B) \to \mathcal{N}\left(0, \sigma^2\right), \qquad \textit{as } m \to \infty$$

Here, based on Chung, N. C.[**?**]'s work, we present a rigorous statistical testing to evaluate the similarity in presence-absence data, deriving statistical asymptotic properties and estimation of significance of the Jaccard coefficient.

Because the exact solution for a large m is computationally expensive or small m for lack of power, the bootstrap procedure is proposed to approximated the distribution of Jaccard coefficient. The bootstrap procedure has gained popularity for its wide applicability and statistical learning. The basic idea of bootstrap is that, by using resampling method, we could create an empirical distribution that converge to exactly distribution almost surely, And, it allows for estimation for p-values. Thus, we can access the significance of $J^c(A, B)$. In particular, resamping method with replacing A and B separately, breaks the potential dependency and make the independent assumption valid. Thus, we would be able to calculate an empirical distribution of Jaccard coefficients under the null hypothesis.

The advantages of using bootstrap procedure is (1). The expectation of Jaccard coefficient can be estimated directly from resampled vectors $A^\star$ and $B^\star$; (2). Each iteration provides randomness, which helps avoid a bias related to using an estimated expectation based only on observation. (3). Under the setting: N-to-1 clinical trails, we could avoid the request of large samples and be able to performing the statistical hypothesis testing for each unique patient.

Algorithm 1: Bootstrap Procedure for Jaccard coefficient

```
Input: Two binary therapy $A$ and $B$;
1. Calculate a centered Jaccard coefficient;
while $k = 0, 1, \cdots$ do
    Resample with replacemebt $A$ and $B$, resulting in $A^\star$ and $B^\star$;
    Calculated boottrap null coefficients
end
Compute the $p$-value by
    $p\text{-value} = \frac{\mathbf{1}\{|t_b^*| \geq |t|; b = 1, \ldots, B\}}{B}$
```

3.2.3 Population Hypothesis Testing

In this section, instead of considering patient level analysis, we focus on the hypothesis test for group of patients. Let assume we have a vector of Jaccard coefficient, J. we consider to using the minimum extreme value distribution to evaluate the significance of Jaccard coefficient borrow the idea of Rahman et al. (2014). Rahman et al. (2014)[**?**] proposes a method to compute a p-value of a Jaccard coefficient using an extreme value distribution.

For the statistic hypothesis, we need to find a statistic that characterize the samples and can be used of testing. We are often interested in extreme values of a parameter, like minimum Jaccard coefficient in our study and minimum strength, minimum force, minimum net income in a stock, because they are the values that determine whether a system will potentially fail or the minimum benefit that guaranteed. For example: minimum net income in a stock - it must be arranged to be at least greater than zero to prevent the cost; minimum risk of prescribed therapy that ensure patients is in safe side; modeling the extremes of meteorological events is shown to be necessary since these cause the greatest impact. It worth noting that the extreme value distributions are asymptotic results, meaning that the probability distribution of the minimum of a set of m independent values drawn from some distribution approaches the extreme value distributions only as n approaches infinity.

$$\text{Probability Density Function: } f(x) = \left(\frac{1}{b}\right) \exp\left(-\frac{x-a}{b}\right) \exp\left[-\exp\left(-\frac{x-a}{b}\right)\right]$$

Measuring statistical significance of the hits: The significance of the hits returned from the database can be inferred from the p-values derived from the z scores of the similarity.

The mean (μ) and s.d. (σ) of the similarity scores are used to define the z score, $z = (J - \mu)/\sigma$. For the purpose of calculating the p-value, only hits with $J > 0$ are considered. The p-value is derived from the z score using an extreme value distribution. And the p-value is calculated as below:

$$P = 1 - \exp\left(-e^{-z\pi/\sqrt{6}-\Gamma'(1)}\right), \text{ where the Euler-Mascheroni constant } \Gamma'(1) \approx 0.577215665.$$

3.3 Statistical Analysis

In this section, we give the pipeline of the statistical analysis. Note that we only provide the general framework and some necessary adjustment will be made based on the data set received. The primary outcome is to examine the impact of the matching score on the clinical measure, such as survival time. Although, it is a indirect evidence, it provides the idea of accuracy of the the system and could help the improvement. Before go that deeper, we first start with some exploratory analysis which can help locate the population and give the general picture of the samples.

Preliminary Screening: Prior to conducting the exploratory factor analysis, preliminary screening will be conducted. Data will be first screened for inclusion/exclusion, consent, end of study completed, missing data. All the inclusive and exclusive condition need to be satisfied and the data of participants can only be used after consent.

Descriptive Statistics: Patients' demographic information and clinical characteristics will be examined with descriptive statistics, using frequency (percentage) for categorical variables and mean (standard derivation), median (interquartile range) and range for continuous variables as appropriated.

Survival Analysis: Survival analysis will be used to study the association between the similarity of therapies and clinical measurement (time-to-death). Logrank test, Wilcoxon test, Fleming test and Kaplan-Meier analysis will be used to visualize the estimated probability of survival given specific time point and compare groups of patients (patients with similarity $\leq \alpha$ vs. the rest). p-values ≤ 0.05 are considered significant. And p-values-values will be adjusted for multi-comparison if needed.

In order to adjusting the validation cased by patients, mixed effect univariate and multivariate cox proportional hazards regression models will be used to estimate the hazard ratio of similarity coefficient to the survival time. The proportional hazards assumption will be tested by assessing Schoenfeld residuals and by plotting the negative logarithm of the estimated survivor function against the log time using log plots.

3.3.1 Mixed effects Cox Regression Models

Mixed effects cox regression models are used to model survival data when there are repeated measures on an individual or some other reason to have both fixed and random effects. The mixed effect cox regression model fits the model

$$\lambda(t) = \lambda_0(t)e^{X\beta+Zb}, \qquad \text{where } b \sim G(0, \Sigma(\theta))$$

where λ_0 is the baseline hazard function, X and Z are the design matrices for the fixed and random effects, respectively, β is the vector of fixed-effects coefficients and b is the vector of random effects coefficients. The random effects distribution G is modeled as Gaussian with mean zero and a variance matrix Σ, which in turn depends a vector of parameters θ.

The main idea of mixed effects cox regression models is that they make specific assumptions about the variation in observations attributable to variation within a subject and to variation among subjects. Under our setting, we are consider the variation is coming from the uniqueness of the patients.

3.3.2 Mixed Effect Model

To make the explain the mixed-effect in a easy way, we break the cox regression model into two part, linear model and link function, where we have linear model as $Y = X\beta + Zb$ with link function: $g(t) = \lambda_0(t)e^Y$. It's not hard to see the cox regression is kind like perform a map (link function) on a linear model.

We use a simple notation for convenient and ignore the link function for now. let Y_{ij} denote the response of subject $i, i = 1, \ldots, n$ at time $X_{ij}, j = 1, \ldots, n_i$ and $\beta_{i0} + \beta_{i1}X_{ij}$ denote the line that characterizes the observation path for i. Note that each subject has an individual-specific intercept and slope. Note that

- The within-subject variation is seen as the deviation between individual observations, Y_{ij}, and the individual linear trajectory, that is $Y_{ij} - (\beta_{i0} + \beta_{i1}X_{ij})$.

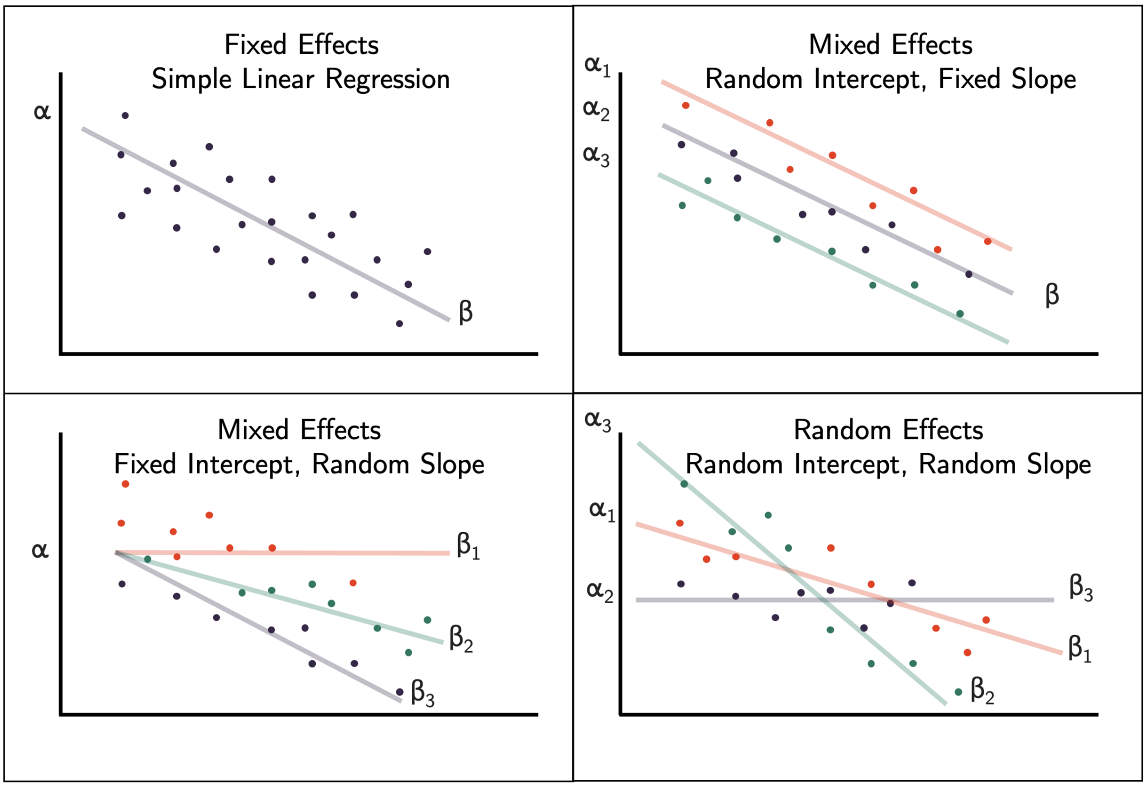
$$E\left(Y_{ij} \mid \beta_i\right) = \beta_{i,0} + \beta_{i,1}X_{ij}, \quad Y_{ij} = \beta_{i,0} + \beta_{i,1}X_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

- The between-subject variation is represented by the variation among the intercepts, $\operatorname{var}(\beta_{i0})$ and the variation among subject in the slopes $\operatorname{var}(\beta_{i1})$.

$$\begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} \sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix}\right]$$

  where D is the variance-covariance matrix of the random effects, with $D_{00} = \operatorname{var}(b_{i,0})$ and $D_{11} = \operatorname{var}(b_{i,1})$

From the following figure, we can say that (A) A randomintercepts model where the outcome variable Y_{ij} is a function of predictor X_{ij}, with a random intercept for study ID. Because all individuals have been constrained to have a common slope for predictor X, their regression lines are parallel. Solid lines are the regression lines fitted to the data. Point colour corresponds to study ID of the data point. The black line represents the global mean value of the distribution of random effects. (B) A random intercepts and random slopes model, where both intercepts and slopes are permitted to vary by group. Random slope models give the model far more flexibility to fit the data, but require a lot more data to obtain accurate estimates of separate slopes for each group.



3.3.3 Multivariable-Level Analysis

Mixed Effects Cox Regression Models will be used in multivariable-level analysis in order to adjust patients' individual effect and to quantify the effect of selected variables when they cooperate by presenting the coefficient (standard error), 95 % confidence Interval and P-value. The variables included in the multivariable analysis was selected if they were statistically significant on the univariable level analysis and considered clinically significant by the research team.

Multivariable-Level Analysis help to identify the risk factor and quantify their impact when they work together. All the risk factor that shows significant could be consider as the main factor to be used in the system to predict the gene and help increase the accuracy. Besides, several advantages can be used to help identify the factor, such as variable selection technique.

4 Conclusion and Limitations

In this paper, we present a general frame for analysis the data. Specially, we propose to using Jacard coefficient as well as bootstrap procedure for statistical testing. Although, bootstrap has its advantage of lower computation cost, unfortunately, when the size of samples are extreme small, there is no statistical guarantee and the result is obvious lack of reliability. Such restriction should be take into consideration when it comes to the real-data analysis and other method can be consider to address this issue. Another changeling is that potential problem with this score based on graph theory: We know that drugs in general target more than only one gene. Potentially, this gene could very well be outside of our principal subgraph G, but could have interactions. How can we incorporate this into our matching score? We leave these problems for later work

Acknowledgements This works was supported by the PIMS Math industry work shop and we would like to express our appreciations to our academic mentor, Dr.Sang, andindustrial mentors, Ali Hashemi and Giannoula Lakka, for advice and assistance with their knowledge on this project.