Enhancing diagnostics for invasive Aspergillosis using machine learning

Simone Romano[a], James Bailey[a], Lawrence Cavedon[a,b], Orla Morrissey[c,d], Monica Slavin[e,f], Karin Verspoor[a,b]

[a]The University of Melbourne, Dept. of CIS
[b]NICTA (National ICT Aust.) VRL
[c]Alfred Health
[d]Monash University
[e]Peter MacCallum Cancer Centre
[f]Melbourne Health

SUMMARY

Invasive Aspergillosis (IA) is a serious fungal infection and a major cause of mortality in patients undergoing allogeneic stem cell transplantation or chemotherapy for acute leukaemia[1]. The major contributing factor to the high mortality rates is that culture methods have limited sensitivity, only detecting 40-50% of IA cases. The currently accepted criteria used for diagnosing IA are CT scan findings, microbiology and risk factors. However, because of ease of use and improved sensitivity, biomarkers such as Aspergillus PCR and Galactomannan (GM) assays are used both increasingly and alternatively. These tests are performed at least twice weekly. Two consecutively positive PCR or GM results or greater than two intermittently positive PCR or GM results within a two-week timeframe is taken as an indicator of Probable IA irrespective of the results of other tests[4]. The frequent testing coupled with moderate specificity of these biomarkers can result in a number of false positive results. It can be difficult to ascertain whether an individual result is a true or false positive. Large amounts of data are collected during the treatment of high-risk haematology patients and we propose leveraging such data to produce more accurate predictions of IA diagnosis. We describe here the application of machine learning techniques to predict probability of IA, which can be used to enhance the interpretation of biomarker results. When the estimated probability of infection is low, we identify 26.5% PCR/GM positive tests in our data that did not lead to an IA diagnosis within a week (TNR = 28.9%, NPV=100%). For such cases, antifungal treatment may be safely avoided, minimising over-treatment and drug toxicity, and reducing associated costs.

INTRODUCTION

Invasive Aspergillus (IA) has been associated with a 34-43% mortality rate[2] and a patient with IA incurs an added 7 days of hospital stay and extra $AU30,957 in hospital costs[3]. In a randomised controlled trial comparing different strategies for diagnosing IA[4], large amounts of data were collected from 240 patients undergoing allogeneic stem cell transplantation or chemotherapy for acute leukaemia between September 2005 and November 2009 at six Australian centres. All patients were tracked for 26 weeks from the beginning of their treatment, providing rich longitudinal data on daily and weekly tests for each patient. In total, the data consists of $240 \times 26 \times 7 = 43{,}680$ records, a large number that makes bed-side, interpretation a challenging task. In one strategy arm of the study twice weekly Aspergillus PCR and GM results were used to diagnose IA according to pre-defined criteria[4]. However, in some cases these biomarkers produced single positive results. Knowing that in these patients IA is associated with high mortality rates, a single positive result did trigger treatment with antifungal drugs which may have been unnecessary in some cases and may have resulted in avoidable toxicity and expense. We aim to aid the interpretation of a single positive result by providing information on the likelihood that it is a false positive by using machine learning techniques over the collected data to compute a value of infection probability when a single positive PCR/GM test is recorded.

DESCRIPTION

We focus on discriminating between the positive PCR/GM tests that are associated with an IA diagnosis within a week and those that are not associated with an immediate IA diagnosis. When a single positive PCR/GM is detected, two types of data could be used to improve the reliability of prediction: data related to the subject characteristics (constant along the treatment) and data related to events occurring in the recent past. The latter includes results of daily tests (e.g. blood tests) as well as individualised treatment determined by the clinicians during the days preceding the single positive PCR/GM result. Due to the temporal aspect of the data, it is not possible to employ a simple statistical prediction model such as logistic regression. Rather, it is necessary to tailor a machine learning model that exploits the trajectories of values over time.

Simone Romano

PhD Student
University of Melbourne

sromano@student.unimelb.edu.au

Simone Romano is a PhD student at the Computing and Information System department at the University of Melbourne. His main interests are in data mining and machine learning applied to biomedical problems. He is currently working on prediction of invasive aspergillosis using classification techniques.

METHODS

We used the Random Forests[5] machine learning algorithm for prediction due to its ability to cope with a large number of heterogeneous features. Our approach used both features that did not vary during the treatment, (e.g. gender, age, BMI, and smoking status) as well as features that varied over time, including neutrophil count, body temperature, corticosteroids doses, haemoglobin and platelet count. We generated the following features reflecting clinical intuitions about strategies for capturing the variation of data over time:

- Duration features: We counted the number of days the value each parameter lay within specified ranges. We limited our analysis to 30 days prior to a single positive PCR/GM reading, reflecting upper bound of the incubation time for IA[6]. We partitioned values for each parameter into percentiles or, for cases where we had specific knowledge about data type, we used pre-specified thresholds. For example, we divided temperature in 1 Celsius degree intervals, e.g. [36,37], (37,38] etc., and we counted number of days temperature occurred for a patient in each interval;
- Trajectories: In order to capture changes in the sampled value for a given parameter, we used the methodology proposed in[7]. We selected two days in the 3 week window preceding a single positive PCR/GM test and computed the mean value, the standard deviation, and the relative difference between those values. We performed this operation for each possible pair of days in the 3 week window. For example, we computed the mean value of temperature readings for the week prior to the single positive PCR/GM result, the standard deviation and the relative difference between the temperature measured one week prior to the positive test result and the day of the positive test result. We repeated this process for all possible intervals in the 3-week window. This technique has the capacity to capture subtle changes in the trajectories of values.

RESULTS

Our training set was a collection of 358 single positive PCR/GM tests that precede the earliest label of IA infection according to either diagnostic strategy (standard culture-based or biomarker strategy). Respectively, 284 cases were related to positive PCR and 83 cases to positive GM tests. Just 29 of the positive PCR/GM were associated with a Proven IA or Probable IA label within a week according to either diagnostic test. Thus 329 results might be considered as false positives. We built a predictive model that exploits the data available at the occurrence of a positive PCR/GM to output a probability of infection within a week value. We validated this model by a patient-level cross-validation framework, where we took care not to use data relative to one patient to both model training and validation. Setting a low threshold on the model output probability to achieve high NPV (100%) we were able to identify 95 such tests that do not lead to an IA infection (TNR = 28.9%) within a week.

CONCLUSIONS AND FUTURE WORK

The implemented predictive model seems to enable safe avoidance of antifungal therapy for a significant number of cases. It accurately discriminated between true and false single positive PCR/GM results. This tool can be an aid to further avoid over-treatment, reduce drug-toxicity, and reduce antifungal drug costs. Future work will aim to make the model more accurate in predicting when a positive PCR/GM is associated with an immediate infection to trigger the antifungal treatment earlier in time; search for alternative diagnosis when the outcomes are equally probable according to the model; and make the model output more interpretable to clinical practitioners, e.g. by identifying the trajectories in the data which generate a low or high probability of IA.

[†]NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

REFERENCES

1. D. Neofytos, D. Horn, et al., Epidemiology and outcome of invasive fungal infection in adult hematopoietic stem cell transplant recipients: analysis of Multicenter Prospective Antifungal Therapy (PATH) Alliance registry, Clinical Infectious Diseases 48(3): 265-273, 2009.
2. WJ. Steinbach, K. A. Marr, et al., Clinical epidemiology of 960 patients with invasive aspergillosis from the PATH Alliance registry. J. Infection 65(5): 453—456, 2012.
3. MR. Ananda-Rajah, A. Cheng, et al., Attributable hospital cost and antifungal treatment of invasive fungal diseases in high-risk haematology patients: an economic modeling approach. Antimicrob Agents and Chemotherapy, 2011: 55(5): 1953-60.
4. C. O. Morrissey, SC. Chen, et al.,Galactomannan and PCR versus culture and histology for directing use of antifungal treatment for invasive aspergillosis in high-risk haematology patients: a randomised controlled trial, Lancet Infectious Diseases 13(6), 519-528, 2013.
5. L. Breiman, Random forests, Machine Learning 45(1), 5-32, 2001.
6. H. Deng, G. C. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Information Sciences 239, pp. 142-153, 2013.
7. T. Benet, N. Voirin, et al., Estimation of the incubation period of invasive aspergillosis by survival models in acute myeloid leukaemia patients, Medical Mycology 51(2), 214-218, 2013.