Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210853

Early Prediction of Neoplasms Using Machine Learning: A Study of Electronic Health Records from the Ministry of National Guard Health Affairs in Saudi Arabia

Asma ABDULLAH ALFAYEZ [a,b,c], Alvina GRACE LAI [a] and Holger KUNZ [a,1]
[a] *Institute of Health Informatics, University College London, London, United Kingdom*
[b] *King Abdullah International Medical Research Center, Riyadh, Saudi Arabia*
[c] *King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia*

Abstract. The early detection and treatment of neoplasms, and in particular the malignant, can save lives. However, identifying those most at risk of developing neoplasms remains challenging. Electronic Health Records (EHR) provide a rich source of "big" data on large numbers of patients. We hypothesised that in the period preceding a definitive diagnosis, there exists a series of ordered healthcare events captured within EHR data that characterise the onset and progression of neoplasms that can be exploited to predict future neoplasms occurrence. Using data from the EHR of the Ministry of National Guard Health Affairs (MNG-HA), a large healthcare provider in Saudi Arabia, we aimed to discover health event patterns present in EHR data that predict the development of neoplasms in the year prior to diagnosis. After data cleaning, pre-processing, and applying the inclusion and exclusion criteria, 5,466 patients were available for model construction: 1,715 cases and 3,751 controls. Two predictive models were developed (using Decision tree (DT), and Random Forests (RF)). Age, gender, ethnicity, and ICD-10-chapter (broad disease classification) codes as predictor variables and the presence or absence of neoplasms as the output variable. The common factors associated with a diagnosis of neoplasms within one or more years after their occurrence across all the models were: (1) age at neoplasms/event diagnosis; (2) gender; and patient medical history of (3) diseases of the blood and blood-forming organs and certain disorders involving immune mechanisms, and (4) diseases of the genitourinary system. Model performance assessment showed that RF has higher Area Under the Curve (AUC)=0.76 whereas the DT was less complex. This study is a demonstration that EHR data can be used to predict future neoplasm occurrence.

Keywords. Machine learning, decision tree, random forest, neoplasms, cancer, artificial intelligence, big data.

[1] Corresponding author, Holger Kunz, Institute of Health Informatics, University College London, 222 Euston Road NW12DA, London, United Kingdom; E-mail: h.kunz@ucl.ac.uk

1. Introduction

The objective of this study was to discover health event patterns present in Electronic Health Record (EHR) data that predict the development and occurrence of neoplasms (malignant and benign) in one year prior to diagnosis based on the patient's medical history by developing predictive models. This will improve neoplasms prediction to support the early diagnosis and as a warning system. After developing machine learning models, we compared models' performances and complexity to discuss their likely applicability in practice.

2. Methodology

This study used data from the Ministry of National Guard Health Affairs (MNG-HA) in Saudi Arabia. It is one of the largest healthcare organisations and has pioneered healthcare advances in Saudi Arabia [1, 2]. The data covered all patient encounters: in-patient, out-patient, and emergency department (ED) visits in the primary and secondary care settings between 2015 to the beginning of 2019 were included. Patient data were extracted from the EHR and included: patients' demographics and diseases history. We selected patients according to the inclusion and exclusion criteria to obtain the case group (patients with neoplasms) and control group (patients without neoplasms). We chose the first neoplasm incident/diagnosis as the index event before which the medical history was retrieved for the case group. For the control group, the last disease diagnosis was used as the index since there was no mutual diagnosis/event for the entire control group that could be used as the index event. For modelling the impact of prevalent comorbidities on early neoplasms prediction, we modelled the medical history using the International Classification of Diseases, Tenth Revision (ICD-10) codes in the EHR. We then implemented two machine learning predictive algorithms to build the models then compared their performances and complexities. The algorithms were: Decision Tree (DT) and Random Forests (RF). We chose these models as they are used to predict binary outcomes based on a set of observed characteristics and explore the effect of these characteristics on the outcome. The outcome was the presence or absence of neoplasm whereas the characteristics were the medical history beside the social demographic variables: age, gender, and ethnicity. We used the following performances metrics: Area Under the Curve (AUC); balanced accuracy; sensitivity; specificity; positive predictive values; negative predictive values; and F-score for accuracy.

3. Results

The total population was 17,631 patients including both: the case and control groups. The case group has 11,204 patients and the control group has 6,427 patients. After data cleaning, pre-processing, and applying the inclusion and exclusion criteria, 5,466 patients remained: 1,715 cases and 3,751 controls. Table 1 shows the top ten neoplasms represented in the study sample. It shows the ICD10 codes, the corresponding neoplasms types for each ICD10 code according to the World Health Organisation (WHO) [3], and the number of cases for each.

Table 1. Top ten neoplasm types represented in our study sample

ICD10 code	Type	N(%)
C50	malignant neoplasms of breast	153(9%)
C73	malignant neoplasms of the thyroid gland	147(8.6%)
C22	malignant neoplasms of the liver and intrahepatic bile ducts	140(8.2%)
C18	malignant neoplasms of the colon	99(5.8%)
C64	malignant neoplasms of the kidney, except renal pelvis	87(5%)
D43	neoplasm of uncertain or unknown behaviour of brain and central nervous system	85(5%)
C61	malignant neoplasms of the prostate	61(3.6%)
C67	malignant neoplasms of the bladder	58(3.4%)
D24	benign neoplasms of the breast	55(3.2%)
C25	malignant neoplasms of the pancreas	46(2.7%)

The mean and standard deviation of the follow-up time for the case and control groups one year before the event were 624 days (SD: 188) and 731 days (SD: 231) respectively. Table 2 shows the 10-fold cross-validation prediction performance assessment results for the models. RF had a slightly higher predictive performance with identical performance in terms of NPV (0.90) with the DT.

Table 2. 10-fold cross-validation model performance assessment results. AUC: area under the curve; BA: balanced accuracy; SE: sensitivity; SP: specificity; PPV: positive predictive values; NPV: negative predictive values; and F1-score

Model	AUC	BA	SE	SP	PPV	NPV	F1-score
DT	0.74	0.77	0.73	0.82	0.58	0.90	0.64
RF	**0.76**	0.79	0.74	0.84	0.61	0.90	0.67

4. Discussion

The aim of this study was to develop models that predict the occurrence of a neoplasm within one year using medical history data obtained from the patients' EHRs. Two models: DT and RF for neoplasm prediction were used. The common factors associated with a diagnosis of neoplasms within one or more years after their occurrence across all the models were: (1) age at neoplasms/event diagnosis; (2) gender; and patient medical history of (3) diseases of the blood and blood-forming organs and certain disorders involving immune mechanisms; and (4) diseases of the genitourinary system. These factors were associated with subsequent neoplasm development but predicted differently: age and male gender increased the likelihood of developing neoplasms, which are known risk factors as older people and males are more likely to develop neoplasms compared to younger and female patients [4-6].

This study used a small subset of the data available in the EHR. Therefore, this analysis is unlikely to have captured all of the variables that predict neoplasm occurrence. For example, while smoking and alcohol status were available and are known risk factors

for malignant neoplasms [7], we did not use them as most of their values in the EHR were either not recorded after being collected from the patient, were missing, or were not collected from the patients at all.

RF showed a better performance in terms of AUC. RF model, by definition, used all the provided predictors, twenty-seven in our case, making it more complex in this study. In contrast, the DT model used only four of the input predictors, making it less complex model. Therefore, while RF showed better performance it was also more complex, whereas DT showed lower performance but was less complex model.

5. Conclusion

In this study, two ML models have been used to predict neoplasms (malignant and benign) one year prior to diagnosis. The two ML algorithms used to build the models were: (1) DT and (2) RF. The medical history was obtained from the EHR to construct a feature space for training and testing the two models. The study showed that ML algorithms applied to data from EHR's could improve and support the early diagnosis of neoplasms.

References

[1] King Abdulaziz Medical City in Riyadh [Internet]. Kingdom of Saudi Arabia Ministry of National Guard Health Affairs. 2020 [cited 2021Sep17]. Available from: https://www.ngha.med.sa/English/MedicalCities/AlRiyadh/Pages/default.aspx

[2] Kingdom of Saudi Arabia Ministry of National Guard Health Affairs. 2018 [cited 2021Sep15]. Available from: https://www.ngha.med.sa/English/AboutNGHA/Pages/profile.aspx

[3] International Statistical Classification of Diseases and Related Health Problems 10th Revision [Internet]. ICD-10 Version:2010. World Health Organisation (WHO); [cited 2021Jul14]. Available from: https://icd.who.int/browse10/2010/en

[4] White MC, Holman DM, Boehm JE, Peipins LA, Grossman M, Henley SJ. Age and cancer risk: a potentially modifiable relationship. American journal of preventive medicine. 2014 Mar;46(3):S7-15.

[5] Laconi E, Marongiu F, DeGregori J. Cancer as a disease of old age: changing mutational and microenvironmental landscapes. British journal of cancer. 2020 Mar;122(7):943-52.

[6] García-Martín JM, Varela-Centelles P, Gonzále M, Seoane-Romero JM, Seoane J, García-Pola MJ. Epidemiology of Oral Cancer. In: Oral Cancer Detection. Cham: Springer International Publishing; 2019. p. 81–93.

[7] Hydes TJ, Burton R, Inskip H, Bellis MA, Sheron N. A comparison of gender-linked population cancer risks between alcohol and tobacco: how many cigarettes are there in a bottle of wine?. BMC public health. 2019 Dec;19(1):1-8.