RESEARCH ARTICLE

Prognostic Evaluation of Categorical Platelet-based Indices Using Clustering Methods Based on the Monte Carlo Comparison for Hepatocellular Carcinoma

Pi Guo[1], Shun-Li Shen[2], Qin Zhang[3], Fang-Fang Zeng[1], Wang-Jian Zhang[1], Xiao-Min Hu[1], Ding-Mei Zhang[1], Bao-Gang Peng[2], Yuan-Tao Hao[1]*

Abstract

Objectives: To evaluate the performance of clustering methods used in the prognostic assessment of categorical clinical data for hepatocellular carcinoma (HCC) patients in China, and establish a predictable prognostic nomogram for clinical decisions. Materials and Methods: A total of 332 newly diagnosed HCC patients treated with hepatic resection during 2006-2009 were enrolled. Patients were regularly followed up at outpatient clinics. Clustering methods including the Average linkage, k-modes, fuzzy k-modes, PAM, CLARA, protocluster, and ROCK were compared by Monte Carlo simulation, and the optimal method was applied to investigate the clustering pattern of the indices including platelet count, platelet/lymphocyte ratio (PLR) and serum aspartate aminotransferase activity/platelet count ratio index (APRI). Then the clustering variable, age group, tumor size, number of tumor and vascular invasion were studied in a multivariable Cox regression model. A prognostic nomogram was constructed for clinical decisions. Results: The ROCK was best in both the overlapping and non-overlapping cases performed to assess the prognostic value of platelet-based indices. Patients with categorical platelet-based indices significantly split across two clusters, and those with high values, had a high risk of HCC recurrence (hazard ratio [HR] 1.42, 95% CI 1.09-1.86; $p<0.01$). Tumor size, number of tumor and blood vessel invasion were also associated with high risk of HCC recurrence (all $p<0.01$). The nomogram well predicted HCC patient survival at 3 and 5 years. Conclusions: A cluster of platelet-based indices combined with other clinical covariates could be used for prognosis evaluation in HCC.

Keywords: Clustering method - Categorical data - Hepatocellular carcinoma - Monte Carlo - Prognostic

Asian Pac J Cancer Prev, **15 (14)**, 5721-5727

Introduction

Hepatocellular carcinoma (HCC) is a leading cause of cancer-related death worldwide, and the burden of this devastating cancer is expected to increase further in coming years (Nguyen et al., 2009; Venook et al., 2010). In Asian region, the incidence of HCC exceeds 30 cases per 100, 000 residents annually, which is due to the high prevalence of chronic viral hepatitis, mainly chronic hepatitis B (Teo et al., 2002; Gao et al., 2012; Guo et al., 2012).

Although many factors such as tumor size, number of tumor, vascular invasion and resection margin status are associated with the prognosis of HCC resection, it is necessary to find a potential prognostic cluster that is available before surgery, because it can be used to predict and assess the prognostic status for HCC patients who received tumor resection. In addition, the preoprative platelet count and serum aspartate aminotransferase activity/platelet count ratio index (APRI) have shown to be independent prognostic factors for patients after resection of HCC (Ichikawa et al., 2009; Maithel et al., 2011). Although the single APRI or platelet count indicator presents obvious prognostic value for HCC, the prognostic value of platelet-based indices as a panel has not been studied. It will be meaningful to evaluate the prognostic value of this panel of platelet-based indices for HCC.

The main purpose of this study is to evaluate the prognostic value of a panel of categorical platelet-based indices including platelet count, platelet/lymphocyte ratio (PLR) and APRI in HCC after hepatic resection using a clustering method. First, we will determine which clustering method is suitable for analyzing categorical prognostic factors. Second, after detecting the clustering patterns, we will establish a predictable model for evaluating the prognosis of HCC in clinical practice. On the basis of these two points, a Monte Carlo simulation will be performed to compare the performance of the clustering

*[1]Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-sen University, [2]Department of Hepatobiliary Surgery, the First Affiliated Hospital, Sun Yat-sen University, Guangzhou, [3]Good Clinical Practice Office, Cancer Hospital of Shantou University Medical College, Shantou, Guangdong, China *For correspondence: haoyt@mail.sysu.edu.cn*

methods for categorical data and the most robust clustering method will be selected. Besides that, multivariable analysis will be conducted to investigate the significant prognostic factors, and a predictable nomogram for HCC after resection will be constructed for clinical decisions.

Materials and Methods

Patients, treatments and follow-up

This study enrolled a total of 332 newly diagnosed HCC patients treated with hepatic resection in the First Affiliated Hospital of Sun Yat-sen University during 2006-2009. A confirmed diagnosis of HCC was made through histopathological examination of the specimen. Patients with coexistent hematologic disorders, and mixed hepatocellular carcinoma and cholangiocarcinoma were excluded. Every patient signed an informed consent form before enrolling in the study, and all the procedures were performed in accordance with the requirements of the medical research ethics. The enrolled subjects were more than 18 years of age with complete clinical and laboratory data. Patients with intent to cure were treated with hepatectomy, and regularly followed up at outpatient clinics every 3 months for the first 2 years, every 6 months for the next 3 years, and once a year thereafter. At each follow-up, patients received a physical examination, liver ultrasound and other corresponding solutions if needed. In addition, abdominal CT scans were given every 6-12 months or when recurrence was suspected.

To evaluate the prognostic value of platelet-based indices including platelet count, PLR and APRI in HCC after hepatic resection, we obtained the original laboratory data about these three indices for each patient. The three indices were then calculated to stand for a platelet-based prognostic cluster of HCC recurrence. The disease-free survival (DFS) was calculated from the date of surgery to the date of HCC recurrence. Due to no validated cutoff value existed for both PLR and APRI before the analysis, initially the receiver operating characteristic curve analysis (Zweig et al., 1993) was used to identify the most appropriate cutoff points of both the indices to classify patients into high-risk and low-risk groups of HCC recurrence. Thus the cut-off values of 115 and 0.62 corresponded to the maximum joint sensitivity and specificity for PLR and APRI were determined. Therefore, the categorical indices including the platelet count (<300 mm3, ≥300 mm^3), the PLR (<115, ≥115) and the APRI (<0.62, ≥0.62) were constructed.

Statistical analysis

Evaluation of prognostic factors: The panel of platelet-based indices including platelet count, PLR and APRI were integrated as a whole into the proposed clustering method to assess the prognostic value for HCC, acting as a prognostic cluster rather than a single indicator in this study. The cluster center representing by the most frequent category for each indicator was characterized according to the indicator distribution in each cluster. Covariates including the age group, tumor size, number of tumor, vascular invasion were analyzed. Estimates of the probability of DFS for different clusters were calculated with the Kaplan-Meier method and compared using the log-rank test. Multivariable analysis was conducted with stepwise Cox proportional hazards regression to investigate the significant factors for HCC prognosis and a nomogram (Derici et al., 2012) was constructed for clinical decisions based on this multivariable Cox model. A calibration plot was used to graphically assess the agreement between the predicted probabilities and observed outcomes. For a prediction model with good calibration, the curve virtually followed a 45-degree slope. For all analyses, a 2-sided $p<0.05$ was considered significant.

Clustering methods for categorical data: To cluster the platelet-based prognostic factors for HCC, the representative methods for categorical data including the Average linkage (Everitt et al., 2001), k-modes (Huang et al., 1998), fuzzy k-modes (Huang et al., 1999), CLustering LARge Applications (CLARA) (Wei et al., 2000), Partitioning Around Medoids (PAM) (Kaufman et al., 1987), RObust Clustering using linKs (ROCK) (Guha et al., 1999), protocluster (Bien et al., 2011) were selected in this study. Monte Carlo simulation was performed to compare the clustering methods for determining the most robust method for our study.

The Average linkage (Everitt et al., 2001) starts with each object (a sample or variable) as a separate cluster. The dissimilarity measures of

$$d(C_i,C_j)=\frac{\sum_{X_q\in C_i}\sum_{X_p\in C_j} d(X_q,X_p)}{n_i n_j}$$

between clusters are calculated. In the above formula, the

$$d(X_q,X_p)$$

dissimilarity measure between the elements of X_q and X_p. Based on the dissimilarity measure, the two most similar clusters are merged. The merging step is repeated iteratively till the desirable number of clusters is obtained.

The k-modes (Huang et al., 1998) method is an extension of k-means for clustering categorical data. It uses a dissimilarity measure, modes instead of means, to investigate the proximity of clusters. This method executes as follows: (i) k initial modes are generated and the dissimilarity measure

$$d(x_{i,j},q_{l,j})=\begin{cases}0 \ if \ x_{i,j}=q_{l,j}\\ 1 \ if \ x_{i,j}=q_{l,j}\end{cases}$$

($x_{i,j}$ stands for the observation of the domain of each categorical variable A_j and $q_{l,j}$ for the modes of the cluster l) is calculated, where $x_{i,j}$ stands for the observation of the domain of each categorical variable A_j and $q_{l,j}$ for the modes of the cluster l . Each object is compared to the modes and is assigned to the most similar group; (ii) after allocating, the modes are updated and the update-step is repeated iteratively till there is no reallocation of objects needed. The fuzzy k-modes (Huang et al., 1999) method acts as an extension of the k-modes based on the fuzzy theory, and the fuzzy parameters and the degree of membership of the observations to each cluster are estimated. These two parameters are used as weights for updating the k modes.

The CLARA (Wei et al., 2000) extends the k-medoids approach for a large number of objects. The CLARA initially calculates the optimal medoids using the PAM method based on a small set of random samples drawn from the whole dataset. The quality of resulting medoids is measured by the cost function:

$$Cost(M,D) = \frac{\sum_{i=1}^{n} d(O_i, rep(M_i, O_i))}{n}$$

where M is a set of medoids, $d(O_i, O_j)$ is the dissimilarity between objects, $rep(M, O_i)$ and returns a medoid in M which is closest to O_i n is the number of clusters.

The PAM (Kaufman et al., 1987) method is similar to the k-means algorithm in terms of partitioning and minimizing the overall dissimilarity between the representants of each cluster and its members, but the PAM works with medoids instead of centroids. Generally the PAM starts with choosing k entities to become the mediods and then calculates the dissimilarity measurement (e.g., the metric of euclidean or manhattan distance) between the mediods. By iteratively allocating every object to its nearest medoid, the mediod of each cluster is updated till all the medoids remain unchanged.

The ROCK (Guha et al., 1999) clustering method is carried out based on the measure of links instead of distance between cluster objects. Let x_q and x_r be two observations. The ROCK uses the $link(x_q, x_r)$ to represent the number of neighbors the two observations have in common: a higher value of the $link(x_q, x_r)$ suggests a higher probability of x_q and x_r belonging to the same group. Initially the ROCK merthod computes the number of links between objects, and then merges the objects into clusters till no links present or the predefined number of clusters is achieved.

As one type of agglomerative hierarchical clustering methods, the protocluster (Bien et al., 2011) generates a hierarchical structure from dataset depending on a minimax linkage rather than a complete linkage, and naturally associates a prototype chosen from the original dataset with every interior node of the dendrogram. For any point x and cluster C, the formula

$d_{max}(x, C) = max\, d(x, x')$ defines the distance to the farthest point in C from x. The minimax radius of the cluster C, $r(C) = min\, d_{max}(x, C)$, is defined to find the prototype point from which all points x C in C are as close as possible. The minimax linkage d (G, H)=r (G H) denotes the distance between clusters G and H, and the allocation of objects is iteratively implemented based on the linkage.

Monte Carlo simulation: We referred to the Monte Carlo simulation scheme established by Mingoti et al. (Mingoti et al., 2012), and extended it in this study. In this simulation, different degrees of overlapping among clusters (Degree 1, 2, 3), different number of clusters (k=2, 3, 5), categorical variables (m=2, 4) and categories of each variable (c=2, 3, 5) were considered. Therefore, a total of 60 population structures of clusters were simulated. Every 50 observations were randomly generated according to the uniform distribution for each cluster (Table 1) presents the simulated population structures by cases (non-overlapping cases: 1-4; overlapping cases: 5-8). The simulation study aimed to assess the changes in the performance of clustering methods in different situations in terms of the overlapping degree, and number of clusters, variables and categories, and then identify the most robust clustering method.

Table 1. Monte Carlo Simulation Structure Generated According to Different Number of Clusters, Variables, Categories and Degree of Overlapping

Case	k	m	X1	X2	X3	X4
Degree 1 non-overlapping in the first variable						
1 (9 situations)	2	2	2	2, 3, 5		
	3	2	3	2, 3, 5		
	5	2	5	2, 3, 5		
2 (9 situations)	2	4	2	2, 3, 5	3, 5, 2	5, 2, 3
	3	4	3	2, 3, 5	3, 5, 2	5, 2, 3
	5	4	5	2, 3, 5	3, 5, 2	5, 2, 3
Degree 2 non-overlapping in the first and second variables						
3 (9 situations)	2	4	2	2	2, 3, 5	3, 5, 2
	3	4	3	3	2, 3, 5	3, 5, 2
	5	4	5	5	2, 3, 5	3, 5, 2
Degree 3 non-overlapping in the first, second and third variables						
4 (3 situations)	5	4	5	5	5	2
	5	4	5	5	5	3
	5	4	5	5	5	5
Degree 4 overlapping in the first variable						
5 (9 situations)	2	2	2	2, 3, 5		
	3	2	3	2, 3, 5		
	5	2	5	2, 3, 5		`
6 (9 situations)	2	4	2	2, 3, 5	3, 5, 2	5, 2, 3
	3	4	3	2, 3, 5	3, 5, 2	5, 2, 3
	5	4	5	2, 3, 5	3, 5, 2	5, 2, 3
Degree 5 overlapping in the first and second variables						
7 (9 situations)	2	4	2	2	2, 3, 5	3, 5, 2
	3	4	3	3	2, 3, 5	3, 5, 2
	5	4	5	5	2, 3, 5	3, 5, 2
Degree 6 overlapping in the first, second and third variables						
8 (3 situations)	5	4	5	5	5	2
	5	4	5	5	5	3
	5	4	5	5	5	5

Non-overlapping clusters were generated in cases 1 and 2 (Degree 1: non-overlapping in the first variable), case 3 (Degree 2: non-overlapping in the first two variables) and case 4 (Degree 3: non-overlapping in the first three variables). Overlapping clusters were generated in cases 5 and 6 (Degree 4: overlapping in the first variable), case 7 (Degree 5: overlapping in the first two variables) and case 8 (Degree 6: overlapping in the first three variables). For example, in case 1, k=2, m=2 and c=2, and suppose that {A1, A2} are the categories of the first variable. The two clusters were constructed based on the following steps: for the first variable the category {A1} was assigned to all the observations of the first cluster and the category {A2} for all observations of the second cluster. This step generated non-overlapping observations in the first variable between the two clusters. Each category of the second variable observations was randomly generated for both clusters. For cases 2, 3 and 4, the similar procedure was followed to generate uniform random observations for different situations. For overlapping cases 5-8, all the categories of the overlapping variables were proportionally generated and the simulation procedure ensured the same frequency of each category for each cluster. For example, in case 7, k=2 and m=4, the first two variables X1 and X2 were built

with 2 categories, the third X3 with (2, 3, and 5) categories, and X4 with (3, 5, and 2) categories, respectively. For the first two variables, the simulation procedure ensured the proportionality of each category of the overlapping variables accounting for 50% from all observations in the respective cluster. The samples for the other two variables were generated at random. For the other overlapping cases, the similar procedure was followed.

For each run of the Monte Carlo simulation, the pre-specified two clusters were used to in the execution of the clustering methods, and the initial random seed 201403 for program execution was used.

Results

Monte Carlo Comparison

The average prediction accuracy of each clustering method based on the overlapping degree and the number of clusters is shown in (Table 2). The simulated results were grouped into the "non-overlapping" cases (Degree 1-3) and "overlapping" cases (Degree 4-6) according to the number of clusters (k=2, 3, 5). The overall mean accuracy for all clustering algorithms was also calculated. It is shown that in average, for the non-overlapping group, the Average-linkage and ROCK were the best algorithms (overall means over 99%) compared to k-modes, fuzzy k-modes, CLARA, PAM, protocluster (overall means between 39% and 91%). For the overlapping group, the ROCK was the best with the overall mean accuracy of around 51.1% larger than the other clustering methods. The efficiency loss (Effloss) measured by the difference between the "non-overlapping" and "overlapping" average accuracy showed that Average linkage and fuzzy k-modes were the most affected by overlapping (average Effloss standed at approximately 62% and 54%) but the average Effloss rates for CLARA, PAM and protocluster methods were similar, standing at around 35%. The ROCK had the medium average Effloss rate of 48.8% among all the algorithms. The average Effloss for k-modes was the smallest, but its accuracy for both the non-overlapping and overlapping situations were quite small.

The average accuracy for m=2 and m=4 categorical variables was compared in (Table 3). Comparing to the results in (Table 2), the increased number of categorical variables had less impact on the accuracy than the increased number of clusters. For each clustering algorithm, the larger is the degree of the overlapping the smaller are the average accuracy values, as shown in (Figure 1).

In the non-overlapping cases, three algorithms including Average linkage, fuzzy k-modes and ROCK had the best predictive performance. In the overlapping cases, the ROCK outperformed the other methods in terms of

Table 2. Average Accuracy of Each Clustering Method According to the Overlapping Degree and the Number of Clusters

Algorithm	Non-overlapping k=2	k=3	k=5	overall mean
Average linkage	1	1	0.9993	0.9998
k-modes	0.562	0.3777	0.2208	0.3869
Fuzzy k-modes	0.9938	0.9229	0.8211	0.9126
CLARA	0.7616	0.752	0.7784	0.764
PAM	0.7536	0.7357	0.7633	0.7509
ROCK	1	0.9977	1	0.9992
protocluster	1	0.6667	0.4	0.6889
	Overlapping k=2	k=3	k=5	overall mean
Average linkage	0.5662	0.3632	0.2119	0.3804
k-modes	0.5164	0.3464	0.2074	0.3567
Fuzzy k-modes	0.5466	0.3456	0.2116	0.3679
CLARA	0.5179	0.3688	0.3329	0.4065
PAM	0.5179	0.3719	0.3402	0.41
ROCK	0.6921	0.516	0.325	0.511
protocluster	0.6009	0.263	0.1052	0.323
	Difference (Effloss) k=2	k=3	k=5	overall mean
Average linkage	0.4338	0.6368	0.7874	0.6193
k-modes	0.0456	0.0313	0.0134	0.0301
Fuzzy k-modes	0.4473	0.5773	0.6095	0.5447
CLARA	0.2437	0.3832	0.4455	0.3575
PAM	0.2357	0.3639	0.4232	0.3409
ROCK	0.3079	0.4817	0.675	0.4882
protocluster	0.3991	0.4037	0.2948	0.3659

Table 3. Average Accuracy of each Clustering Method According to the Overlapping Degree and the Number of Categorical Variables

Algorithm	Non-overlapping m=2	m=4	overall mean
Average linkage	0.9991	1	0.9995
k-modes	0.405	0.3554	0.3802
Fuzzy k-modes	0.7041	0.9889	0.8465
CLARA	0.674	0.8046	0.7393
PAM	0.6892	0.7791	0.7342
ROCK	1	0.999	0.9995
protocluster	0.6888	0.6476	0.6682
	Overlapping m=2	m=4	overall mean
Average linkage	0.3489	0.3699	0.3594
k-modes	0.3421	0.3417	0.3419
Fuzzy k-modes	0.3492	0.3536	0.3514
CLARA	0.3513	0.4197	0.3855
PAM	0.3515	0.4251	0.3883
ROCK	0.5958	0.4482	0.522
protocluster	0.2851	0.3082	0.2967
	Difference (Effloss) m=2	m=4	overall mean
Average linkage	0.6502	0.6301	0.6402
k-modes	0.063	0.0137	0.0383
Fuzzy k-modes	0.3549	0.6353	0.4951
CLARA	0.3227	0.3849	0.3538
PAM	0.3378	0.354	0.3459
ROCK	0.4042	0.5508	0.4775
protocluster	0.4037	0.3394	0.3716

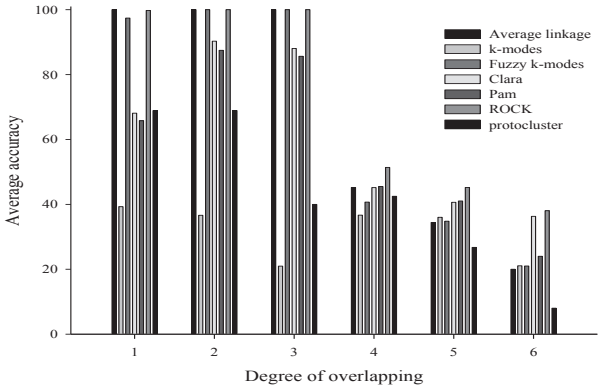

Figure 1. Average Accuracy for All Clustering Methods According to the Overlapping Degree (m=4)

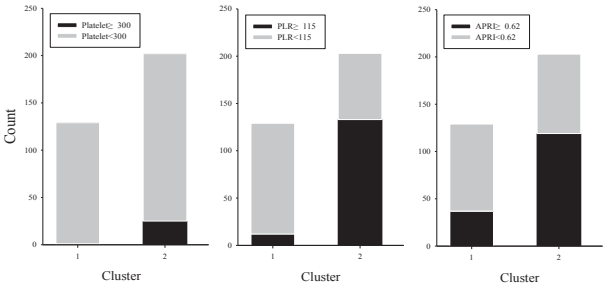

Figure 2. Distribution of three Categorical Platelet-Based Indices in Two Clusters Calculated by the ROCK Method

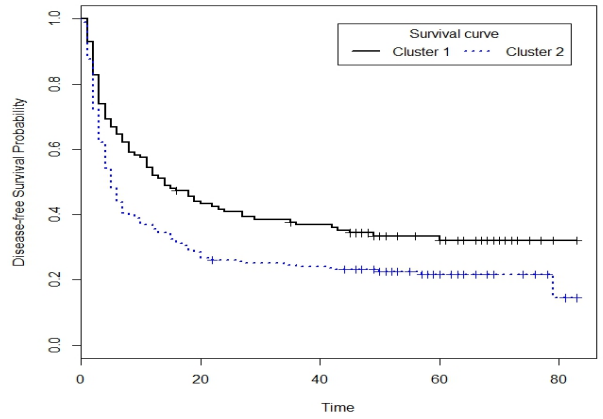

Figure 3. Disease-Free Survival Probability Curves of Two Clusters of Patients Based on the ROCK Clustering Method

prediction accuracy. Taking both the average accuracy and the Effloss rate, the ROCK was the best method according to our simulation. Therefore, the ROCK method was chosen to assess the prognostic value of the platelet-based indices for HCC in this study.

Prognostic value of platelet-based indices

The panel of categorical platelet-based indices including the platelet count, PLR and APRI was clustered by using the ROCK method, and two clusters were generated to assess the prognostic value for HCC. This panel of indices worked as a prognostic cluster rather than a single indicator to show its joint effect. The cluster center represented by the most frequent category for each indicator was characterized according to the indicator distribution in each cluster, as shown in (Figure 2). It can

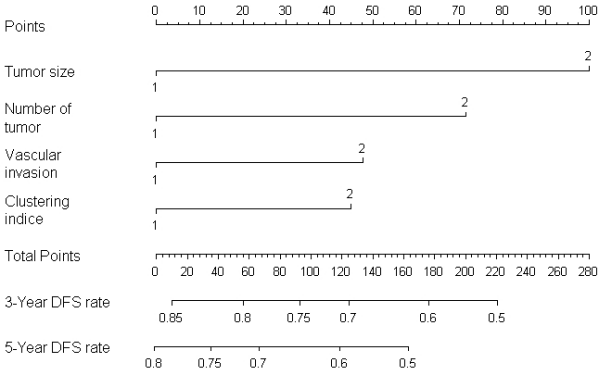

Figure 4. Prognostic Nomogram of Predicting 3- and 5-Year Survival Probability for HCC Patients after Resection Based on The Constructed Multivariate Cox Regression Model

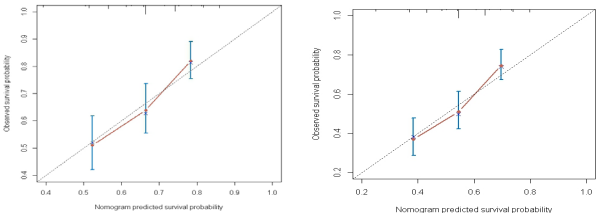

Figure 5. The Calibration Curve for Predicting Patient Survival at 3 Years (A) and 5 Years (B). The nomogram-predicted probability of the overall survival is plotted on the x-axis. The actual observed survival probability is plotted on the y-axis

be seen that patients with categorical platelet-based indices significantly split across two clusters. Patients with high values of indices came into being Cluster 2, especially for PLR≥115, APRI≥0.62.

(Figure 3) showed the DFS probability for the two clustered patients using three platelet-based indices according to the ROCK method. The DFS of patients with lower values of platelet-based indices, especially for PLR<115 and APRI<0.62, were significantly better compared to patients with the elevated values, suggesting that high values of the platelet-based cluster were associated with poor prognosis for HCC (the log-rank test p=0.0029). Patients with high values of platelet-based measures in Cluster 2 had high risk of HCC recurrence (hazard ratio [HR] 1.42, 95% CI 1.09-1.86; p<0.01) according to the Multivariate Stepwise Cox regression model. The tumor size, number of tumor and blood vessel invasion were associated with high risk of HCC recurrence (HR 2.01, 95% CI 1.42-2.85; HR 1.64, 95% CI 1.22-2.19; HR 1.38, 95% CI 1.04-1.82; respectively).

Nomogram for predicting HCC survival

To provide clinicians with a quantitative method to predict a patient's probability of HCC recurrence, we constructed a nomogram that integrated the platelet-based cluster and other covariates (Figure 4). The contribution of each covariate to the total score in the nomogram plot can be visually appreciated. To use the nomogram in (Figure 4), locate patient's variable on the corresponding axis; draw a line to the points axis, sum the points, and draw a line from the total points axis to the 3- and 5-year DFS

probability axis to get the predicted survival rate. (Figure 5) showed the calibration plots of each model in terms of the agreement between the predicted and the observed survival probabilities. Model performance was evaluated, relative to the 45-degree line, which represented perfect prediction. Compared with an ideal model, the established nomogram did well for predicting patient survival at 3 and 5years.

Discussion

In this present study, the ROCK clustering method was shown to be the most robust among the selected algorithms based on the Monte Carlo simulation when the average accuracy and Effloss rate were considered together. Hence, the ROCK method was performed to assess the prognostic value of the platelet-based indices as a whole rather than a single variable for HCC after resection. Patients with higher values of platelet-based indices clustered together, especially for PLR≥115 and APRI≥0.62. The result indicated that an elevated value of platelet-based set was associated with poor prognosis for HCC after resection. To better guide the clinical practice, a prognostic nomogram with high predictable performance was established. The analysis showed that the nomogram did well for predicting patient survival at 3 and 5years for HCC after resection.

Previous studies showed that the increased platelet count was associated with poor prognosis in nasopharyngeal carcinoma (Gao et al., 2013), gastric cancer (Hwang et al., 2012), colorectal cancer (Lin et al., 2012), and endometrial carcinomas (Gorelick et al., 2009). Our study found that the elevated values of platelet-based indices predicted poor survival for HCC patients after resection, which was consistent with the findings in other cancers. Besides that, the indicators of PLR and APRI were also used to predict the prognosis for patients with epithelial ovarian cancer and chronic hepatitis in other studies (Lin et al., 2011; Raungkaewmanee et al., 2012). In another study published on APJCP in 2014, elevated PLR was reported as useful biomarkers for diagnosis in lung cancer patients before treatment (Kemal et al., 2014). The platelet-based indices have been shown to be robust discriminative factors for predicting both recurrence and survival of cancer patients. Although the single indicator of platelet count, PLR or APRI presented significant prognostic value for different kinds of cancer, the panel of platelet-based indices as a whole and its prognostic value was not reported in previous studies. We evaluated the prognostic value of this panel of categorical platelet-based indices for HCC using clustering method, and found that patients with elevated values of platelet-based factors congregated in a cluster. The panel of indices acted as a prognostic cluster rather than a single indicator to show its joint effect on HCC recurrence.

Clustering analysis is a main technique of data preprocessing (Mukti et al., 2013). As a kind of unsupervised learning approach, clustering analysis is the task of grouping a set of objects in the same cluster where the objects are more similar to each other than to those in other clusters. Data clustering algorithms have been used to analyze the prognostic factors for survival of cancer patients. Generally, the survival data of cancer patients contain much categorical prognostic information such as the tumor grade, metastasis status, complications, surgical margin status and so on. To investigate the survival characteristics of cancer patients, clustering methods for clinical data, especially for the categorical information, could be applied to find some interesting patterns hidden in the data. In addition, the performance of clustering methods to analyze categorical prognostic factors for cancer patients should also be comprehensively evaluated and compared. By means of Monte Carlo simulation, we showed that overlapping was the factor with the major impact on the accuracy of all the clustering methods and the impact of the increased number of clusters on the performance of the methods was large.

As a quantitative method to predict a patient's probability of an event, such as death or recurrence, prognostic nomogram provided an efficient way to facilitate patient counseling and individualism management of cancer patients (Iasonos et al., 2008; Zhang et al., 2013; Koca et al., 2014). Nomograms are widely used, primarily because of their ability to reduce statistical predictive models into a single numerical estimate of the probability of death or disease recurrence. As Iasonos et al. (Iasonos et al., 2008) pointed out, the nomogram construction mainly included the following steps: identify the source population, define the outcome, identify potential covariates, constructing the nomogram, validating the constructed model, interpret the final nomogram and apply the nomogram. We developed a predictable nomogram for clinical use in predicting patient survival at 3 and 5 years for HCC after resection. The predictive accuracy and discriminative ability of the nomogram were determined by calibration curve in this study. However, our current study is limited because it is retrospective, with limited sample size and the Han people just studied. Clearly, our results should be further validated by prospective study in multicentre clinical trials as well as in different racial groups.

In summary, our study showed that the platelet-based cluster established by the ROCK method was significantly associated with the prognostic value for HCC. Patients with the elevated platelet count, PLR and APRI presented poor survival for HCC after resection. The prognostic nomogram constructed in this study could be used in clinical practice.

Acknowledgements

We thank Mr. Zhenli Zhu for his great effort in assisting the manuscript revision. We also thank Mr. Zan Ding for his professional suggestions in the statistical programming. The authors declare that no competing interests exist

References

Bien J, Tibshirani R (2011). Hierarchical clustering with prototypes via minimax linkage. *J Am Stat Assoc*, **106**, 1075-84.

Derici S, Sevinc A, Harmancioglu O, et al (2012). Validation of three breast cancer nomograms and a new formula for predicting non-sentinel lymph node status. *Asian Pac J Cancer Prev,* **13**, 6181-5.

Everitt BS, Landau S, Leese M, et al (2001). Hierarchical Clustering. Cluster Analysis, 5th Edition, 71-110.

Gao J, Xie L, Yang WS, et al (2012). Risk factors of hepatocellular carcinoma--current status and perspectives. *Asian Pac J Cancer Prev,* **13**, 743-52.

Gao J, Zhang HY, Xia YF (2013). Increased platelet count is an indicator of metastasis in patients with nasopharyngeal carcinoma. *Tumour Biol*, **34**, 39-45.

Gorelick C, Andikyan V, Mack M, et al (2009). Prognostic significance of preoperative thrombocytosis in patients with endometrial carcinoma in an inner-city population. *Int J Gynecol Cancer,* **19**, 1384-9.

Guha S, Rastogi R, Shim K (1999). ROCK: A robust clustering algorithm for categorical attributes. Paper read at Data Engineering, 1999. Proceedings., 15th International Conference on.

Guo P, Huang Z, Yu P, et al (2012). Trends in cancer mortality in China: an update. *Ann Oncol,* **23**, 2755-62.

Huang Z (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min Knowl Discov,* **2**, 283-304.

Huang Z, Ng MK (1999). A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, **7**, 446-52.

Hwang SG, Kim KM, Cheong JH, et al (2012). Impact of pretreatment thrombocytosis on blood-borne metastasis and prognosis of gastric cancer. *Eur J Surg Oncol,* **38**, 562-7.

Iasonos A, Schrag D, Raj GV et al (2008). How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol,* **26**, 1364-70.

Ichikawa T, Uenishi T, Takemura S, et al (2009). A simple, noninvasively determined index predicting hepatic failure following liver resection for hepatocellular carcinoma. *J Hepatobiliary Pancreat Surg* **16**, 42-8.

Kaufman L, Rousseeuw P (1987). Clustering by means of medoids. Statistical Data Analysis Based on the L1 Norm and Related Methods. *North-Holland*, p405-16.

Kemal Y, Yucel I, Ekiz K, et al (2014). Elevated serum neutrophil to lymphocyte and platelet to lymphocyte ratios could be useful in lung cancer diagnosis. *Asian Pac J Cancer Prev* **15**, 2651-4.

Koca B, Kuru B, Ozen N, et al (2014). A breast cancer nomogram for prediction of non-sentinel node metastasis-validation of fourteen existing models. *Asian Pac J Cancer Prev,* **15**, 1481.

Lin MS, Huang JX, Zhu J, et al (2012). Elevation of platelet count in patients with colorectal cancer predicts tendency to metastases and poor prognosis. *Hepatogastroenterology* **59**, 1687-90.

Lin ZH, Xin YN, Dong QJ, et al (2011). Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: An updated meta-analysis. *Hepatology,* **53**, 726-36.

Maithel SK, Kneuertz, PJ (2011). Importance of low preoperative platelet count in selecting patients for resection of hepatocellular carcinoma: a multi-institutional analysis. *J Am Coll Surg*, **212**, 638-48.

Mingoti SA, Matos RA (2012). Clustering algorithms for categorical data: a Monte Carlo study. *Int J Stat Appl,* **2**, 24-32.

Mukti MZR, Ahmed F (2013). Early detection of lung cancer risk using data mining. *Asian Pac J Cancer Prev,* **14**, 595-8.

Nguyen V, Law M, Dore G (2009). Hepatitis B-related hepatocellular carcinoma: epidemiological characteristics and disease burden. *J Viral Hepat* **16**, 453-63.

Raungkaewmanee S, Tangjitgamol S, Manusirivithaya S, et al (2012). Platelet to lymphocyte ratio as a prognostic factor for epithelial ovarian cancer. *J Gynecol Oncol,* **23**, 265-73.

Teo E, Fock K (2002). Hepatocellular carcinoma: an Asian perspective. *J Dig Dis,* **19**, 263-8.

Venook AP, Papandreou C, Furuse J, et al (2010). The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. *Oncologist,* **15**, 5-13.

Wei CP, Lee YH, Hsu CM, et al (2000). Empirical comparison of fast clustering algorithms for large data sets. Paper read at System Sciences. Proceedings of the 33rd Annual Hawaii International Conference on.

Zhang JX, Song W, Chen ZH, et al (2013). Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol,* **14**, 1295-306.

Zweig MH, Campbell G (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* **39**, 561-77.