Octagon Health Data Science Competition

Robert Ciborowski
Computer Science
University of Toronto, St. George

Derek Wang
Computer Science
University of Toronto, Mississauga

Abstract

Data science techniques can be used in the medical field to provide useful insights on the effectiveness of treatments. The goal of this paper is to find how many Canadians use Sorafenib, a Tyrosine Kinase Inhibitor (TKI) drug, for hepatocellular carcinoma, and for how long. This paper explores a dataset of all Canadians who used Sorafenib between 2016 and 2019. The paper will also determine factors related to success and failure of the drug for each patient.

1. Dataset

The dataset of Canadians who used Sorafenib between 2016 and 2019 is provided to us by the Institute of Management and Innovation, University of Toronto Mississauga. We are interpreting the Measure-Tx rows of the dataset as the number of patients being treated in this current study and Measure-event as the number of patients that leave the study due to adverse events caused by the drug in question. We are also interpreting Measure-censored to represent interval censoring. That is, Measure-censored represents patients who finish the study successfully for a given month. In this paper, we will refer to the variable representing whether the patient is using another anti-cancer treatment concurrently as “Con_ACT”.

2. Number of Nine Month Patients

First, we will try to answer how many patients stay on treatment for at least 9 months. We analyzed the retention rate over the 39 month program by counting all patients who remained in the study during a given month. Our results can be seen in Figure 1.

Afterwards, we examined the 9th month to see how many patients remained in the study for at least 9 months. There were 704 patients out of the initial 1441 who remained in the trial. Thus, about 49% of patients left the trial after 9 months.

3. Predictors of Successful Therapy

Using the dataset, it is possible to examine which factors can be used to predict successful therapy. We define a successful therapy to be one where the patient recovers from liver cancer because of Sorafenib. The rows in the dataset under “Measure-censored” represent the patients who have been cured by the drug. The following features from the dataset which may affect the chance of a successful therapy are province, sex, age, and Con_ACT.

To find which of these features are the most correlated with a greater chance of successful therapy, one may use the SelectKBest algorithm, which determines the dependence between each factor in the dataset and the success percentage. To do this, we used the chi-square test along with scikit-learn's SelectKBest implementation. Afterwards, we found which values for the most correlated features provide the greatest chance of success. We used scikit-learn's LinearRegression model, which performs linear regression on the most correlated factors against the success rate. We also excluded any rows with the values "ALL", "UNKWN" or "Null" because their lack of information cannot help us find which factors are the most correlated with success.

	Column Name(s)	**5 Column Values With Greatest Success (in order from greatest)**
Most Correlated Column	Prov	QC, Prairies, ON, BC, Atlantic
2 Most Correlated Columns (in order of correlation)	Prov, Age	(QC, 18-19), (QC, 20-24), (Prairies, 18-19), (QC, 25-29), (Prairies, 20-24)
3 Most Correlated Columns (in order of correlation)	Prov, Age, Con_ACT	(QC, 18-19, No), (QC, No, 20-24), (Prairies, No, 18-19), (QC, No, 25-29), (Prairies, No, 20-24)

Table 1. The most correlated columns to success.
Prov = Province, QC = Quebec, ON = Ontario, BC = British Columbia

Our results conclude that location is the most important factor of success. Quebec patients are the most likely to see Sorafenib succeed. Age is the second most important factor of success. Younger patients who are between the ages of 18 and 24 are the most likely to see success. Patients who do not take another anti-cancer therapy are also more likely to see Sorafenib succeed. Sex is not as correlated with successful treatment.

4. Monthly Rate of Discontinuation

The values in the dataset under "Measure-event" represent how many patients experienced adverse events, which caused them to discontinue the use of the drug. These patients have not seen success in using the drug because they ended its use prematurely. Thus, we will use the "Measure-events" data to observe this discontinuation rate.

The bars labelled “Events” under Figure 2 show the number of discontinuations for each month. Figure 3 shows the percentage of patients out of the total remaining patients who discontinued their use of Sorafenib in a given month.

As shown by Figure 3, the majority of patients who discontinued their use of the drug do so in the earlier months. As time passes, patients’ discontinuation numbers drop, but due to the lower number of people left, the discontinuation rate decreases slowly. In the final few months, no patients discontinued their use of the drug. The average monthly discontinuation rate is about 2.6%.

5. Reasons of Discontinuation

Using a similar method to Section 3, one can determine some potential reasons for discontinuation of the therapy. We also used scikit-learn’s implementation of SelectKBest and LinearRegression. Using this method, we found which factors were the most correlated with high average monthly discontinuation:

	Column Name(s)	**5 Column Values With Greatest Average Discontinuation (in order from greatest)**
Most Correlated Column	Prov	AB, Atlantic, BC, ON, Prairies
2 Most Correlated Columns (in order of correlation)	Prov, Age	(AB, 65+), (AB, 60-64), (Atlantic, 65+), (AB, 55-59), (Atlantic, 60-64)
3 Most Correlated Columns (in order of correlation)	Prov, Age, Con_ACT	(AB, 65+, Yes), (AB, 60-64, Yes), (Atlantic, Yes, 65+), (AB, 55-59, Yes), (Atlantic, 60-64, Yes)

Table 2. The most correlated columns to success.
Prov = Province, AB = Alberta, ON = Ontario, BC = British Columbia

Table 2 shows that location is the most correlated with discontinuation. The province of Alberta has the greatest probability of stopping their use of the drug due to an adverse event. Age is also correlated with discontinuation. Older patients have a greater probability of stopping their use of the drug prematurely. Patients who take another anti-cancer therapy are also more likely to discontinue. Sex is not as correlated with discontinuation. From this information, one can conclude that residing in Alberta or one of the Atlantic provinces increases a patient’s chance of experiencing an adverse event with Sorafenib. Also, an older age and the use of other anti-cancer therapies are both causes of discontinuation.

Appendix

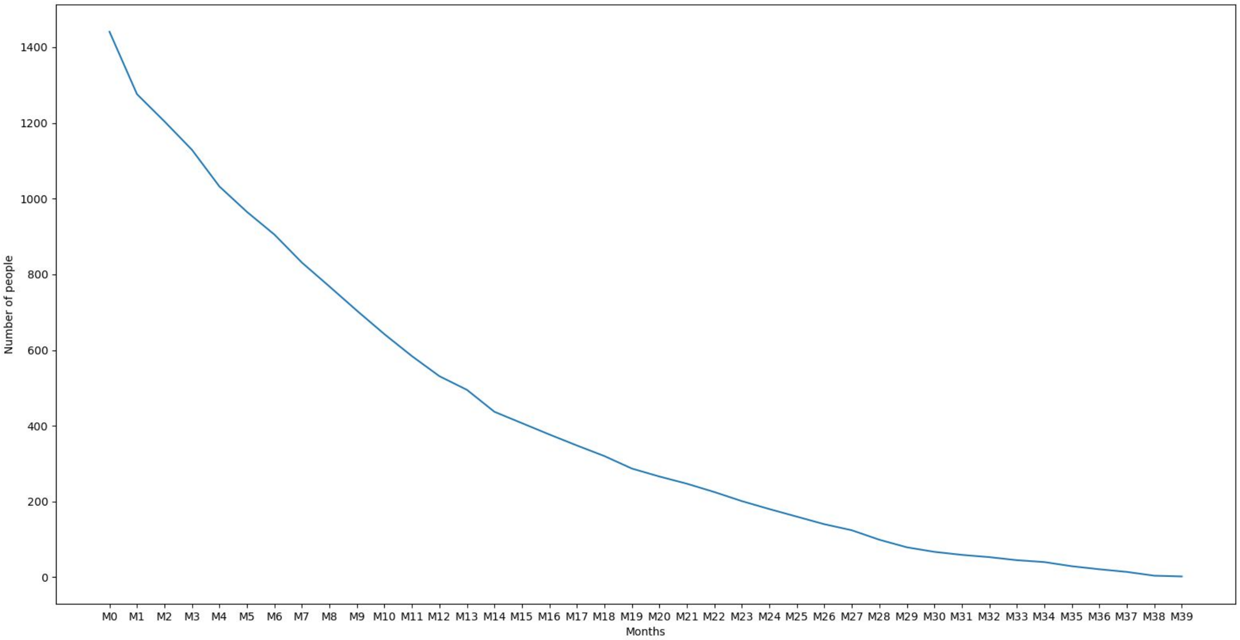

Figure 1. The number of people continuing treatment after each month.

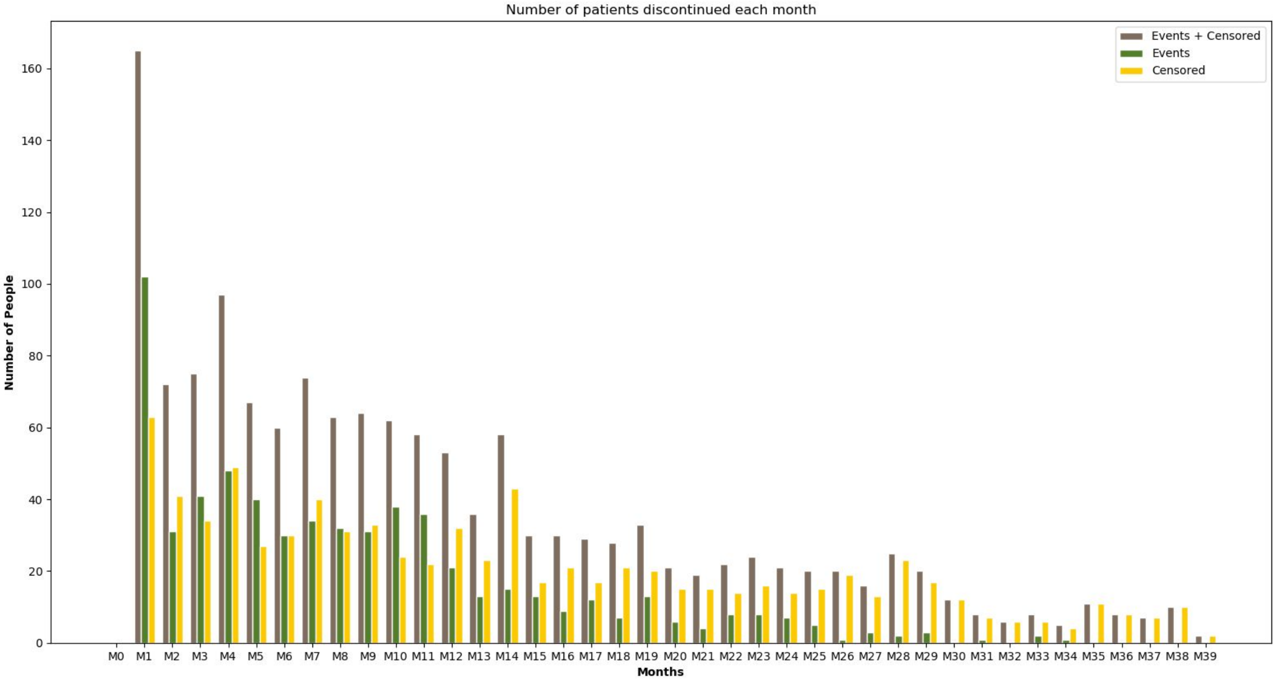

Figure 2. The number of people who discontinued the treatment (events) or succeeded in the treatment (censored) for a given month.

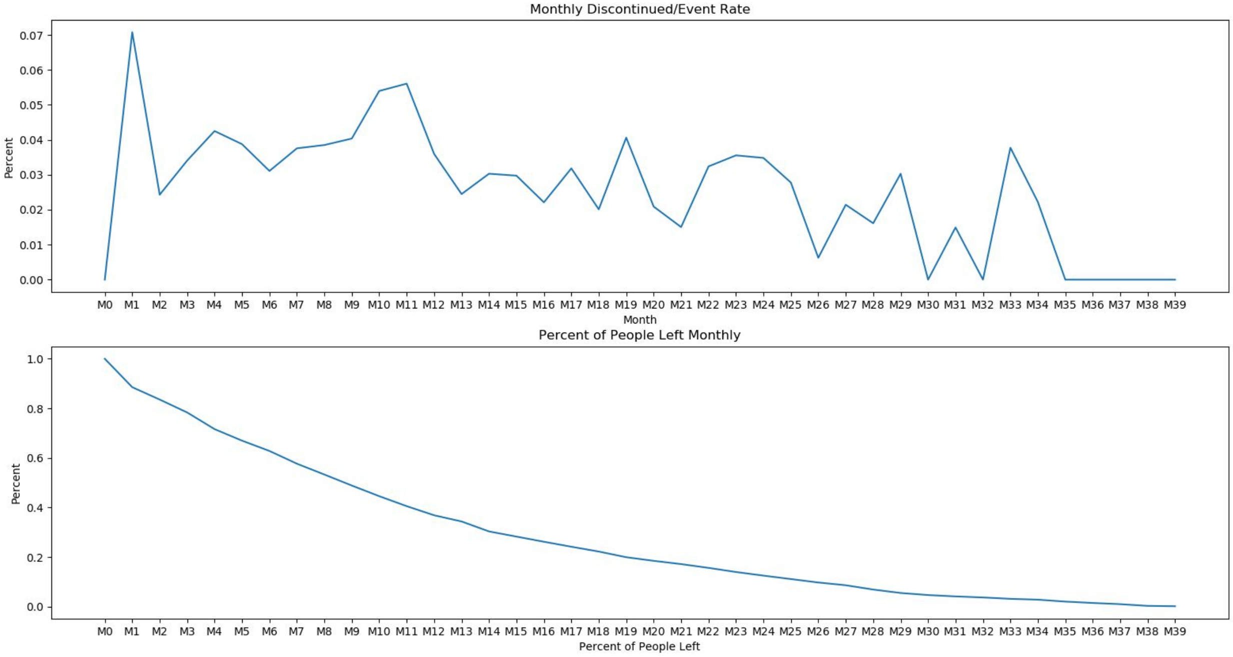

Figure 3. Each month’s discontinuation rate versus the number of people using the drug that month.

Our definitions for Measure-event and Measure-censored were obtained from https://www.quantics.co.uk/blog/introduction-survival-analysis-clinical-trials/ and https://www.fda.gov/patients/clinical-trials-what-patients-need-know/glossary-terms#E-1.

More information about the programming portion of this project can be found here.