MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190259

Development of Deep Learning Algorithm for Detection of Colorectal Cancer in EHR Data

Yu-Hsiang Wang[a], Phung-Anh Nguyen[b,*], Md. Mohaimenul Islam[b], Yu-Chuan Li[b,c], Hsuan-Chia Yang[b]

[a] College of Medicine, Taipei Medical University, Taipei, Taiwan
[b] International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan
[c] College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
[*] Corresponding author

Abstract

We aimed to develop a deep learning model for the prediction of the risk of advanced colorectal cancer in Taiwanese adults. We collected data of 58152 patients from the Taiwan National Health Insurance database from 1999 to 2013. All patients' comorbidities and medications history were included in the development of the convolution neural network (CNN) model. We also used 3-year medical data of all patients before the diagnosed colorectal cancer (CRC) as the dimensional time in the model. The area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were computed to measure the performance of the model. The results showed the mean (SD) of AUC of the model was 0.922 (0.004). Moreover, the performance of the model observed the sensitivity of 0.837, specificity of 0.867, and 0.532 for PPV value. Our study utilized CNN to develop a prediction model for CRC, based on non-image and multi-dimensional medical records.

Keywords:
Colorectal Neoplasms; Algorithms; Electronic Health Records;

Introduction

Cancer has been the leading cause of death since 1986 [2] and colorectal cancer (CRC) has been ranked as the third cause of cancer-related death in Taiwan since 2008 [5]. The incidence rate of CRC was 43.6 per 100,000 population in 2015, the age-adjusted incidence rate of male was higher than female. In fact, it increased by 13% for male and 6% for a female since 2002. CRC is now a burden and challenge to Taiwan's health care [1]. Therefore, it is important to find the potential biomarkers that can contribute to early diagnosis of colorectal cancer.

Artificial intelligence (AI) and machine learning are taking all fields of science by storm including clinical medicine [7; 10]. Deep learning is a type of the highly flexible machine learning approach, has emerged a groundbreaking algorithm to enhance the performance of current machine learning techniques and to solve complex problems [7]. Several studies already showed that machine learning model can predict cancer that help physicians to reduce the number of false decisions and improve health outcomes.

In this study, we aimed to develop a deep learning model by applying all patients phenotypic information to detect the onset of colorectal cancer.

Methods

Data Sets

In this study, we used the reimbursement data from Bureau National Health Insurance (NHI) system in Taiwan that had registered all medical records since 1996 [8]. We obtained a randomization of two million population of the NHI claim data from January 01, 1999 to December 31, 2013. The study was approved by the institutional review board committee at Taipei Medical University and the need for informed consent was waived.

Study population

We initially identified all patients who were diagnosed with a primary diagnosis of colorectal cancer (CRC) (International Classification of Disease, Ninth Revision [ICD-9] codes 153, 154) between January 01, 1999, and December 31, 2013. The diagnosis accuracy of CRC was confirmed by both specific ICD-9 codes and inclusion in the Registry for Catastrophic Illness Patient (RCIP), a subpart of the NHI database [6]. Surgical pathological confirmation or undergoing treatment process of CRC is required for patients to be registered in the RCIP. However, the index date was defined as the date of diagnosis. Patients were excluded if they were younger than 20 years of age at the date of diagnosis CRC or if they did not have any outpatient claims at each calendar year of four years before the index date.

We then randomly selected 50,000 individuals without any cancer diagnosis during 15 years in the NHI claim database. All medical records of those were also collected. We defined the date of the last visit observed in the NHI database as the index date of patients with cancer-free. We then used the same conditions as CRC patients to cross-check their age and the history of medical data.

In this study, our purpose was to predict CRC in one year ahead; we, therefore, used the data of patients with and without cancer diagnosis from the fourth year to the first year before the index date and left out a year data to be the gap of developing CRC cancer **(Figure 1).**

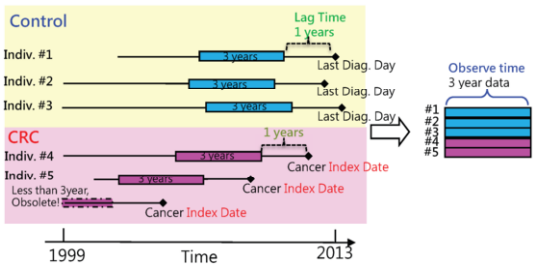

Figure 1. Observation Window and Lag Time

Feature definition

First, we checked the index date to observe the age and sex of overall individuals (i.e., with and without cancer diagnosis). We included the ICD9-CM 3-digit codes such as 001-999, V01-V91 in the features of the development of the model. The supplementary classification of external causes of injury and poisoning codes, E000-E999 were excluded in the feature list.

For the medications, the Taiwan medication codes were collected and mapped to WHO-ATC (Anatomical Therapeutic Chemical Classification System) codes. The ATC-4 (ie. ATC 5-digit) codes were used in the further development of the model. A total of 1931 features (ie. age, sex, 1099 ICD9-CM codes, and 830 ATC-4 codes) were used in the development of the CRC prediction model.

Time definition and data processing

We collected the data of all individuals in 1095 days (i.e., 3-year data). For each disease, the patient was considered as a diagnosis if at least three visits were observed with its ICD9 code during the 3-year period. For each medication, the patient was defined as a user if its ATC code was observed in the prescriptions, and the medication observation time was the number of days prescribed by physicians. For each patient, the 7-day data was summed up continuously, and the final data of 157 grids were derived from 1095-day data set **(Figure 2)**. The value of each feature of each patient was then normalized to scale up to the range in between 0 and 1 in order to make the similar severity of all individuals as the same cause. It expresses mathematically in (1) as: $x_{new} = {x - x_{min}}/{x_{max} - x_{min}}$ (1), where x = (x_1, …, x_n) and x_{new} is now the new normalized value. **Figure 3** shows an example of the normalized data.

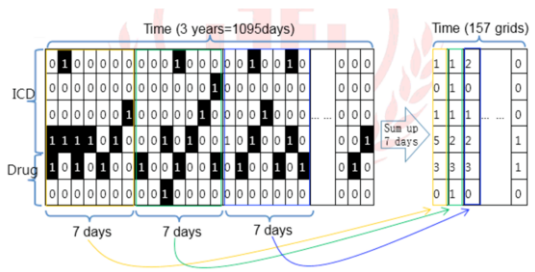

Figure 2. The processing of time data

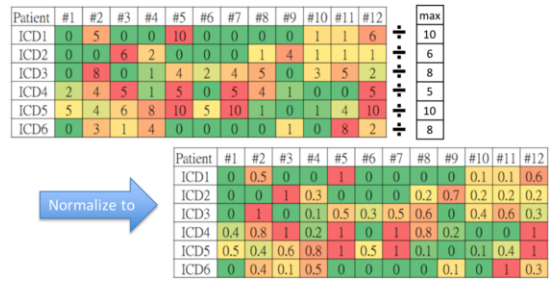

Figure 3. An example of normalized data

Development of the model

Deep learning is the process of training a neural network (a large mathematical function with millions of parameters) to perform a given task. **Figure 4** shows the convolution neural network (CNN) structure of the study. We created the function that computed the CRC severity from a large set of the matrix of all individuals. During the training process, the parameters of the neural network were initially set to random values. The total of 1929 features data including 1099 ICD9 codes and 830 ATC codes, were categorized into 19 groups (i.e., 18 groups for ICD9, for example, 001-139-infectious and parasitic disease group, and one ATC group) for all patients. We created two convolutional layers with 32 feature maps for each category. Two max-pooling layers with the filters of size 1x3 were applied to reduce the number of parameters, and to control overfitting. After that, we flattened the pooled feature maps with the addition of age and sex into another layer in order to process the artificial neural network. A hidden layer of 400 neurons was used in this fully-connected layer.

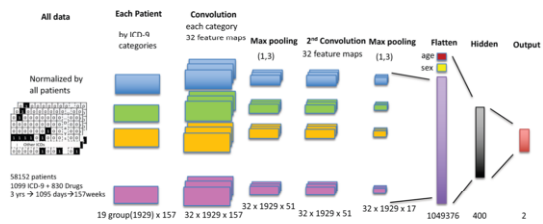

Figure 4. CNN structure of the study

The optimization algorithm used to train the network weights was a distributed stochastic gradient descent with a mini-batch size of 32. The model was optimized using Adadelta optimizer [13]. The input and hidden layers used a Rectified Linear Unit (ReLU) activation function (refs), while the output layer used the Softmax activation function. The dropout value of 0.1 used to prevent the overfitting and was applied to the input and hidden layer [9].

The average k-fold cross-validation accuracy, with a k-value of 5, was used as the metric to determine the best performance, optimizer, and loss functions of the model. We computed the area under the receiver operating curve (AUC) value, sensitivity, specificity, and positive predictive value (PPV) to observe the model performance.

The software was implemented using python v.3.7 and the model was created and trained with the Tensorflow framework v.1.9.

Results

The demographic characteristics of cancer and non-cancer patients are summarized in **Table 1**. Following the conditions, there were 58152 patients including in the study, of which 10185 diagnosed with CRC and 47967 patients were without a cancer diagnosis. The mean age (SD) of patients with CRC diagnosis was 62.7 (14.2) compared to patients without cancer of 47.6 (17.3). Most of the patients diagnosed with CRC were male (54.1%). In addition, the average number of diagnosis of a CRC patient was significantly higher compared to a non-cancer patient before the index date (e.g. 41.2 vs. 26.4 in the first observed year).

Table 1. Characteristics of patients in this study

	Case (n=10185)	Control (n=47967)
Age mean (sd)	62.7 (14.2)	47.6 (17.3)
Male sex (%)	5513 (54.1%)	23109 (48.2%)
Diagnosis count	1119580	3735982
Average in 1 year before index date*	41.2	26.4
Average 1~2 year before index date*	39.9	28.6
Average 2~3 year before index date*	37.7	27.4

*Note: * Number per patient per year*

Figure 5 shows the AUROC of the model with 5-fold cross-validation. The mean (SD) of AUC of the model was 0.922 (0.004). The performance of the model observed with the sensitivity of 0.837, specificity of 0.867, and 0.532 for PPV value.

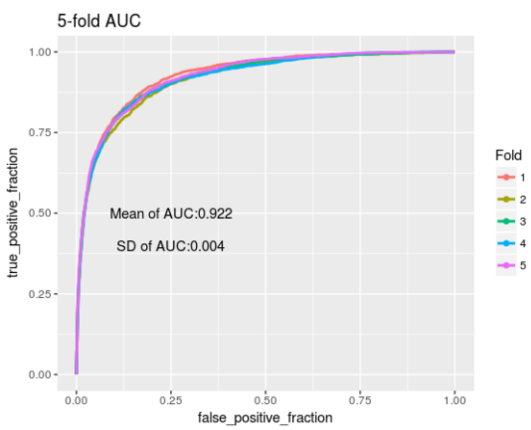

Figure 5. AUROC of the model with 5-fold cross-validation

Discussion

Our study findings demonstrate that the CNN model could appropriately predict CRC based on non-image and multi-dimensional medical records (ie. phenotype, diseases and medication). The results showed higher AUC (0.922) with very good sensitivity and specificity. This automated system for the detection of CRC offers several advantages, including consistency of interpretation (it could be a good screening tool for CRC for any person before deciding to take further some examination tests).

In this study, we tried to observe the features weight by using the odds ratio (ORs) and stepwise feature extraction (i.e., ANN model). We also selected top 10 features and most of the features were related to cardiovascular diseases. Comparing to other studies [3; 4; 12], our study showed higher performance (Area under receiver operating curve, sensitivity, and specificity). Since, our government has already launched a nationwide screening program of CRC since 2004 and a free fecal immunochemical test is offered biennially to individuals aged 50 to 75 [11]. According to our findings, the program may extend to cover all patients beyond this age range.

The main strength of this study lies in a use of a population-based cohort with a large and nationally representative sample, which increases its generalizability in Taiwan. However, there are several limitations in our study that need to be addressed. First, this is a retrospective cohort study, so selection bias and misclassification may be introduced. Second, inherent limitations of NHIRD hinder our ability to get some information related to the CRC, such as smoking habits, alcohol consumption, body mass index (BMI), family history of CRC, diet, and physical activity. However, we have included the comorbidities as surrogates for some risk factors of CRC, such as COPD and a stop-smoking clinic for smoking, alcohol-related illness for alcohol, and obesity for BMI. Finally, many chronic diseases such as diabetes mellitus (DM) and hypertension are not labeled every time on diagnosis, which may lessen the importance of those diseases.

Future direction

The ultimate goal of our study is to develop a web service for the public, helping them make predictions of risk of getting CRC one year before based on prior three-year-diagnosis and medication records. Therefore, this model provides a cheap and instant first-line screening for CRC which doesn't require medical examination in the first place. In the future, we would like to include more features in our study, including lifestyle, physical characteristics, smoking history, lab examination, and genetic information that help predict the risk more accurately.

Conclusions

In our study, we developed a deep learning model to predict CRC based on non-image and multi-dimensional medical records (diagnoses and medications). Our model showed higher performance for detecting CRC one year ahead. Indeed, previously published literatures did not show any time frame to predict CRC and traditional methods are not able take time frame consideration for predicting CRC. But, time sequence matters in the real world clinical setting. Therefore, our model can be considered in the real clinical settings.

Acknowledgements

This research was sponsored in part by Ministry of Science and Technology (MOST) 107-2634-F-038-002-, and Ministry of Science and Technology (MOST) 106-2634-F-038 -001 -CC2.
We would like to thank Mr. Chia-Wei Liang for his assistant in this study.

References

[1] T.C.R. Center, Report data, in.
[2] B.K. Chen and C.Y. Yang, Temporal trend analysis of avoidable mortality in Taiwan, 1971-2008: overall

progress, with areas for further medical or public health investment, *BMC Public Health* **13** (2013), 551.
[3] J. Grundner, H.U. Prokosch, M. Sturzl, R. Croner, J. Christoph, and D. Toddenroth, Predicting Clinical Outcomes in Colorectal Cancer Using Machine Learning, *Stud Health Technol Inform* **247** (2018), 101-105.
[4] M.-H. Hsieh, L.-M. Sun, C.-L. Lin, M.-J. Hsieh, K. Sun, C.-Y. Hsu, A.-K. Chou, and C.-H. Kao, Development of a Prediction Model for Colorectal Cancer among Patients with Type 2 Diabetes Mellitus Using a Deep Neural Network, *Journal of clinical medicine* **7** (2018), 277.
[5] T.N.H. Insurance, Top 10 cancer in Taiwan 2008-2014, in.
[6] U. Iqbal, P.A. Nguyen, S. Syed-Abdul, H.C. Yang, C.W. Huang, W.S. Jian, M.H. Hsu, Y. Yen, and Y.C. Li, Is long-term use of benzodiazepine a risk for cancer?, *Medicine (Baltimore)* **94** (2015), e483.
[7] C. Naylor, On the prospects for a (deep) learning health care system, *Jama* **320** (2018), 1099-1100.
[8] P.A. Nguyen, S. Syed-Abdul, U. Iqbal, M.-H. Hsu, C.-L. Huang, H.-C. Li, D.L. Clinciu, W.-S. Jian, and Y.-C.J. Li, A Probabilistic Model for Reducing Medication Errors, *PLOS ONE* **8** (2013), e82401.
[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research* **15** (2014), 1929-1958.
[10] B. Wahl, A. Cossy-Gantner, S. Germann, and N.R. Schwalbe, Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?, *BMJ Global Health* **3** (2018).
[11] Y.W. Wang, H.H. Chen, M.S. Wu, and H.M. Chiu, Current status and future challenge of population-based organized colorectal cancer screening: Lesson from the first decade of Taiwanese program, *J Formos Med Assoc* **117** (2018), 358-364.
[12] H.-J. Yang, C.W. Cho, S.S. Kim, K.-S. Ahn, S.-K. Park, and D.I. Park, Application of deep learning to predict advanced neoplasia using big clinical data in colorectal cancer screening of asymptomatic adults, *Gastrointestinal endoscopy* (2018).
[13] M.D. Zeiler, ADADELTA: an adaptive learning rate method, *arXiv preprint arXiv:1212.5701* (2012).

Address for correspondence

Phung-Anh Nguyen, Ph.D.
International Center for Health Information Technology
College of Medical Science and Technology
Taipei Medical University
Address: 250-Wuxing Str., Xinyi Dist., Taipei 11031, Taiwan
Email: alexnthhp@tmu.edu.tw
Phone: +886 266382736 #1528