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

November 2, 2021

Abstract

Cancer treatments have been developed according to the 'precision medicine' paradigm of cancer therapy for years. While there has been success for specific cancers, many other patients still face cycles of remission and relapse. At this point, there is abundant scientific evidence that shows that cancer has a polygenic basis. Due to the genetic complexity and heterogeneity of physical tumors, the design of personalized therapies have been suggested to provide a potential avenue for effective treatment. To date, precision oncology trials have been performed and, unfortunately, several of these trials have been hindered by very low agreement between what is suggested and the actually therapy being prescribed. The commonly suggested reasons for this failure are because of the use of limited gene panels, the use of a restrictive matching algorithm, the lack of drug availability and limited clinician 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 DNA and RNA sequences. For each patient, our system identifies the best set of target genes and active hallmarks for the cancers, each of which have sets of available corresponding therapies associated with them. Unfortunately, the suggested therapies are not actually prescribed in the clinical setting; however, a clinician prescribes the therapy they believe is most appropriate for a given patient. In this paper, we would like to evaluate the level of agreement between our Aiomic therapies with therapies prescribed by oncologists through a set of similarity measures. We are interested in evaluating the association between the adoption rate of Aiomic therapies and clinical measurements, which provide us the evidence about the performance of the system and help to make improvements.

1 Introduction and Background

Cancer is a disease that affects 14 million people each year. While we have had some success treating specific cancers, but many patients face cycles of remission and relapse. 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 medical history and genomic profile. Multiomic profiles of the individual patients can characterize the genetic predisposition to cancer: which genes are driving the cancer, and which of the hallmarks of cancer are active. To date, the therapeutic management has shifted the weight from the selection of patients with cancer towards their 'precision medicine' approach based on the mechanisms of tumorigenesis, patients' demographic characteristics, and comorbidites. The genetic complexity, heterogeneity of physical tumors, DNA profiling, proteomic, RNA analysis, and immune mechanisms should to be taken into consideration. Accurate bioinformatic analysis is essential to optimize patient's treatment.

Precision medicine has significantly improved the diagnostic and therapeutic landscape of cancer. Successful implementation of precision medicine requires translational and bioinformatics infrastructure to support optimization of treatment selection. Several of these trials have been hindered by very low agreement between what is suggested and the actual therapy is prescribed. The commonly suggested reasons for this failure are because of the use of limited gene panels, the use of a restrictive matching algorithm, the lack of drug availability and limited clinician knowledge. Traditionally in medicine, drugs are designed and approved by testing on large populations. This design and approval process of results is only a portion of the population actually responding to a given therapy. However, due to different patients demographics, such as gender, age group and medical history, patients may respond differently to the same drugs. Targeted therapy alone or in combination with cytotoxic or other effective therapeutic strategies and innovative clinical trials with adaptive design should be offered to all patients.

Data sharing and N-of-1 clinical trials hold the promise to optimize the personalized treatment of individual patients and to expedite drug approval for rare alterations and tumor types. However, this presents a novel statistical problem, different from the double-blind clinical trial with thousands of patients, it is difficult to achieve statistical power. Because each patient in a trial is receiving a unique therapy, these therapies are not obviously comparable. In one study, the I-PREDICT trial [3] 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.

In order to facilitate implementation of precision medicine in the clinic, we have developed a computational system that can help to identify personalized cancer therapy for every cancer patient, given their unique DNA and RNA sequences. For each patient, our system identifies the best set of target genes and active hallmarks for the cancers, each of which have sets of available corresponding therapies associated with them. The goal of the system is to develop an N-of-1 treatment plan that could be initiated by the patient's physician under the auspices of a master protocol. Of note, the system served as an advisory tool, with the final decision made by the patient's physician. Therefore, the system reflects an experience with molecular profiling and patient treatment based on the historical data in the context of an academic medical center. Herein, we hypothesized that adherence to the system recommendations to match patients with targeted therapies was associated with higher degrees of matching and improved 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, the system will facilitate accurate utilization of data sequencing to perform algorithm analysis, identify the best set of target genes and active hallmarks, each of which have sets of available therapies associated with them. As noted previously, the system is playing the role as an advisor rather than a decision maker. In a clinical setting however, a clinician prescribes the therapy they believe is most appropriate for a given patient. Several of these trials have been hindered by very low matching rates, often in the 5–10% range, and low response rates. In this paper, we would like to evaluate of the level of agreement between our Aiomic therapies with those actually prescribed by oncologists through a set of similarity measures. We are interested in the evaluating the association between the adoption rate of Aiomic therapies and clinical measurements. This will provide us the evidence about the performance of the system and help to make improvements. We hypothesized that personalized treatment with combination therapies would improve outcomes in patients. However, some problems arise with this strategy these problems are broken 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, we are interested not only in just matching gene targets, but also evaluating the success of Aiomic therapies when given therapies are not an exact match. The problems can be stated as follows:

- How can we compare Aiomic therapies to prescribed therapies by the physician?
- Across all patients, can we say how often and to what degree the suggested Aiomic therapies were adopted?
- 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 only one drug is used in the given therapy, what percentage is attributable to the recommendation.
- Can we create a predicable model for partial adoption of our therapies?

3 Methods

3.1 Similarity measure: Jaccard similarity coefficient

Analysis of similarity of personalized cancer therapy identified by the system and the therapy clinician prescribed can help us understand the gap between "precision medicine" and the actual therapy that patients received. The presence and absence of drug lists are surveyed using clinical research, imaging, and other techniques. The Jaccard coefficient is one of the most fundamental measures for quantifying similarity using presence-absence data [7]. In section 3.1.1, we give a brief introduction of the Jaccard coefficient and we present a hypothesis test for similarity for presence-absence data, using Jaccard coefficient based on the bootstraps procedure [2] in section 3.1.2. The bootstraps procedure is considered in order to overcome the computational burden due to the high-dimensionality. We claim that, if the asymptotic distribution of similarity exist and is shown to be normal distribution, the present bootstrap hypothesis test can be used for any such similarity. Finally, we close the section by introducing the population hypothesis test using extreme value distribution [4].

3.1.1 Jaccard Similarity Coefficient for Presence-Absence Data

We consider a hypothetical data set in which two potential therapies are recommended for each patients, one is recommended by our system and the other is the one prescribed by the clinician. The two therapies are represented in the format of presence-absence vectors [5], which could be the list of the targeted gene or of the drugs recommended. Ideally, any meaningful similarity measure to match two therapies should display the following desirable properties:

- **Quantification:** Different similarity measures force different types of association, such as Pearson's correlation coefficient measuring the linear relationship between two variables. It is important to select the one that satisfies our request regarding the measure of overlap between the two variables and can be used to generate the interpretation.
- **Interpretations:** Thanks to machine learning techniques, we are able to design a specific algorithm that can be used to calculate similarity. Unfortunately, most of these techniques are black box computing, which is hard to interpret and lack meaning. Thus, the similarity measure could be used for the interpretation would be the preferred one, especially for clinical research.
- **Statistical guarantees:** Most similarity measures lack probability interpretations or statistical error control. Statistical properties have been inadequately studied. Thus, the develpoment of rigorous statistical test for evaluating the uncertainty similarity is necessary.

The similarity measure we consider here is the 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. 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 or drug. To explain the basic idea of the Jaccard coefficient, as an example, suppose we have 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 $A \cap B = \{1, 1, 1, 1\}$. The 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 closer the ratio is to 1, the more similar between the two sets are. The formula to find the Index is:

$$\text{Jaccard Index} = (\text{the number in both sets}) \ / \ (\text{the number in either set}) * 100$$

The formula in notation is:

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

The steps are:

1. Count the number of members which are shared between both sets.
2. Count the total number of members in both sets (shared and un-shared).
3. Divide the number of shared members (1) by the total number of members (2).
4. 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.

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

Statistical hypothesis testing using this similarity coefficient provides the confidence of the result. To evaluate whether A and B are independent, the following statistical hypothesis test 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 saying that the conventional Jaccard coefficient equals an expected value under independence. However, like most similarity coefficients, the Jaccard coefficient lacks probabilistic interpretations and statistical error control. Such problems could make results lack generalization and confidence. Another challenge is that it is difficult to achieve statistical power, because each patient in a trial is receiving a different therapy making these therapies comparable. This is the N-to-1 problem.

In order to utilize the Jaccard similarity coefficient, Chung, N. C. (2019) [1] 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 is shown to provide accurate p-values.

Proposition 1 *(Asymptotic property)[1] if A and B are independent then*

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

Here, based on Chung, N. C.'s work[1], we present a rigorous statistical test method to evaluate the similarity in presence-absence data, deriving statistical asymptotic properties and estimation of significance of the Jaccard coefficient.

Since the exact solution for a large m is computationally expensive and 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. It allows for estimation for p-values. 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 makes 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 the bootstrap procedure are (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 to avoid 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 bootstrap null coefficients
end
Compute the p-value by

$$p\text{-value} = \frac{\mathbf{1}\left\{|t_b^*| \geq |t|; b = 1, \ldots, B\right\}}{B}$$

3.1.3 Population Hypothesis Testing

In this section, we focus on the hypothesis test for group of patients. Assume we have a vector of Jaccard coefficient, J. We consider using the minimum extreme value distribution to evaluate the significance of Jaccard coefficient borrowing the idea from Rahman et al. (2014)[6]. 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 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 is 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 standard deviation. (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. The p-value is calculated as below:

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

3.2 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 indirect evidence, it provides the idea of accuracy of the the system and could be an improvement. Before we go into 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 under the inclusion and exclusion criteria, consent, end of study completion status, and missing data. All the inclusive and exclusive criteria 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 will be adjusted for multi-comparison if needed.

In order to adjust the validation caused 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.

3.2.1 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

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

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 consider that the variation is coming from the uniqueness of the patients.

3.2.2 Mixed Effect Model

To explain the mixed-effect model in an 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 is 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, $\text{var}\left(\beta_{i0}\right)$ and the variation among subject in the slopes $\text{var}\left(\beta_{i1}\right)$.

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

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

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.

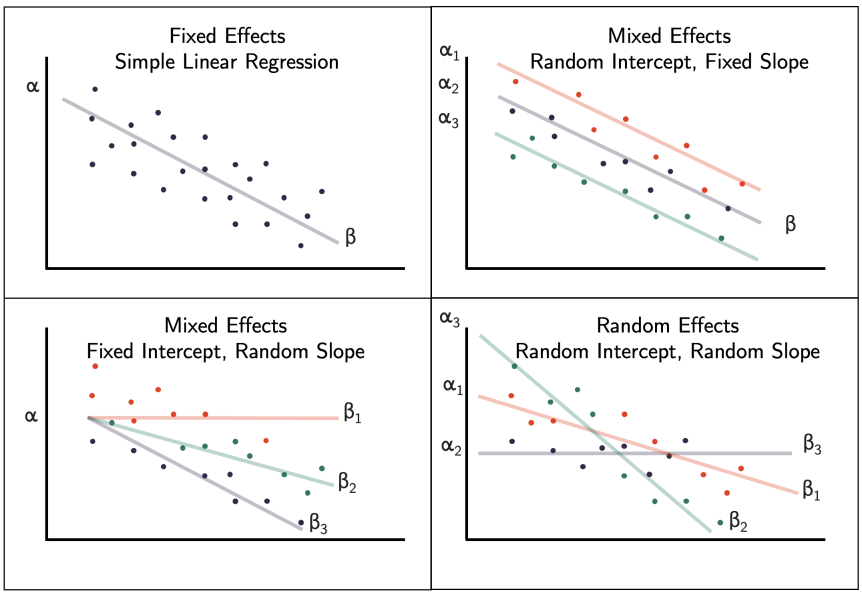

Figure from https://bookdown.org/steve midway/DAR/

3.2.3 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 were selected if they were statistically significant on the univariable level analysis and considered clinically significant by the research team.

Multivariable-Level Analysis helps to identify the risk factor and quantify their impact when they work together. All the risk factors that are significant could be considered as the main factor to be used in the system to predict the gene and help increase the accuracy. 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 framework for data analysis. Specifically, we propose using the Jacard coefficient as well as bootstrap procedure for statistical testing. Bootstrap has its advantage of lower computation cost, unfortunately, when the sample sizes are extremely small, there is no statistical guarantee and the result lacks of reliability. Such limitations should be taken into consideration when it comes to the real-data analysis and other methods to address this issue. Another challenge is that 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 the interactions. How can we incorporate this into our matching score? We leave these challenges for further exploration.

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

References

[1] Neo Christopher Chung, BłaŻej Miasojedow, Michał Startek, and Anna Gambin. Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data. *BMC bioinformatics*, 20(15):1–11, 2019.

[2] Nicholas I Fisher and Peter Hall. On bootstrap hypothesis testing. *Australian Journal of Statistics*, 32(2):177–190, 1990.

[3] Shumei Kato, Ki Hwan Kim, Hyo Jeong Lim, Amelie Boichard, Mina Nikanjam, Elizabeth Weihe, Dennis J Kuo, Ramez N Eskander, Aaron Goodman, Natalie Galanina, et al. Real-world data from a molecular tumor board demonstrates improved outcomes with a precision n-of-one strategy. *Nature communications*, 11(1):1–9, 2020.

[4] Samuel Kotz and Saralees Nadarajah. *Extreme value distributions: theory and applications*. World Scientific, 2000.

[5] János Podani and Dénes Schmera. A new conceptual and methodological framework for exploring and explaining pattern in presence–absence data. *Oikos*, 120(11):1625–1638, 2011.

[6] Syed Asad Rahman, Sergio Martinez Cuesta, Nicholas Furnham, Gemma L Holliday, and Janet M Thornton. Ec-blast: a tool to automatically search and compare enzyme reactions. *Nature methods*, 11(2):171–174, 2014.

[7] TT Tanimoto. An elementary mathematical theory of classification and prediction, ibm report (november, 1958), cited in: G. salton, automatic information organization and retrieval, 1968.