Analyzing Patient Symptoms and Treatment Effectiveness of Lung Cancer Drug Therapies with Machine Learning Algorithms

Victoria Yu

Abstract

Cancer patients often experience various severe side effects from cancer treatment. Specifically, lung cancer patients commonly experience Cisplatin-induced nephrotoxicity–deterioration in kidney function due to toxic effects of the chemotherapy drug Cisplatin. However, it is difficult to predict how a patient will respond to Cisplatin treatment, as patients respond differently depending on their unique clinical-demographic features. This project analyzes data for lung cancer patients undergoing Cisplatin treatment by using machine learning methods to determine the most important clinical features that affect nephrotoxicity, and to predict the probability of a patient experiencing toxic symptoms given their characteristics. Three different clinically relevant patient groups were used, and age, dose given, and pre-treatment GFR were found to be positively correlated and the most relevant features for determining treatment effectiveness and possibility of nephrotoxicity. This information can help doctors and patients predict treatment response and optimize treatment plans.

Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death in Canada. Cisplatin, a type of chemotherapy drug, is considered the front-line treatment for patients with advanced, non-squamous, non-small-cell lung cancer. Therapeutic effects of Cisplatin are improved by increasing dose, but high-dose therapy leads to nephrotoxicity–rapid deterioration in the kidney function due to toxic effects of medications and chemicals–and possibly acute kidney injury, which leads to extreme swelling, shortness of breath, nausea, and more.

Notably, it is often difficult to predict how a patient will respond to Cisplatin treatment, as a patient's characteristics, such as their age, BMI, sex, blood pressure, and more all play a role, meaning each patient and the side effects they experience are unique. However, understanding the factors that contribute to Cisplatin induced nephrotoxicity in patients is crucial to optimize treatment plans.

Thus, in this experiment, I analyzed patient data and its effects on nephrotoxicity and glomerular filtration rate (GFR) using data from a study conducted by C. Máthé et al. The study compares Cp-induced nephrotoxicity in 3 different groups: no comorbidity (NC), hypertension plus ischaemic heart disease (CD), and diabetes and ischaemic heart disease (DMIH). I implemented Random Forest classifier and regression Machine Learning algorithms to determine the treatment effectiveness for each patient and their probability of experiencing severe nephrotoxic effects depending on their clinical-demographic features. Given this algorithm, I explored the most important clinical features that determine whether a patient is likely to experience Cp-induced nephrotoxicity.

Current research shows possible correlations between age, sex, and smoking as risk factors for experiencing Cisplatin-induced nephrotoxicity. However, there is currently no method of predicting how a patient will respond to treatment according to their unique clinical-demographic features.

Study Design

The study compares Cp-induced nephrotoxicity in three patient groups: no comorbidity (NC) (n=80), hypertension plus ischaemic heart disease (CD) (n=110), and diabetes and ischaemic heart disease (DMIH) (n=52). I used important feature variables from the publicly available data, and as specific individual patient data was unable to be obtained due to personal health information restrictions, I populated a larger dataset assuming a Gaussian distribution for each feature group. This data was then visualized graphically and compared between the NC, CD, and DMIH groups.

Methodology

I performed a correlation analysis to determine the most important feature variables for determining treatment effectiveness that was common to all three patient groups. Knowing these variables, I trained a regression model for each patient group (NC, CD, DMIH) to predict treatment effectiveness, and to account for the additional comorbidities present in the CD and DMIH groups that make them more susceptible to experiencing nephrotoxicity. Furthermore, I trained a classification model to determine if patients were likely to experience nephrotoxicity, which was defined by a 10% or more decrease in glomerular filtration rate (GFR) after Cisplatin treatment. A 10% or more decrease would cause a patient's GFR to be lower than 60 ml/min per 1.73 m^2, which causes them to be definitionally at risk for moderate chronic renal disease and nephrotoxicity. The regression and classifier machine learning algorithms are random forest models–supervised machine learning algorithms that combine the output of multiple decision trees to make a more accurate overall prediction.

	Age	BMI	Pre_Cp_GFR	Post_Cp_GFR	Dose_per_treatment	Total_dose	Number_Cycles	Systolic_BP	Diastolic_BP	Cardiac_Frequency	Group
0	51.540475	23.863450	100.781451	82.833759	82.944284	321.691057	3.553522	114.941980	79.713461	74.856965	NC
1	55.888848	25.822634	101.276571	92.012412	81.553877	318.109727	2.813919	120.537366	74.559078	72.722759	NC
2	53.840836	21.701588	99.289548	87.630903	81.333528	317.084763	3.310616	120.620886	75.579108	67.538334	NC
3	54.368493	21.662195	101.326272	87.896178	81.447351	319.926941	5.355136	127.875788	81.133563	73.292381	NC
4	54.121306	27.176767	101.459675	90.961832	78.525647	324.692705	4.095209	117.573830	76.201441	73.077028	NC
5	51.311160	24.416514	95.120755	88.482411	80.106205	316.717366	4.600999	118.740757	81.267118	68.293687	NC
6	55.723071	25.002973	100.973811	94.388442	78.068429	324.855715	3.922170	122.807926	79.895958	74.730508	NC
7	58.344712	24.613926	103.384382	87.985302	76.969067	315.870592	2.703968	123.000956	83.258313	71.395656	NC
8	56.209765	25.834263	101.149947	86.525631	83.834255	317.461334	4.397857	124.347361	82.647314	68.204084	NC
9	53.851960	23.456916	96.316068	87.181547	75.133954	316.769809	3.539014	119.786307	76.057017	71.193261	NC
10	55.324405	25.205552	97.060142	88.057502	80.118291	322.659689	4.038968	116.929359	82.041118	72.505401	NC
11	56.079268	22.583737	94.080439	86.125384	78.864068	322.107481	3.668228	114.271558	76.025372	64.354227	NC
12	52.903500	25.693993	100.162264	90.561572	81.903678	313.428716	4.007684	120.210010	80.582551	70.207995	NC
13	58.082250	24.780562	98.724105	87.873161	82.724793	331.330829	4.086127	121.886679	84.513597	65.245226	NC
14	56.786162	22.261616	95.319472	92.430475	78.987169	328.707463	3.550772	120.897588	81.746931	67.565147	NC

Fig. 1: Sample of data generated, using age, BMI, pre- and post-Cp GFR, dose per treatment, total dose, number of cycles, blood pressure, cardiac frequency, and patient group (NC, CD, or DMIH) as feature variables.

Results

Based on the available data from the study, each feature was given as a mean with a standard error of mean (SEM). Therefore, the best method of populating a larger dataset was assuming a Gaussian distribution for each feature. By visualizing this data graphically, it was observed that between the NC, CD, and DMIH groups, age, BMI, blood pressure, and cardiac frequency beats per minute increased as the number of comorbidities increased.

To work with a smaller set of features, a correlation analysis was performed and found that age, pre-Cp GFR, and total dose given were the most important features that affected toxicity on an overall level, common to all three patient groups. A correlation coefficient ranges from -1 to 1, and the coefficients for age, pre-Cp GFR, and total dose were all above 0.8, suggesting a strong positive relationship to nephrotoxicity.

Knowing that age, pre-Cp GFR, and total dose were the three major features that affect treatment effectiveness, I trained a regression and classifier model to further analyze how these features vary in importance in the three patient groups with different comorbidities and baseline characteristics. In the NC (no comorbidity) group, the age of the patient was the most important factor (72.3%), followed by the dose they received (27.5%). In the CD group, age (36.4%), total dose (32.1%), and pre-Cp GFR (31.1%) were of relatively equal importance for determining treatment effectiveness. However, in the DMIH group, the patient's pre-Cp GFR was significantly the most important (97.4%), followed by their age (2.5%). In all groups, a patient's pre-Cp GFR was the most important factor that determined whether or not they would experience nephrotoxicity.

There are some further important trends to note. With the dataset generated, it was observed that patients in the CD and DMIH groups could tolerate less Cisplatin dose, had a greater nephrotoxicity frequency, and were more likely to drop out of treatment due to being too sick. This aligns with the previously discussed results, showing how age, pre-Cp GFR, and dose are positively correlated with nephrotoxicity. Furthermore, the trained regression model proves to be very accurate, as the R^2 value–coefficient of determination–is very high for both the training and test data, showing a good fit. The accuracy to which the classification model correctly predicts whether a patient is likely to experience severe side effects is also high: 73% for the NC group, 74% for the CD group, and 79% of the time for the DMIH group.

Lastly, a confusion matrix and receiver operating characteristic (ROC) curve was created to further test the performance of the classification model. A confusion matrix compares the predicted values of the algorithm to the true values, while an ROC curve plots the true positive rate of the classification model against the false positive rate at all classification thresholds. Both evaluation tools showed good performance of the classification model.

Discussion and Conclusions

There are two main conclusions that can be drawn from the results.

First, my data analysis provides a better understanding of the important clinical-demographic features that influence treatment effectiveness and nephrotoxicity from Cisplatin treatment,

especially within patients with additional comorbidities such as cardiovascular disease and diabetes. Thus, if a patient was older, had a lower pre-Cp GFR, and more comorbidities, the patient will likely have reduced tolerance to Cisplatin treatment and should be given a lower dose to optimize treatment.

Second, with the relative accuracy of the trained regression and classification models, medical professionals can have more confidence predicting how a patient will likely respond to treatment. By advancing the field of predictive modeling in oncology and medicine, the machine learning models I created provide clinical professionals with more information about individual patient risk and toxicity profiles, allowing them to make more clinically informed decisions that optimize treatment plans to mitigate adverse side effects.

In the future, the methodology used in this project can be extended to predict nephrotoxicity progression overtime, or consider other side effects of Cisplatin treatment such as nausea and infection. I also look forward to improving the performance of the model I created, such as training the model on a larger dataset, which was difficult to do due to limitations of personal health data accessibility, which should be taken into account.

References

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random Forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics, 19(1).

Gao, Y., Dorn, P., Liu, S., Deng, H., Hall, S. R., Peng, R.-W., Schmid, R. A., & Marti, T. M. (2019). Cisplatin-resistant A549 non-small cell lung cancer cells can be identified by increased mitochondrial mass and are sensitive to pemetrexed treatment. Cancer Cell International, 19(1).

McSweeney, K. R., Gadanec, L. K., Qaradakhi, T., Ali, B. A., Zulli, A., & Apostolopoulos, V. (2021). Mechanisms of cisplatin-induced acute kidney injury: Pathological mechanisms, pharmacological interventions, and genetic mitigations. Cancers, 13(7), 1572.

Miller, R. P., Tadagavadi, R. K., Ramesh, G., & Reeves, W. B. (2010). Mechanisms of cisplatin nephrotoxicity. Toxins, 2(11), 2490–2518.

Máthé, C., Bohács, A., Duffek, L., Lukácsovits, J., Komlosi, Z. I., Szondy, K., Horváth, I., Müller, V., & Losonczy, G. (2010). Cisplatin nephrotoxicity aggravated by cardiovascular disease and diabetes in lung cancer patients. European Respiratory Journal, 37(4), 888–894.

Statistics Canada. (2022, January 4). Lung cancer is the leading cause of cancer death in Canada.
Retrieved from https://www.statcan.gc.ca/o1/en/plus/238-lung-cancer-leading-cause-cancer-death-canada

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43(6), 1947–1958.

Thakwani, J. (2019, January 3). Symptoms and Signs of Nephrotoxicity. Medindia. Retrieved from https://www.medindia.net/health/drugs/symptoms-and-signs-of-nephrotoxicity.htm