Machine learning methods in predicting chemotherapy-induced neutropenia in oncology patients using clinical data

Authors: Alexander Holborow, Bryony Coupe, Mark Davies and Shangming Zhou

Introduction

Chemotherapy-induced neutropenia (CIN) incidence varies depending upon diagnosis and treatment regimen; however, in patients with solid tumours it has been reported to be as high as 15–22%.[1,2] The resulting dose adjustments or omissions can result in significant deviation from optimal treatment regimes, with the potential for failure to achieve expected rates of remission or duration of survival.[3] Prophylactic granulocyte colony-stimulating factor (G-CSF) can effectively reduce the incidence of CIN in these patients; however, prescription of CSF to all patients is not cost effective.[4]

Existing models of CIN risk are based primarily on treatment regimen with additional consideration of individual patient factors, including age, disease, and performance status.[5] Prophylactic G-CSF is routinely recommended if the expected risk of severe neutropenia is ≥20%; however, individual risk is difficult to quantify in the absence of an agreed mechanism.[6]

Due to the early adoption of electronic chemotherapy prescribing, a rich source of historical patient data has been developed relating to chemotherapy regimen and a subset of key variables. Modern data mining methods, incorporating machine learning approaches, can be utilised to analyse these data with the aim of producing more accurate, personalised predictions of the risk of neutropenia. We performed a comparison of several machine learning algorithms with a logistic regression of CIN risk in patients undergoing chemotherapy for solid tumours.

Methods

We performed a retrospective analysis of 15,119 patients aged 18 years or older with a diagnosis of cancer between 1 January 2000 and 31 December 2018 who were treated at the Singleton Cancer Centre in Swansea, South Wales, and whose data was recorded utilising the ChemoCare system.

Variables extracted included age, sex, cancer type, use of G-CSF, chemotherapy treatment regimen including dose and treatment date, total cycle number, history of prior chemotherapy, as well as several laboratory values (those from routine full blood count, liver function test and bone profile testing). Computed variables included presence or absence of neutropenia, defined as an absolute neutrophil count (ANC) of $<1 \times 10^9$/L.

Utilising these variables, we trained logistic regression, random forest, support vector machine, artificial neural network, naive Bayes and K-nearest neighbour algorithms in classification utilising the cycle 1 data with the presence or absence of a neutropenic event as the outcome variable. Data normalisation, binarisation and discretisation were performed where necessary. All models were optimised utilising parameter tuning and variable selection where appropriate. Error estimation was performed utilising tenfold cross-fold validation with three repeats, with algorithm performance tested on an external test dataset.

Table 1. Validation and test predictions by the different algorithms.

Algorithm	Validation AUC	Test AUC	Sensitivity	Specificity
Logistic regression	74.22	71.54	66.52	65.89
Artificial neural network	76.25	72.32	67.22	68.32
Naive Bayes	74.25	71.14	69.56	63.55
K-nearest neighbour	74.12	72.02	71.23	66.21
Random forest	77.22	73.24	68.99	72.25
Support vector machine	68.55	67.55	61.25	62.35

AUC = area under the curve

Results and discussion

Each of the algorithms achieved a comparable level of accuracy in predicting a neutropenic event (Table 1). The best overall performance was achieved using the random forest algorithm (77.22% validation and 73.24% test). The random forest algorithm also had the lowest loss in performance when applied to the test dataset.

Authors: Swansea University, Abertawe Bro Morgannwg Health Board, Wales

Conclusion

We have demonstrated that integrating data mining and machine learning approaches with routinely collected clinical data can be useful in developing classification algorithms that may aid clinical decision making. Utilising the best-performing algorithm, random forest, we have developed a web application that can offer individual risk predictions at point of care. With further improvements and validation, such a tool could be used to target G-CSF to those patients at greatest risk of neutropenia. ■

References

1 de Melo Gagliato D, Celloso Medrado Santos JP, Dino Cossetti RJ *et al*. Febrile neutropenia risk with adjuvant docetaxel and cyclophosphamide (TC) chemotherapy regimen in two Brazilians cancer centers. *J Integr Oncol* 2017;6:195.

2 Weycker D, Li X, Edelsberg J *et al*. Risk and consequences of chemotherapy-induced febrile neutropenia in patients with metastatic solid tumors. *J Oncol Pract* 2015;11:47–54.

3 Nagel CI, Backes FJ, Hade EM *et al*. Effect of chemotherapy delays and dose reductions on progression free and overall survival in the treatment of epithelial ovarian cancer. *Gynecol Oncol* 2011;124:221–4.

4 Rajan S, Carpenter WR, Stearns SC, Lyman GH. Short-term costs associated with primary prophylactic G-CSF use during chemotherapy. *Am J Manag Care* 2013;19:150–9.

5 Dang CT, Fornier MN, Hudis CA. Risk models for neutropenia in patients with breast cancer. *Oncology (Williston Park)* 2003;17(11 Suppl 11):14–20.

6 Pawloski PA, Thomas AJ, Kane S *et al*. Predicting neutropenia risk in patients with cancer using electronic data. *J Am Med Inform Assoc* 2017;24:e129–35.