Chung Shan Medical Journal 2025; 36:111-118

airiti

Original Article

Risk factors predicting rectal cancer recurrence by machine learning model

Jiun-Yi Hsia[1,2], Chi-Chang Chang[3,4*], Ying-Chen Chen[5]

[1] *Division of Thoracic Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taichung City, Taiwan.*

[2] *School of Medicine, Chung Shan Medical University, Taichung, Taiwan.*

[3] *School of Medical Informatics, Chung Shan Medical University & IT office, Chung Shan Medical University Hospital, Taichung City, Taiwan.*

[4] *Department of Information Management, Ming Chuan University, Taoyuan, Taiwan*

[5] *Institute of Medicine, Chung Shan Medical University, Chung Shan Medical University Hospital Taichung City, Taiwan.*

Purpose: Despite advances in treatment modalities, the recurrence event is still a clinical challenge for patients with rectal cancer. Previous studies have shown that fewer studies have compared the clinical benefits of different predictive models to improve predictive outcomes for individual conditions.

Method: A total of 3,403 hospital-based cancer registry records by six cancer registries in Taiwan, between 2011 and 2020. Different from traditional statistics analysis, the approach is mainly based on the predicted classification models, which were classification and regression trees, random forests. The modelling performance of predictive models was evaluated by accuracy, sensitivity, specificity, F1 score, Kappa, Matthews correlation coefficient (MCC), Area under the curve (AUC). Using decision curve analysis (DCA) and clinical impact curves, predictive models to evaluate the best clinical benefits.

Results: The results of this study showed that risk factors for recurrence are related to lifestyle, genetics, environmental factors, and the use of anti-cancer therapy. Unlike traditional regression model analysis, the risk factor analysis by machine learning methods used in our study could deal with the existence of complex interactions between different predictor variables. Furthermore, these results suggest that effective identification of treatment-associated factors that increase the risk of recurrence is a growing concern for cancer survivors.

Conclusion: The results of this study can be used as a reference for clinicians to assist in clinical decision making and provide maximum clinical benefit to patients.

Keywords: rectal cancer, recurrence, machine learning, chemotherapy, decision curve analysis, clinical impact curve.

1. Introduction

One of the most difficult problems for cancer survivors is recurrence. The number of cases of rectal cancer was the seventh most common cancer in the world (IARC) (1), (2). However, despite advances in treatment technology, recurrent events remain a critical clinical problem for survivors of rectal cancer (3). In the past, survival has been

* Corresponding Author: Chi-Chang Chang
Address: No.110, Sec.1, Jianguo N.Rd., Taichung City 40201, Taiwan
Tel: +886-4-2473-0022 ext. 12218
E-mail: threec@csmu.edu.tw

airiti

a priority in the clinical management of rectal cancer. Due to advances in treatment methods, periodic events have become a key problem in risk management. Consequently, the ability to conduct integrated analyses from an empirical medical point of view is currently an important issue in the prevention and treatment of recurrence in rectal cancer survivors. Manfredi et al. (2001) found that the stage at which the disease was diagnosed played an important role (4). Asoglu et al. In 2007, CEA levels and overall survival ($p = 0.004$) had a significant correlation (5). Chen et al. (2009) investigated the correlation between clinical pathological factors and rectal recurrence (6). In recent years, the use of machine learning technology in the clinical analysis of evidence-based medicine has become increasingly important (7). The purpose of this study was to identify important risk factors for cancer from the rectal cancer registry database, evaluate the prediction models developed for different factors and analyse clinical characteristics. Analysis of clinical features is essential for the development of strategies for the prevention and treatment of recurrence.

2. Materials and Methods

2.1. Study design

The study used cancer registration data from six hospitals between 2011 and 2020. Rectal cancer data sets include Chang Guang Medical Hospital (5,725), Chung Shan University Medical Hospital (259), Taichung Tzu Chi Hospital (118), Far East Memorial Hospital (221), Jen-Ai Hospital (192), and Chi Mei Medical Center (715). Between 2011 and 2020, Taiwan has 7,230 hospital-based cancer registry records. All data were collected according to the following criteria: (1) Primary location of rectal cancer, (2) hospital treatment. The exclusion criteria were: (1) cases that were never free of disease, (2) incomplete records of variable data. To ensure data quality, we include conditions as criteria for data cleaning. With the exception of data that are not the first repeats and incomplete records in variable data, 3827 cases have been excluded.

2.2. Analytical methods

In this study, the models are used as prediction models and are used to classify important risk factors based on their importance. Breiman (1984) proposed the CART algorithm. It can be used to solve various problems, such as classification and regression. This method uses a binary partition rule to summarise and analyze data sets with a large number of complex variables. CART algorithms classify data into a tree-like process with roots, nodes, and leaves. It is a widely used method to learn decision trees. The C4.5 algorithm was proposed in 1992 and the ID3 algorithm was proposed by Quinlan in 1986 (8, 9). The core of the ID3 algorithm is to use the information gain criterion to select the characteristics of each node in the decision tree and to construct the decision tree recursively. C4.5 will reflect the training sequence and gradually form the tree structure of attribute variables based on the highest information-winning ratio. C5.0 is a method used to analyse large data sets, classify and analyze hidden data, and display useful data. This algorithm uses a decision tree and the method was obtained mainly from a training group using a recursive division and selection method. C5.0 improves some of the problems of C4.5. It is suitable for processing large data sets and uses Boosting to improve model accuracy. You can use algorithms to generate decision trees or rule sets. According to the C5.0 model, the samples are classified according to their maximum information gain. The samples under the first classification are then reclassified according to different columns and the classification is repeated until no other samples are classified. Finally, the lowest classification level is reaffirmed, and sample subsets that do not contribute significantly to model values are cut.

In this study, prediction models were evaluated on the basis of accuracy, sensitivity, specificity, accuracy, F1 score, kappa, Matthew's correlation coefficient (MCC), curve area (AUC) and clinical impact curve. The performance of a machine learning algorithm is usually measured by its predictive accuracy. Precision measurement is the percentage of accurately predicted predictions in all classes. This study was conducted in the environment of the Weka software version 3.8.5 and the R programming language software version 4.1.3. The SMOTE method oversamples data sets to

airiti

solve the category imbalance problem. Data sets were divided into 70% of the training set and 30% of the testing set.

3. Results

The sample for the analysis of rectal cancer was 3,403 valid cases between 2011 and 2020, of which 497 (13.8%) were confirmed recurrences. The description of the population of recurrent rectal cancer samples and the results of the statistical analysis to predict the overall risk factors for recurrence are shown in Table 3.

Based on the literature of rectal cancer survivors and discussions with clinical experts, 16 risk factors were selected: sex, age at diagnosis, tumor size, regional lymph nodes examined, combine stage group, surgical margins of the primary site, radiotherapy, BMI (Body mass index), smoking behavior, betel nut chewing behavior, Alcohol use, carcinoembryonic antigen (CEA) lab value, carcinoembryonic antigen (CEA), obstruction, perforation, chemotherapy (Table 1). The combined stage group is based on the clinical stage, or pathological stage depending on the status of the treatment. The coding criteria for the combine stage group are: (1) The clinical stage is used for patients who have surgery after neoadjuvant therapy, (2) the pathological stage is used for patients without neoadjuvant therapy and with surgery, (3) The clinical stage is used for patients without neoadjuvant therapy and without surgery.

To further validate the above statistical risk factor and check the results as an auxiliary illustration, the model of this study used CART, C4.5, RF and C5.0 to calculate the importance ranking of the risk factors, as shown in Table 2. This study also analyzed the ranking of important risk factors according to the average Borda count (Tseng et al., 2017). In Table 2, the results of the three groups of data analysis for rectal cancer showed that the most

Table 1. Predicted risk factors and coding content in this study

Risk Factors	Coding Standards	
Sex	1：Male	2：Female
Age at Diagnosis	1：< 65 years	2：≧65 years
Tumor Size	1：≦5cm	2：>5 cm
Regional Lymph Nodes Examined	1：No	2：Yes
Combine stage group	1：≦ 2	2：> 2
Surgical Margins of The Primary Site	1：Free	2：Non-free
Radiotherapy	1：No	2：Yes
BMI(Body mass index)	1：< 24 kg/m2	2：≧24 kg/m2
Smoking Behavior	1：No	2：Yes
Betel Nut Chewing Behavior	1：No	2：Yes
Alcohol use	1：No	2：Yes
Carcinoembryonic Antigen (CEA) Lab Value (ng/ml)	1：≦50 ng/ml	2：>50 ng/ml
Carcinoembryonic Antigen (CEA)	1：>Normal value	2：≦Normal value
Obstruction	1：Negative	2：Positive
Perforation	1：Negative	2：Positive
Chemotherapy	1：No	2：Yes
Recurrence	1：N	2：Y

airiti

Table 2. Predicting risk factors for recurrence by comparing four models

Rank	CART	C4.5	RF	C5.0	Borda count (Overall)
1	Surgical Margins of The Primary Site	Combine stage group	Combine stage group	Sex	Chemotherapy
2	Combine stage group	Chemotherapy	Chemotherapy	Carcinoembryonic Antigen (CEA) Lab Value	Combine stage group
3	Chemotherapy	Carcinoembryonic Antigen (CEA) Lab Value	Carcinoembryonic Antigen (CEA) Lab Value	Smoking Behavior	Carcinoembryonic Antigen (CEA) Lab Value
4	Carcinoembryonic Antigen (CEA) Lab Value	Carcinoembryonic Antigen (CEA)	Carcinoembryonic Antigen (CEA)	Tumor Size	Surgical Margins of The Primary Site
5	Perforation	Radiotherapy	Radiotherapy	Radiotherapy	Regional Lymph Nodes Examined
6	Carcinoembryonic Antigen (CEA)	Surgical Margins of The Primary Site	Surgical Margins of The Primary Site	Combine stage group	Carcinoembryonic Antigen (CEA)
7	Regional Lymph Nodes Examined	Tumor Size	Tumor Size	Alcohol use	Tumor Size
8	Radiotherapy	Regional Lymph Nodes Examined	Regional Lymph Nodes Examined	Age at Diagnosis	Radiotherapy
9	Obstruction	Obstruction	Obstruction	Chemotherapy	Obstruction
10	Tumor Size	Perforation	Age at Diagnosis	BMI	Age at Diagnosis
11	Betel Nut Chewing Behavior	Smoking Behavior	Smoking Behavior	Obstruction	Betel Nut Chewing Behavior
12	Smoking Behavior	Betel Nut Chewing Behavior	Alcohol use	Regional Lymph Nodes Examined	Smoking Behavior
13	Alcohol use	Alcohol use	Sex	Carcinoembryonic Antigen (CEA)	Sex
14	Age at Diagnosis	Age at Diagnosis	Perforation	Betel Nut Chewing Behavior	Perforation
15	Sex	Sex	BMI	Surgical Margins of The Primary Site	Alcohol use
16	BMI	BMI	Betel Nut Chewing Behavior	Perforation	BMI

important risk factors were: overall stage, SSF1: CEA test value.

The results of the rectal cancer recurrence prediction are shown in Table 3, and the AUCs of the five prediction models are shown in Figure 1. In this study, the highest AUC value and accuracy for the overall recurrence prediction model was C5.0 (0.8635).

airiti

Table 3. The performance evaluation results of recurrence predictive models

Method	Accuracy	Sensitivity	Specificity	F1 score	MCC	Precision	Kappa	AUC
C5.0	0.8128	0.8409	0.7720	0.7707	0.6125	0.7694	0.6125	0.8635
RF	0.7928	0.8409	0.7230	0.7398	0.5683	0.7575	0.5679	0.8507
C4.5	0.7880	0.7886	0.7872	0.7516	0.5692	0.7191	0.5674	0.8390
CART	0.7832	0.8153	0.7365	0.7346	0.5514	0.7328	0.8153	0.8293

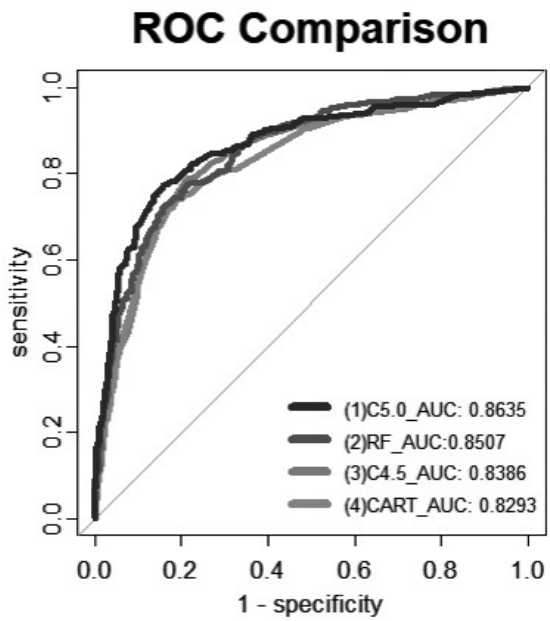


Figure 1. ROC (Receiver Operating Characteristic) Curves for recurrence in overall rectal cancer

4. Discussions

The purpose of this study was to identify important disease risk factors from the rectal cancer registry database, evaluate the prediction models developed for different factors, and analyze clinical features. Based on the result of data analysis, risk factors for recurrence are related to lifestyle, genetics, environmental factors, and the use of anti-cancer therapy, which are the most important risk factors in this study. The results of this study show that chemotherapy is the same as Chin et al. (2020) and Ting et al. (2020). Lifestyle (betel-nut chewing behavior, betel-nut chewing behavior, smoking behavior) and BMI are important factors that affect recurrence prediction, the same result as Chin and others. (2020). In comparison with Yu et al. (2022), the group of combined stages is the most important risk to predict recurrence, and primary site and tumor surgical margins are the same important risk factors as Ting et al. (2020)

Different from traditional regression model analysis, the risk factor analysis by machine learning methods used in our study could deal with the existence of complex interactions between different predictor variables. The identification of treatment related-factors that increase the risk of developing recurrence is of growing concern for cancer survivor surveillance and management.

5. Conclusion

The results of this study indicate that tumor size, regional lymph nodes examined, combine stage group, surgical margins of the primary site, radiotherapy, chemotherapy, carcinoembryonic antigen (CEA) lab value, carcinoembryonic antigen (CEA), Obstruction, and perforation were all risk factors for recurrence of rectal cancer survivors. These findings may be helpful in developing effective prevention and surveillance programs for high-risk rectal cancer survivors during their follow-up. Additionally, multi-institutional data from rectal cancer survivors provided authentic hospital-based cancer registry records and pathology data for this study. Real-world medical datasets usually have class imbalance problems, which are composed of a larger number of normal instances and a smaller number of abnormal ones. In the future, a more advanced approach can be taken to address this issue of class imbalance. Furthermore, future research can suggest incorporating more personalised patient

airiti

characteristics for further analysis.

Reference

1. (IARC) IAfRoC. Word Health Oraganization. 2020.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.
3. Ganeshan D, Nougaret S, Korngold E, Rauch GM, Moreno CC. Locally recurrent rectal cancer: what the radiologist should know. Abdominal Radiology. 2019;44(11):3709-25.
4. Manfredi S, Benhamiche AM, Meny B, Cheynel N, Rat P, Faivre J. Population-based study of factors influencing occurrence and prognosis of local recurrence after surgery for rectal cancer. Br J Surg. 2001;88(9):1221-7.
5. Asoglu O, Karanlik H, Muslumanoglu M, Igci A, Emek E, Ozmen V, et al. Prognostic and predictive factors after surgical treatment for locally recurrent rectal cancer: a single institute experience. Eur J Surg Oncol. 2007;33(10):1199-206.
6. Chen W, Chen M, Liao Z, Wang Y, Zhan Q, Cai G. Lymphatic vessel density as predictive marker for the local recurrence of rectal cancer. Dis Colon Rectum. 2009;52(3):513-9.
7. Stanzione A, Verde F, Romeo V, Boccadifuoco F, Mainenti PP, Maurea S. Radiomics and machine learning applications in rectal cancer: Current update and future perspectives. World J Gastroenterol. 2021;27(32):5306-21.
8. Quinlan JR. Induction of decision trees. Machine Learning. 1986;1(1):81-106.
9. Quinlan JR. C4. 5: programs for machine learning. 2014.
10. Achilonu OJ, Fabian J, Bebington B, Singh E, Eijkemans MJC, Musenge E. Predicting Colorectal Cancer Recurrence and Patient Survival Using Supervised Machine Learning Approach: A South African Population-Based Study. Front Public Health. 2021;9:694306.
11. Benson AB, Venook AP, Al-Hawary MM, Arain MA, Chen YJ, Ciombor KK, et al. NCCN Guidelines Insights: Rectal Cancer, Version 6.2020. J Natl Compr Canc Netw. 2020;18(7):806-15.
12. BREIMAN L. <Random Forests.pdf>. 2001.
13. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
14. Buckland M, Gey F. The relationship between Recall and Precision. Journal of the American Society for Information Science. 1994;45(1):12-9.
15. Center TTC. Clinical Guideline: Colorectal Cancer. 2021 [Available from: https://www.cancertaipei.tw/wp-content/uploads/2021/08/02%E5%A4%A7%E8%85%B8%E7%9B%B4%E8%85%B8%E7%99%8C.pdf.
16. Chen P-C, Yeh Y-M, Lin B-W, Chan R-H, Su P-F, Liu Y-C, et al. A Prediction Model for Tumor Recurrence in Stage II–III Colorectal Cancer Patients: From a Machine Learning Model to Genomic Profiling. Biomedicines. 2022;10(2):340.
17. Chen Y-C, & Chang, C.-C. Using Machine Learning Techniques to Predict Recurrent Breast Cancer in Taiwan. 2021.
18. Chiang C-J, You S-L, Chen C-J, Yang Y-W, Lo W-C, Lai M-S. Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. Japanese Journal of Clinical Oncology. 2015;45(3):291-6.
19. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.
20. Choi MS, Huh JW, Shin JK, Park YA, Cho YB, Kim HC, et al. Prognostic Factors and Treatment of Recurrence after Local Excision of Rectal Cancer. Yonsei Med J. 2021;62(12):1107-16.
21. Chun-Ju Chiang Y-WW及W-CL. Taiwan's Nationwide Cancer Registry System of 40 years: Past, present, and future. Journal of the Formosan Medical Association. 2019;118(5):856-8.
22. Farhat W, Azzaza M, Mizouni A, Ammar H, Ben Ltaifa M, Lagha S, et al. Factors predicting recurrence after curative resection for rectal cancer: a 16-year study. World J Surg Oncol. 2019;17(1):173.
23. Hospital CSMU. Clinical Guideline: Rectal Cancer. 2018.
24. Institutes NHR. Clinical Guideline: Colorectal Cancer. 2020.
25. Iseki K, Iseki C, Ikemiya Y, Fukiyama K. Risk of

airiti

developing end-stage renal disease in a cohort of mass screening. Kidney Int. 1996;49(3):800-5.
26. Kim M, Kim CS, Bae EH, Ma SK, Kim SW. Risk factors for peptic ulcer disease in patients with end-stage renal disease receiving dialysis. Kidney Res Clin Pract. 2019;38(1):81-9.
27. NCI, Health NCIatNIo. Dictionary Cancer Terma Recurrence 2022 [Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/recurrence.
28. POWERS DMW. Evaluation, from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.
29. Savargiv M, Masoumi B, Keyvanpour MR. A New Random Forest Algorithm Based on Learning Automata. Comput Intell Neurosci. 2021;2021:5572781.
30. Speybroeck N. Classification and regression trees. Int J Public Health. 2012;57(1):243-6.
31. Ting W-C, Lu Y-CA, Lu C-J, Cheewakriangkrai C, Chang C-C. RECURRENCE IMPACT OF PRIMARY SITE AND PATHOLOGIC STAGE IN PATIENTS DIAGNOSED WITH COLORECTAL CANCER. 品質學報. 2018;25(3):166-84.
32. Tseng CJ, Lu CJ, Chang CC, Chen GD, Cheewakriangkrai C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. Artif Intell Med. 2017;78:47-54.
33. Weller JM, Wu SC, Ferguson CW, Hawthorne VM. End-stage renal disease in Michigan. Incidence, underlying causes, prevalence, and modalities of treatment. Am J Nephrol. 1985;5(2):84-95.