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

doi:10.3233/SHTI210867

Predicting the Risk Factors of Second Primary Cancer in Patients with Hepatocellular Carcinoma

Hsien-Hua LIAO[a,b], Chi-Chang CHANG[c,d,1], Yu-Xiang WANG[c] and Chalong CHEEWAKRIANGKRAI[e]
[a]*Department of Surgery, Chung Shan Medical University Hospital, Taichung, Taiwan*
[b]*School of Medicine, Chung Shan Medical University, Taichung, Taiwan*
[c]*School of Medical Informatics, Chung Shan Medical University & IT office, Chung Shan Medical University Hospital, Taichung 40201, Taiwan*
[d]*Department of Information Management, Ming Chuan University, Taoyuan, Taiwan*
[e]*Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand*

Abstract. Screening for cancer and improved treatments have not only improved treatment outcomes and patient survival but have also led to an increase in the number of second primary cancers (SPCs). Hepatocellular carcinoma has been a common occurrence in Taiwan over the past decade. The mortality rate is second only to malignant tumors of lung cancer, and it also represents the fourth highest cancer medical expenditure. This study aimed to use machine learning to identify the risk factors for Hepatocellular carcinoma survivors. Of 378,445 datasets, including 15,251 from patients with SPCs, were collected; 18 predictive variables were considered risk factors for SPCs based on the physician panel discussion. The machine learning techniques employed included support vector machine, C5 decision tree, and random forest. SMOTE (Synthetic Minority Oversampling Technique) sampling method was used to resolve the imbalance problem. The results showed that the top 5 risk factors for SPCs were tumor size, clinical stage, surgery, total bilirubin, and BCLC Stage. The support vector machine method had the highest predicted accuracy (0.7673). The risk factors extracted from the classification models and association rules will be used to provide valuable information for HCC therapy.

Keywords. second primary cancer, hepatocellular carcinoma, machine learning techniques

1. Introduction

According to the latest data from the International Agency for Research on Cancer, hepatocellular carcinoma is the 7th most common cancer and the 3rd most common cancer. [1] Hepatocellular carcinoma (HCC) has been a common occurrence in Taiwan over the past decade. The mortality rate is second only to malignant tumors of lung cancer, and it also represents the fourth highest cancer medical expenditure, as shown in

[1] Corresponding Author: Chi-Chang Chang, School of Medical Informatics, Chung Shan Medical University & IT Office, Chung Shan Medical University Hospital, 110, Sec. 1, Chien-Kuo N. Rd., Taichung, Taiwan; E-Mail: changintw@gmail.com.

Table 1. [2] There are many treatment options for patients with hepatocellular carcinoma. Often, depending on the tumor, liver function, and physical situation. Screening for cancer and improved treatments have not only improved treatment outcomes and patient survival but have also led to an increase in the number of second primary cancers (SPCs). According to the Bureau of Health Promotion, Ministry of Health and Welfare, hepatocellular carcinoma was ranked second with regards to mortality rates of top 10 cancers and was only ranked below lung cancer, and ranked fourth with regards to medical expenditure, as shown in Table 2.

Table 1. 5-Year Survival Rates for Top 10 Cancers in Taiwan

Cancers / Year	1	2	3	4	5	6	7	8	9	10	All cancers
2014	96.9	82.3	61.1	92.8	59.0	80.4	93.1	97.7	92.2	60.1	77.0
2015	93.7	72.4	45.8	85.5	47.0	67.4	88.2	96.6	85.6	45.6	67.2
2016	90.5	65.5	37.3	79.1	39.6	61.3	84.6	95.7	80.0	38.8	61.3
2017	87.5	60.5	32.0	73.6	33.9	57.1	82.3	94.8	75.1	34.7	57.0
2018	85.2	56.9	28.6	68.5	29.9	53.4	80.3	94.1	71.3	31.6	53.9

Source: The Health Promotion Administration (2018).
Note: 1: breast cancer; 2: colorectal cancer; 3: lung cancer; 4: prostate cancer; 5: liver cancer; 6: oral cancer; 7: uterine cancer; 8: thyroid cancer; 9: skin cancer; 10: stomach cancer.

Table 2. Expenditure for top 10 cancers in Taiwan (unit: NTD/100,000)

Cancers / Year	1	2	3	4	5	6	7	8	9	10
2014	10,311	10,808	10,987	8,471	6,631	4,082	3,162	3,923	2,385	2,538
2015	10,923	11,326	11,138	8,811	6,737	4,298	3,518	4,309	2,446	2,576
2016	11,521	11,745	11,323	8,791	5,382	4,569	3,643	2,696	2,506	2,570
2017	13,217	12,829	12,468	9,666	6,000	4,907	4,158	3,015	2,767	2,755
2018	14,355	15,020	13,845	10,952	6,652	5,433	5,397	5,359	3,122	3,066

Source: The Health Promotion Administration (2018).
Note: 1: breast cancer; 2: lung cancer; 3: colorectal cancer; 4: liver cancer; 5: oral cancer; 6: leukemia; 7: prostate cancer; 8: non-Hodgkin's lymphoma; 9: Esophageal cancer; 10: stomach cancer.

Due to continuous improvements in screening, diagnosis and treatment, the survival rate of newly diagnosed cancer patients is increasing. [3] Clinical information regarding the SPCs patients with HCC is important, because it could explain the cause and may verify the necessity of related secondary cancer during patient follow-up. [4,5,6]

2. Methods

Figure 1 depicts the study framework. Herein, an in-depth analysis was performed based on the tumor size and clinical stage. The HCC dataset containing 378,445 valid records, was used in this study. There were 19 variables in the cancer registry and 18 predictive variables were selected by clinical experts and literature review (Table 3) C5 is a

modified iterative dichotomizer 3 (ID3) algorithm. Based on information theory and probability statistics, the greater the information gain and the higher the information entropy. RF (Random Forest) is based on the statistical learning theory and combines several individual classification trees. RF is a supervised machine learning algorithm that considers the unweighted classified votes. SVM (Support Vector Machine) is a machine learning algorithm based on the principle of structural risk minimization and is used to estimate function by minimizing the upper limit of generalization error.

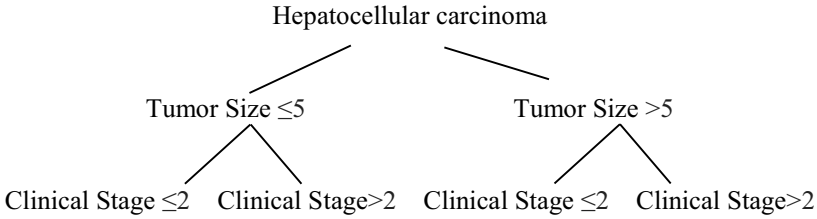

Figure 1. Study framework

Table 3. Important Variables and Coding in this study

Variable	Name	Definition of test data
X1	Age	≤65/>65
X2	Sex	Male/Female
X3	Grade	Well/Moderaterly/Poorly
X4	Tumor number	Single tumor/Multiple tumors
X5	Tumor Size	≤5/>5
X6	Clinical Stage	Stage I、II/Stage III
X7	Pathologic Stage	Stage I、II/Stage III
X8	BCLC Stage	Stage 0、A/Stage B/ Stage C、D
X9	Operation	YES/NO
X10	Surgical Margins	YES/NO
X11	BMI	<18.5/18.5-24/ >=24
X12	Alpha-Fetoprotein	≦400 ng/ml/ >400 ng/ml
X13	Liver Fibrosis	YES/NO
X14	eGFR	≦2 mg/dl/ >2 mg/dl
X15	Total bilirubin	≦2 mg/dl/ >2 mg/dl
X16	INR	<1.5/≧1.5
X17	HBV	YES/NO
X18	HCV	YES/NO
Y	SPCs	YES/NO

3. Results

In Figure 2, the analysis results after stratification of tumor size and clinical stage showed that in the tumor size (>5cm): the accuracy of <= IIb staging was the best with SVM (97.48%), the accuracy for > IIb staging was the best with SVM (90.03%). In the tumor size (<=5cm): the accuracy of <= IIb staging was the best with SVM (81.80%), and the accuracy of > IIb staging was the highest with SVM (83.75%).

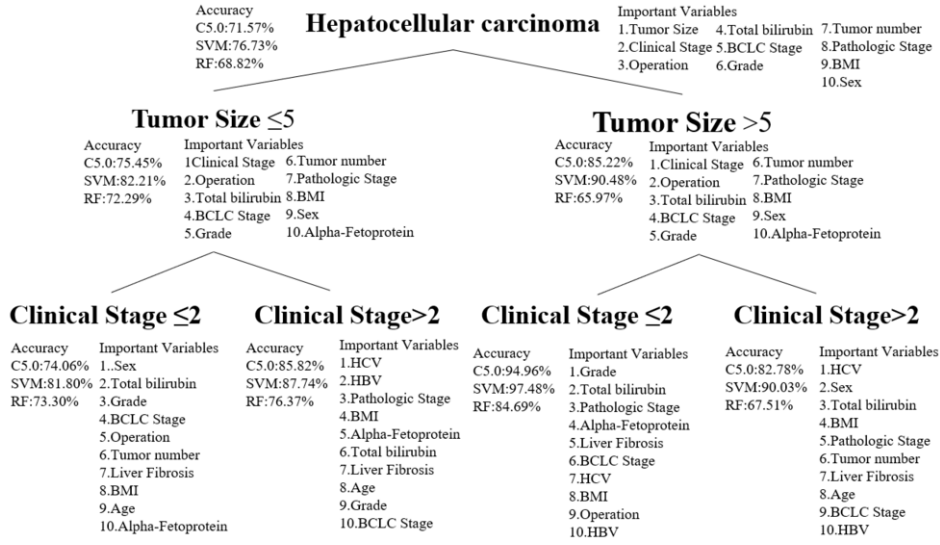

Figure 2. Analysis results after stratification of tumor size and clinical stage

4. Discussion

This study aimed to predict the risk factors of second primary Hepatocellular Carcinoma in survivors of Hepatocellular Carcinoma. The risk factors extracted from the classification models and association rules can be used to provide valuable information for HCC therapy. The models can be used to find the risk factors from database to provide valuable information for improving HCC patients.

Acknowledgements

This work was funded by grant no. CSH-2018-A-015 from the Chung-Shan Medical University Hospital.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021, 71(3):209-249.

[2] The Taiwan Cancer Registry, The age-adjusted incidence rates: 2018. Availiable on https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=269&pid=13498 [Access date: 2021.08.10]

[3] Sun CC, Chang CC, Multiple primary malignant neoplasms: Results from a 5-Year Retrospective Analysis in a Metropolitan Hospital. Formosan Journal of Surgery. 2017, 50: 209-214.

[4] Chin C, Ting WC, Chang CC, Zhang YX, Prediction of risk factors for Synchronous Colorectal Cancer in Patients with Colorectal Cancer, Journal of Quality. 2020, 27: 23-37.

[5] Ting WC, Chang HR, Chang CC, Lu CJ, Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Colorectal Cancer Survivors, Applied Sciences. 2020, 10: 1355.

[6] Chang CC, Chen SH, Developing a Novel Machine Learning-based Classification Scheme for Predicting SPCs in Women with Breast Cancer, Frontiers in Genetics. 2019, 10: 848.