A Warning System for Cytopenias during Chemotherapy using Population-Based Administrative Data

Jiang Chen He
1002348177
jiang.he@mail.utoronto.ca

Supervisors: Rob Grant, Kelvin Chan, Marzyeh Ghassemi

August 2021

Abstract

Introduction: Chemotherapy patients are at increased risk of cytopenias, a condition in which blood count levels are dangerously low. We aimed to assess the utility and feasibility of Machine Learning (ML) algorithms to identify cytopenias in cancer patients prior to their upcoming chemotherapy sessions.

Methods: Four ML algorithms to identify cytopenias were trained using 51 variables from population-based administrative data consisting of data on demographics and cancer diagnosis and treatment, symptom questionnaires, and blood work for 32,567 patients between 2007 and 2020. The target of the models were grade 1 or higher anemia, thrombocytopenia, and neutropenia.

Results: Among the 32,567 patients, 17,591 were female and 14,976 were male, with mean age of 64. Before their next chemotherapy session was due, 2637 had thrombocytopenia, 9360 had neutropenia, and 9838 had anemia. Among the 4 ML algorithms, the Extreme Gradient Boosting Tree performed the best with an area under receiver operating characteristic curve score of 86% (neutropenia), 97% (anemia), and 96% (thrombocytopenia) in the test cohort. The top ten predictors for each blood type included the baseline blood count, chemotherapy regimen, cancer location, and cancer type.

Discussion: ML methods can predict cytopenias in chemotherapy patients before their next scheduled chemotherapy administration, which can aid in early intervention and facilitate tailored treatment.

Keywords: machine learning, cytopenia, neutropenia, anemia, thrombocytopenia, assessment, chemotherapy

Introduction

Chemotherapy is a common cancer treatment that utilizes powerful drugs to destroy cancerous cells in a patient's body. Often these same drugs also damage the patient's bone marrow, where blood cells are produced, leading to reduced blood cell counts. Deficiencies in blood cells, a condition called cytopenia, can cause many complications. For example low neutrophil count (neutropenia) leads to greater risk of infection; low hemoglobin count (anemia) leads to fatigue and shortness of breath; low platelet count (thrombocytopenia) can lead to excessive bleeding. As such, knowing which patients are at risk of cytopenias before administering chemotherapy is important for medical oncologists, who can mitigate the risk through dose modification or growth factors.

In this study, our goal was to develop a machine learning system that predicts neutropenia, anemia, and thrombocytopenia for patients on chemotherapy, enabling more tailored treatments.

Background

With recent advances in machine learning techniques and computing power, numerous papers on predictive modelling in chemotherapy have been published. Researchers have employed machine learning algorithms to predict the existence or absence of chemotherapy-induced neutropenia [1][2][3]. The researchers evaluated the performances of various linear and non-linear models and demonstrated an area under the receiver operating characteristic curve (AUROC) scores ranging from 73% to 91%. Another study compared various machine learning techniques for predicting in-hospital mortality of patients with febrile neutropenia using a dataset of 126,000 patients, 4.6% of them declared as deceased, with a reported AUROC of 92% [4]. In 2017, Palowski et al. presented a neutropenia risk prediction model using a dataset of electronic health records (EHR) of chemotherapy treatments for 126 patients [5]. The authors reported an AUROC score of 75%. In 2020, Culpov and Andre developed a machine learning model to forecast chemotherapy-induced hematological toxicity levels as functions of predicted neutrophil and platelet levels, achieving a mean absolute error of 1.0 and 72.8 G/L respectively with a dataset of 24 patients [6]. Avdic et. al in [7] applied an artificial neural network to predict neutrophil count for childhood cancer patients, achieving a prediction error of 290microL.

Multivariate statistical approaches have also been conducted to predict risk of chemotherapy toxicity [8][9], including chemotherapy-induced neutropenia [10] and myelosuppression [11]. However, these methods are limited by their inability to detect complex multivariable interactions. Machine learning is a more scalable technique that is able to find patterns in high-dimensional datasets. As a result, it can yield higher predictive accuracy and generalize to a higher degree than statistical methods.

To date, all studies suffer from shortcomings, including small sample sizes, limited availability of predictive features, a focus on a subset of cancer patients such as only on breast cancer patients, and the lack of external validation.

In this study, we developed a warning system for cytopenias on chemotherapy by applying machine learning models to population based administrative data.

Methodology

Dataset

The dataset used in this study was provided by ICES, a non-profit independent research institute in Ontario, Canada [12]. They collected population based administrative data from multiple sources, such as Cancer Care Ontario (CCO) and Canadian Institute of Health Information (CIHI). These data were linked using unique identifiers and analyzed by ICES. The analysis cohorts included questionnaires, chemotherapy administration data, blood work, demographic, and clinical data.

Preprocessing Pipeline

The first stage of the preprocessing pipeline involved extracting variables from the different analysis cohorts. The variables selected include information from 1) symptom questionnaires, such as level of anxiety, pain, depression, etc. from 0 to 9. 2) demographic and clinical data, such as age, sex, body surface area, ECOG performance status [13], etc. 3) chemotherapy diagnosis and treatment data, including chemotherapy regimen, cancer type, cancer location, etc. and 4) blood measurement data, including baseline leukocyte count, monocyte count, etc. Appendix Table I shows the full set of variables used.

Each example to train and test the system was a chemotherapy administration. We extracted and forward filled blood measurements from five days prior to the day of chemotherapy administration. We used the forward-filled measurement on day of chemotherapy administration as the baseline blood counts. The target blood count was taken as the forward-filled measurement from the day before to the day after the next chemotherapy administration was due. Figure 1 gives an example of how the baseline and target blood measurements were defined for a regimen with 14-day cycle. Examples without baseline or target measurement of hemoglobin, platelets and neutrophils were excluded. A patient was classified as having a cytopenia if they had a grade 1 event using the Common Terminology Criteria for Adverse Events [14]. Specifically, low platelet count is defined as below 75 x 10e9 / L, low neutrophil count as below 1.5 x 10e9 / L, and low hemoglobin count as below 100 g / L.

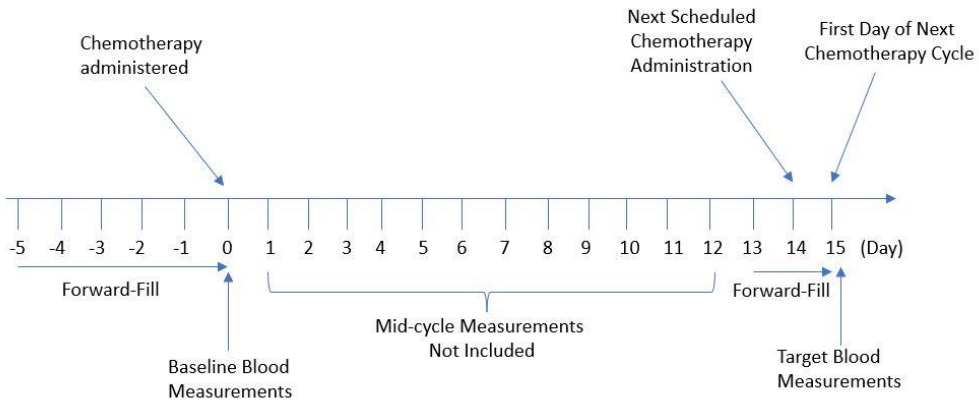

Figure 1: Definition of Baseline and Target Blood Measurement for 14-day Cycle Regimen

Our intent was to develop a system to predict future blood counts at the clinically relevant time points for oncologists, which is typically when the next chemotherapy is due. The number of days until the next chemotherapy administration is due for each regimen was taken from the Cancer Care Ontario Formulary [15]. Given the administrative dataset does not include where patients are in their chemotherapy cycle, where chemotherapy regimens are planned for administration at different intervals, we used the shortest interval between administrations. Where chemotherapy is given on consecutive days, we extracted the target based on when a clinician would typically order the next blood test.

The next stage involved cleaning and filtering the feature data, which included removing erroneous entries (i.e. body surface area of 0 or -99); replacing categorical entries with less than six appearances as 'Other' in accordance with ICES privacy policies; merging small chemotherapy intervals together (i.e. four consecutive days of chemotherapy administration will be merged as a 4-day cycle); and removing chemotherapy sessions where no blood measurements were collected. The patients did not fill out the questionnaire form at every chemotherapy session. Thus we decided to forward-fill any missing questionnaires using responses from previous sessions. Finally, we one-hot encoded the categorical variables, converting each categorical entry into a binary variable, which increased the total number of variables from 51 to 665. A diagram of the overall preprocessing pipeline is shown in Figure 2.

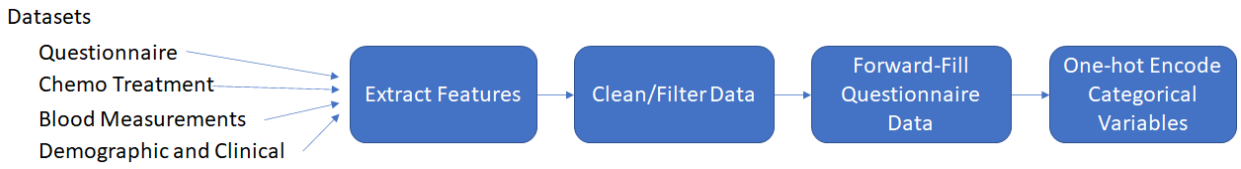

Figure 2: Flowchart of Preprocessing Stages

Training Pipeline

The dataset was split into 60% training set, 20% validation set, and 20% testing set based on patients. Any missing values for numeric variables were replaced with the mean and categorical variables were replaced with the mode. For body surface area, missing data were imputed based on the mean for each sex. The data was then normalized, and the minority examples were oversampled to better balance the data. Specifically, the examples of low platelet counts were increased by 5 folds. The target distribution before and after resampling can be seen in Appendix Table II.

The average precision (AP) score was used to evaluate the models during training. AP score summarizes the precision-recall (PR) curve as the weighted mean of precisions achieved at each threshold. The PR curve was chosen as it focuses more on the positive (minority) class, making it well-suited for an imbalanced dataset.

After training, the models were calibrated such that its predictive probability distribution became closer to the empirical probability distribution of cytopenia occurring using Platt scaling [16]. The best hyperparameters were determined by conducting Bayesian optimization [17]. The models with the best hyperparameters were evaluated. A diagram of the overall training pipeline is shown in Figure 3. All implementations were done using open-source software in Python 3.6 and the sklearn module. The code for reproducing the results is available at https://github.com/KevinJCHe/cytopenia-detector.

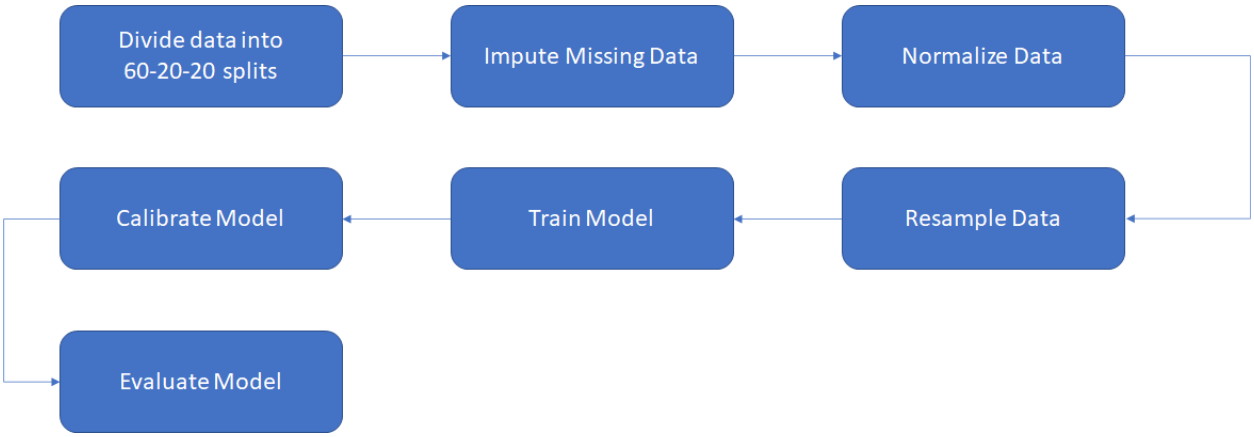

Figure 3: Flowchart of Training Stages

Machine Learning Models

We compared the performance of four different linear and non-linear algorithms, which are 1) logistic regression (LR), with L2 regularization to combat overfitting, 2) random forest (RF), with minimum number of samples required to be at a leaf node at six samples in accordance with ICES privacy policies, 3) extreme gradient boosting (XGB) tree, also with minimum leaf node samples of six, and 4) neural network (NN), with three hidden layers, ReLu activation, and

trained with an Adam optimizer. The models were required to output three predictions for each blood type, but only the NN supports multi output classification from the sklearn module. For LR, RF, and XGB, each model was fitted three times, once for each output.

Results

The overall dataset contained more than 1800 chemotherapy regimens. However, each regimen had its own chemotherapy administration intervals that required manual lookup. At the time of this report, the clinical team had only extracted the chemotherapy administration intervals for the most common 33 chemotherapy regimens, but we plan to repeat this analysis with the complete dataset. The overall dataset had 952,834 chemotherapy patients in total. After excluding patients not under one of the 33 regimens, the number dropped to 137,287 patients. Further processing as described above resulted in a final sample size of 32,567 patients completing 126,183 chemotherapy sessions spanning from 2007 to 2020.

Table 1 provides descriptive statistics of the study population. For statistics on cancer location and chemotherapy regimen, only the top 5 most common are shown. Out of 32,567 patients, 17,591 (54%) were female, 2637 (8.1%) had thrombocytopenia, 9360 (28.7%) had neutropenia, and 9838 (30.2%) had anemia right before their next scheduled chemotherapy administration, and the average age in the population was 64.

Table 1. Statistics of Study Sample.

Patient Characteristics	**Population**
Total N(%)	32567 (100%)
Average Age [IQR]	64 [57 - 73]
Sex N(%)	
Female	17591 (54%)
Low Blood Count N(%)	
Platelet Count < 75 x 10e9 / L	2637 (8.1%)
Neutrophil Count < 1.5 x 10e9 / L	9360 (28.7%)
Hemoglobin Count < 100 g / L	9838 (30.2%)
Cancer Location N(%)	
Bone Marrow	2127 (6.53%)
Rectal	2017 (6.19%)
Lung	1900 (5.83%)

Lymph Node	1862 (5.72%)
Ovary	1756 (5.39%)
Chemotherapy Regimen N(%)	
mfolfox6	5229 (16.06%)
crbppacl	2967 (9.11%)
folfiri+beva	2751 (8.45%)
chop+r	2569 (7.89%)
cape	2013 (6.18%)

The ROC curve and PR curve for each blood type and models based on the test set can be seen in Figure 4. A detailed comparison of the models with reported AUROC score, AP score, F1 score, precision, recall, and accuracy at cutoff-point of 0.5 for the test set is presented in Table 2. The best values are highlighted in bold.

All algorithms achieved a similar level of accuracy, and each algorithm performed better on precision or recall on specific blood types, however XGB achieved the overall best with AUROC of 86%, 97%, and 96%, and AP of 54%, 85%, and 59% for neutrophil, hemoglobin, and platelet, respectively. The hyperparameters used to achieve these results are provided in Appendix Table III.

The XGB algorithm provided the importance scores of each variables based on the average gain over all splits where the variable was used. The top ten important predictors were similar for all 3 blood types, which included the baseline blood count, chemotherapy regimen, cancer location, cancer type, chemotherapy cycle duration, and region of Ontario. Appendix Table IV shows the top ten predictors for each blood type.

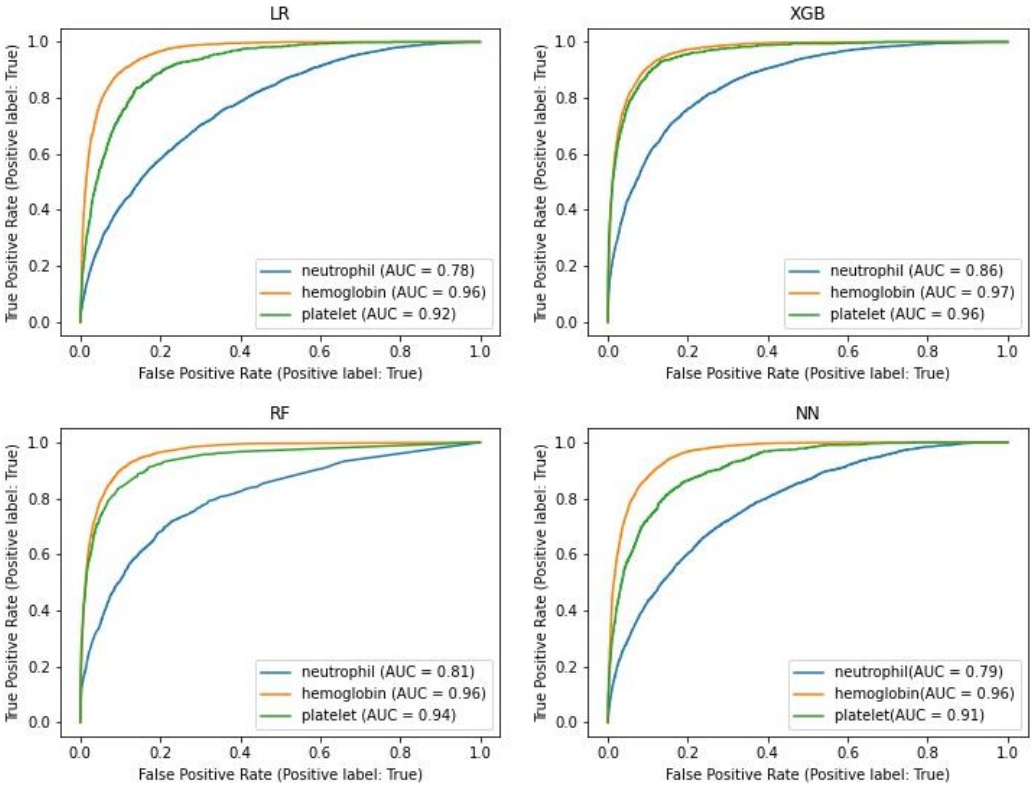

Figure 4A: Receiver Operating Characteristic Curves for Model Comparison

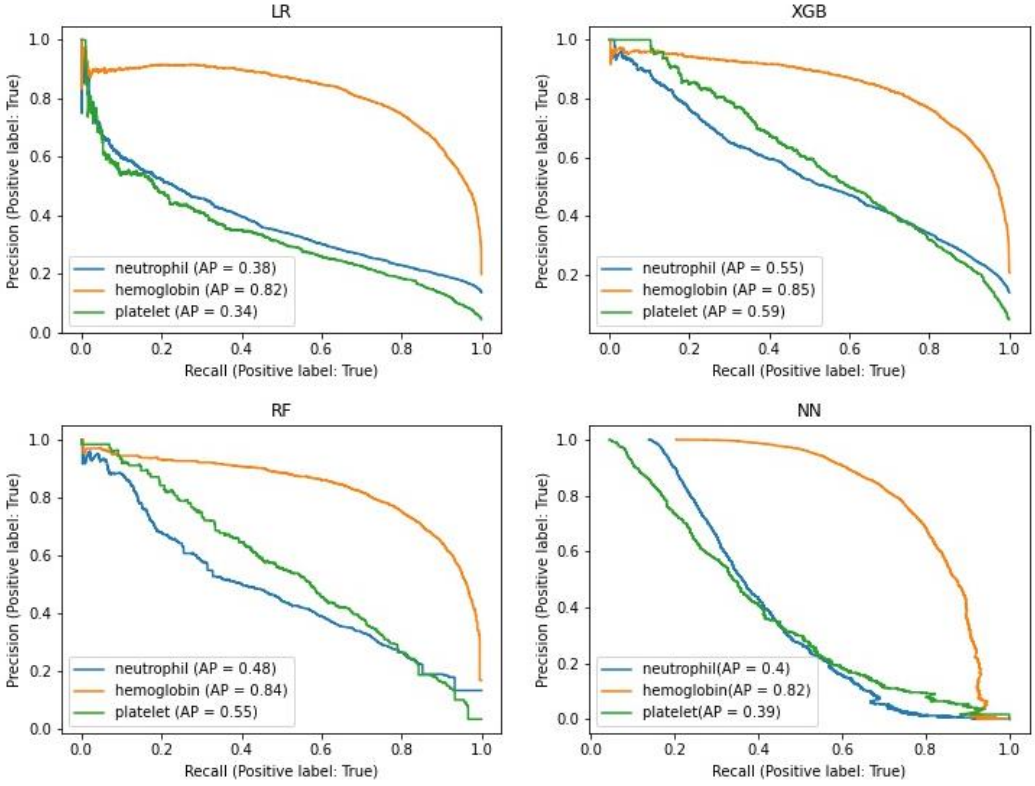

Figure 4B: Precision-Recall Curves for Model Comparison

Table 2. Model Performance Comparison

Models		Accuracy	Precision	Recall	F1	AUROC	AP
Logistic Regression	Neutrophil	87.05	62.14	8.27	14.60	77.58	38.00
	Hemoglobin	91.95	**80.84**	68.51	74.17	96.13	81.99
	Platelet	95.85	38.39	32.76	35.35	91.81	33.56
Random Forest	Neutrophil	88.07	**75.84**	16.01	26.44	80.51	47.86
	Hemoglobin	92.33	78.83	74.51	76.61	96.14	83.63
	Platelet	95.83	43.08	**63.46**	51.32	93.69	55.19
Extreme Gradient Boosting Tree	Neutrophil	**88.59**	69.66	26.21	**38.09**	**86.22**	**54.87**
	Hemoglobin	**92.63**	79.19	**76.39**	**77.77**	**96.55**	**84.81**
	Platelet	**96.37**	**48.20**	63.00	**54.62**	**95.79**	**59.19**
Neural Network	Neutrophil	86.07	46.84	**30.30**	36.80	78.66	39.88
	Hemoglobin	91.80	75.61	75.81	75.71	95.91	81.81
	Platelet	94.70	33.11	52.00	40.46	91.31	38.71

Discussion

The results show that machine learning models, especially XGB, are able to utilize population-based administrative data to predict cytopenias for chemotherapy patients with high AUC and good AP score. The large, rich dataset provided by ICES enabled us to utilize more information than many of the previous studies, allowing our machine learning models to find meaningful patterns among the large number of unique variables. This can be used for early detection systems of cytopenias, which can help improve patient care.

Strengths

Many cytopenia prediction systems use traditional linear statistical approaches. We developed 4 different linear and non-linear machine learning models that can achieve high predictive scores with greater scalability and robustness. We also employed Bayesian optimization to search for the best combination of hyperparameters to mitigate overfitting and increase model performance. We also calibrated the models such that the predictive probability distribution resembles the empirical probability distribution, demonstrated by the calibration plots in Figure 5, also enhancing model performance.

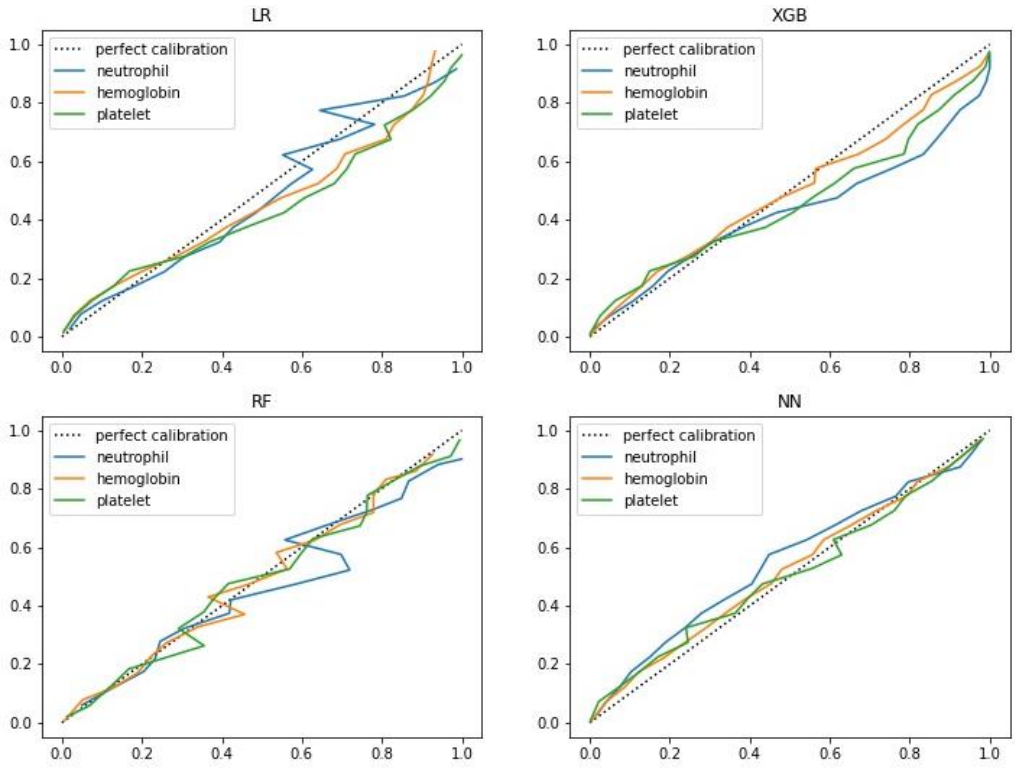

Figure 5: Calibration Plots

Limitations

There are several limitations to this study. With regards to the dataset, we did not include the different types of drugs issued and their dosages for each chemotherapy administration. Although the chemotherapy regimen encompasses what chemotherapy drugs are given and their applied dosages, there are examples where patients were given additional or different drugs with modified dosages. There are thousands of drug identification numbers in the dataset that we would have to manually look up, and due to limited human resources and time we decided to exclude these features. In addition, we did not include whether patients received blood transfusions or agents that stimulate blood growth during the chemotherapy sessions. The dataset is also imbalanced, which can lead to biases towards the majority class. Most importantly, we do not have an external validation set to verify that our algorithm works in other hospital settings. Lastly the data is filled out by human clinicians and nurses, so the data entries are prone to human error and incompleteness.

With regards to the development of the machine learning models, the NN was restricted to three hidden layers. More experiments can be conducted to find the optimal number of hidden layers. Also, we only tested four algorithms. Developing other advanced models such as Neural Controlled Differential Equation (CDE) [18] may potentially yield higher results. Furthermore, we assumed that each chemotherapy session is an independent observation; however, there most likely is an underlying relationship between all of the chemotherapy administrations that a patient recieved. Our current algorithms did not take into account this longitudinal data of a patient's chemotherapy history.

Future Work

Directions for future research can be to include all chemotherapy regimens from the dataset, which will increase the number of patients significantly. We could also include the different drugs and dosages applied and whether patients received blood transfusion or blood growth stimulant agents at each chemotherapy session. In addition, we can identify additional variables that may aid in predicting cytopenias. An area of active investigation is to obtain an external validation set to test our model against other hospital data. Another extension to this study is to develop an application that not only informs doctors of a patient's risk for cytopenias, but also recommends potential mitigating strategies such as prescribing certain medication, administering blood transfusions, or delaying the chemotherapy administration.

Conclusion

This study demonstrated that machine learning algorithms may meaningfully contribute to the assessment of cytopenia risk (specifically neutropenia, anemia, and thrombocytopenia) in chemotherapy patients. Out of the four machine learning algorithms tested, the extreme gradient boosting tree performed the best and showed promising results for cytopenia prediction, achieving AUROC of 93% and AP of 66% on average over all three blood types. This can be used to facilitate mitigating actions, such as administering blood transfusion or delaying chemotherapy administration.

References

[1] Holborow, A., Coupe, B., Davies, M., & Zhou, S. (2019). Machine learning methods in predicting chemotherapy-induced neutropenia in oncology patients using clinical data. Clinical Medicine, 19(Suppl 3), s89–s90.

[2] DeWan, P. A., Inbar, O., Spina, C. S., Rudeen, K., Lagor, C., Walker, M. S., Stepanski, E. J., Nwankwo, J. O., & Hyde, B. (2018). Artificial intelligence methods to predict chemotherapy-induced neutropenia in breast cancer patients. Journal of Clinical Oncology, 36(15_suppl), 6555–6555.

[3] Cho, B.-J., Kim, K. M., Bilegsaikhan, S.-E., & Suh, Y. J. (2020). Machine learning improves the prediction of febrile neutropenia in Korean inpatients undergoing chemotherapy for breast cancer. Scientific Reports, 10(1).

[4] Du, X., Min, J., Shah, C. P., Bishnoi, R., Hogan, W. R., & Lemas, D. J. (2020). Predicting in-hospital mortality of patients with febrile neutropenia using machine learning models. International Journal of Medical Informatics, 139, 104140.

[5] Pawloski, P. A., Thomas, A. J., Kane, S., Vazquez-Benitez, G., Shapiro, G. R., & Lyman, G. H. (2016). Predicting neutropenia risk in patients with cancer using electronic data. Journal of the American Medical Informatics Association, 24(e1), e-e.

[6] Cuplov, V., & André, N. (2020). Machine Learning Approach to Forecast Chemotherapy-Induced Haematological Toxicities in Patients with Rhabdomyosarcoma. Cancers, 12(7), 1944.

[7] Avdic, D., Gallimore, M., Riley, M., & Bingham, C. (2015). Neutrophil count prediction for personalized drug dosing in childhood cancer patients receiving 6-mercaptopurine chemotherapy treatment.

[8] Extermann, M., Boler, I., Reich, R. R., Lyman, G. H., Brown, R. H., DeFelice, J., Levine, R. M., Lubiner, E. T., Reyes, P., Schreiber, F. J., III, & Balducci, L. (2011). Predicting the risk of chemotherapy toxicity in older patients: The Chemotherapy Risk Assessment Scale for High-Age Patients (CRASH) score. Cancer, 118(13), 3377–3386.

[9] Hurria, A., Togawa, K., Mohile, S. G., Owusu, C., Klepin, H. D., Gross, C. P., Lichtman, S. M., Gajra, A., Bhatia, S., Katheria, V., Klapper, S., Hansen, K., Ramani, R., Lachs, M., Wong, F. L., & Tew, W. P. (2011). Predicting Chemotherapy Toxicity in Older Adults With Cancer: A Prospective Multicenter Study. Journal of Clinical Oncology, 29(25), 3457–3465.

[10] Lyman, G. H., Lyman, C. H., & Agboola, O. (2005). Risk Models for Predicting Chemotherapy-Induced Neutropenia. The Oncologist, 10(6), 427–437.

[11] Friberg, L. E., Henningsson, A., Maas, H., Nguyen, L., & Karlsson, M. O. (2002). Model of Chemotherapy-Induced Myelosuppression With Parameter Consistency Across Drugs. Journal of Clinical Oncology, 20(24), 4713–4721.

[12] ICES. Data Discovery Better Health. https://www.ices.on.ca/. Accessed Aug 16, 2021.

[13] ECOG-ACRIN Cancer Research Group. ECOG Performance Status. https://ecog-acrin.org/resources/ecog-performance-status. Accessed Aug 16, 2021.

[14] Cancer Care Ontario. Drug Formulary. https://www.cancercareontario.ca/en/cancer-treatments/chemotherapy/drug-formulary. Accessed Aug 16, 2021.

[15] Cancer Therapy Evaluation Program. Common Terminology Criteria for Adverse Events (CTCAE). https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/CTCAE_v5_Quick_Reference_8.5x11.pdf. Accessed Aug 16, 2021.

[16] Platt, John. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif.. 10.

[17] Practical Bayesian Optimization of Machine Learning Algorithms https://papers.nips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf

[18] Kidger, Patrick & Morrill, James & Foster, James & Lyons, Terry. (2020). Neural Controlled Differential Equations for Irregular Time Series.

Appendix

I. Selected Variables

Questionnaires (from 0 - 9)	**Chemotherapy Treatment Data**	**Demographic and Clinical Data**	**Blood Work Data (Baseline Measurement)**
Anxiety	Cancer Location	Region of Ontario	Hematocrit Count
Pain	Cancer Type	Age	Leukocyte Count
Depression	Intent of Systemic Treatment	Sex	Erythrocyte Mean Corpuscular Volume
Wellbeing	Line of Therapy	Body Surface Area	Erythrocyte Count
Tiredness	Current Chemotherapy Cycle	ECOG Performance Status	Erythrocyte Mean Corpuscular Hemoglobin Count
Drowsiness	Chemotherapy Regimen		Erythrocyte Distribution Width
Lack of Appetite	Chemotherapy Cycle Duration		Erythrocyte Mean Corpuscular Hemoglobin Concentration Count
Shortness of Breath			Lymphocyte Count
Nausea			Eosinophil Count
			Basophil Count
			Monocyte Count
			Platelet Mean Volume
			Hemoglobin Count
			Platelet Count
			Neutrophil Count

II. Target Distribution

Before Resampling

	Training	Validation	Testing

	False	True	False	True	False	True
Low Neutrophil Count	65283	10499	21632	3571	21825	3373
Low Hemoglobin Count	62563	13219	20806	4397	20949	4249
Low Platelet Count	73201	2581	24355	848	24325	873

After Resampling

	Training		Validation		Testing	
	False	True	False	True	False	True
Low Neutrophil Count	72435	13671	23908	4687	21825	3373
Low Hemoglobin Count	68543	17563	22702	5893	20949	4249
Low Platelet Count	73201	12905	24355	4240	24325	873

III. Bayesian Optimization Best Hyperparameters

Models	**Hyperparameters**	**Value**
Logistic Regression	C	0.742299
Random Forest	Max Depth of Tree	7
	Max Features for Split	100%
	Number of Trees	194
Extreme Gradient Boosting Tree	Gamma	0.761009
	Learning Rate	0.0099233
	Max Depth of Tree	6
	Number of Trees	117
	Regularization Lambda	0.081782
Neural Network	Alpha	0.464623

	Batch Size	430
	Momentum	0.137504
	Hidden Layer Sizes	(101, 209, 95)
	Learning Rate	0.004663

IV. Top Ten Variables

Rank	Neutropenia	Anemia	Thrombocytopenia
1	Baseline Neutrophil Count	Baseline Hemoglobin Count	Baseline Platelet Count
2	Chemotherapy Regimen	Chemotherapy Regimen	Chemotherapy Regimen
3	Cancer Type	Region of Ontario	Chemotherapy Cycle Duration
4	Chemotherapy Cycle Duration	Cancer Location	Cancer Location
5	Cancer Location	Cancer Type	Cancer Type
6	Current Chemotherapy Cycle	Chemotherapy Cycle Duration	Baseline Monocyte Count
7	Region of Ontario	ECOG Performance Grade	Baseline Erythrocyte Count
8	Baseline Monocyte Count	Baseline Hematocrit Count	Intent of Systemic Treatment
9	Baseline Leukocyte Count	Intent of Systemic Treatment	Region of Ontario
10	Line of Therapy	Current Chemotherapy Cycle	Baseline Hemoglobin Count