Clinical Biochemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Improve discrimination power of serum markers for diagnosis of cholangiocarcinoma using data mining-based approach

Sirorat Pattanapairoj [a,1], Atit Silsirivanit [b,e,1], Kanha Muisuk [d,e], Wunchana Seubwai [d,e], Ubon Cha'on [c,e], Kulthida Vaeteewoottacharn [c,e], Kanlayanee Sawanyawisuth [c,e], Danaipong Chetchotsak [a,b,*], Sopit Wongkham [c,e,**]

[a] Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand
[b] System Modeling for Industry Research Group, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand
[c] Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand
[d] Department of Forensic Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand
[e] Liver Fluke and Cholangiocarcinoma Research Center, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

ARTICLE INFO

Article history:
Received 9 January 2015
Received in revised form 26 March 2015
Accepted 30 March 2015
Available online xxxx

Keywords:
Tumor markers
Combined analysis
C4.5 decision tree
Single neural network
Hepatobiliary
ALP
Mucin

ABSTRACT

Objective: Cholangiocarcinoma (CCA) is usually fatal because of the absence of tests for early detection and lack of effective therapy. Tumor markers with adequate diagnostic values are of clinical significance. This study is aimed to improve the diagnostic power of serum markers using the computational data mining technique to develop a combined diagnostic model that yielded the best diagnostic values for CCA.

Design and methods: Eight CCA-associated markers—carcinoembryonic antigen, carbohydrate antigen 19-9, alkaline phosphatase (ALP), and gamma glutamyl transferase, biliary-ALP, mucin5AC, CCA-associated carbohydrate antigen (CCA-CA) and CA-S27—were used as the inputs for the C4.5 decision tree classification model and the selected model was confirmed by ANN analyses. Eight serum markers for CCA were determined in the training set of 85 histologically proven-CCA patients and 82 control subjects. The chosen set of combined markers that gave the best diagnostic values for CCA was then validated in the testing set of 22 CCA patients and 60 controls.

Results: A decision tree diagram built by the C4.5 algorithm suggested the serial analysis of CCA-CA and ALP for distinguishing CCA patients from non-CCA subjects with all diagnostic parameters ≥95%. The combined tests showed a precise diagnosis in the testing set.

Conclusions: The C4.5 model indicates the combined markers of CCA-CA and ALP that produced the more precise diagnosis for CCA.

Introduction

Cholangiocarcinoma (CCA), a malignancy of bile duct epithelium, is a relatively rare cancer worldwide; however, its incidence is extremely high in Southeast Asia, especially in the northeastern part of Thailand [1]. CCA is normally difficult to diagnose until the disease develops to the advanced or disseminated stage, which makes CCA a tumor with an extremely poor prognosis. As a result, approximately 35/100,000 or 14,400 CCA patients die per year in Thailand [2]. Therefore, it is an urgent need to search for tumor markers that are sensitive and specific enough to diagnose individuals with CCA.

At present, serum alkaline phosphatase (ALP), carbohydrate antigen 19-9 (CA19-9), carcinoembryonic antigen (CEA) and gamma-glutamyl transferase (GGT) are the most routine markers for diagnosis of CCA in general clinical practices. A number of new CCA-associated markers have also been reported, such as biliary-alkaline phosphatase (BALP) [3], mucin5AC (MUC5AC) [4], CCA-associated carbohydrate antigen (CCA-CA) [5], CA-S27 [6], cytokeratin 19 fragment (CYFRA21-1) [7,8], receptor-binding cancer antigen expressed in SiSo cells IRCAS1 [9], and M2-pyruvate kinase [10]. However, none of these markers alone has impressive clinical diagnostic value [11].

Improvement in the diagnosis of cancer using combined measurements of multiple tumor markers has been demonstrated in many studies [11–13]. This approach provides a better sensitivity, specificity, and accuracy for detection of cancers. Computer-based diagnostic and referral systems, such as artificial neural network and decision tree

* Correspondence to: D. Chetchotsak, Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, 40002, Thailand. Fax: +66 43 362 145.
** Correspondence to: S. Wongkham, Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand. Fax: +66 43 348 386.
E-mail addresses: cdanai@kku.ac.th (D. Chetchotsak), sopit@kku.ac.th (S. Wongkham).
[1] Co-first author.

http://dx.doi.org/10.1016/j.clinbiochem.2015.03.022
0009-9120/

classifications, have been applied to simulate the combined-diagnostic models or equations from the measured tumor markers [12–14].

The computational data mining C4.5 model, and its most recent version C5.0, have been used extensively as a classification model in many different areas such as dental therapy [15], mammogram analysis [16], land-use planning [17], construction accidents [18], and individual credit evaluation [19]. In this study, we aimed to use the computational data mining C4.5 model to develop a combined diagnostic model that yielded the best diagnostic values for CCA from four routinely used serum markers for CCA—CEA, CA19-9, ALP, and GGT, and four recently reported markers—BALP, MUC5AC, CCA-CA and CA-S27. The C4.5 created several combined marker models and the model that had the highest discrimination power was then compared with those of the reference technique, single-hidden-layer feed-forward neural network, commonly known as artificial neural network (ANN). Finally, the discrimination power of the combined tests was validated in the second cohort subjects. The combination test may serve as a tool to provide more precise diagnosis of CCA.

Materials and methods

Patients and serum samples

All serum samples were obtained from the specimen bank of The Liver Fluke and Cholangiocarcinoma Research Center, Khon Kaen University, Khon Kaen, Thailand. Informed consent was obtained from all subjects, and the study protocol was approved by The Khon Kaen University Ethics Committee in Human Research. Serum from 85 histologically proven-CCA patients (Table 1) and from 82 age and sex matched control subjects, including 11 healthy persons (HE), 30 opisthorchiasis subjects (OV), 14 benign hepatobiliary disease patients (BHD), and 27 gastrointestinal cancer patients (GICA) were used in the first cohort as a training set. A second cohort used as a validation set was composed of 22 CCA patients and 60 controls (HE = 22, OV = 7, BHD = 16, GICA = 15). Cancer and BHD patients were diagnosed using histology. Opisthorchiasis in the OV group was confirmed by a positive fecal result for *Opisthorchis viverrini* eggs, and healthy controls were persons whose medical check-up, liver function tests and complete blood count were normal.

Determination of serum markers

ALP (U/L) and GGT (U/L) were determined using enzymatic assay (Roche Diagnostics GmbH, Mannheim, Germany). CEA (ng/mL) and CA19-9 (U/mL) were analyzed using enzyme linked immunosorbent assay (ELISA, Roche Diagnostics GmbH) according to the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC). Serum BALP (U/L), MUC5AC (Absorbance Unit, AU), CCA-CA (AU) and CA-S27 (AU) were determined using lectin capture ELISA as described previously [3–6].

Table 1
Demographical data of CCA patients of the training set.

Characteristics		Number of patients	
Age (years)	≤56	46	54.1%
	>56	39	45.9%
Sex	Male	56	65.8%
	Female	29	34.2%
Histological types	Papillary	17	20.0%
	Non-papillary	68	80.0%
Staging	I–III	12	14.1%
	IVA	25	29.4%
	IVB	48	56.5%
Jaundice[a]	Jaundice	49	57.6%
	Non-jaundice	36	42.4%

[a] Jaundice = total bilirubin >2 mg/dL.

Construction of C4.5 classification model

In this study, C4.5 decision tree algorithm was used as a classification model while an artificial neural network was used for benchmarking purposes. C4.5 is an algorithm used to construct a decision tree classification model in a logical form [20]. The model creates a decision tree based on the information gain concept [21], where information gain reflects the ability to classify the data using a particular marker. Starting at the top node of the tree, C4.5 selects the marker with the highest information gain to classify the data in a "top-down recursive divide-and-conquer manner" [22]. At the branch node, C4.5 selects another marker from the remaining set in the same manner. These procedures take place until no more markers are left. More details of C4.5 are provided in [20].

An artificial neural network known as (ANN) [23] is an information processing and computational system which is modeled after a biological nervous system. Like a human brain, ANN can process, learn, and remember information. The simplest and most common form of ANN is a single-hidden layer feedforward neural network trained using the backpropagation algorithm and was used in this study. ANN can map the relationship between the inputs and outputs through adjusting the connection weights and hence it has been used successfully in both regression and classification tasks.

To construct a classification model using both C4.5 and ANN, all combined markers were used as inputs and the diagnostic result as "CCA" or "nonCCA" was the output. For C4.5, we conducted an experiment to build a model with the best combination of markers. In this step, binary splits of at least five instances per leaf and a confidence threshold for pruning of 0.75 were used as settings. To construct ANN, the following settings were used: 20 hidden units, sigmoid function, and 50,000 learning cycles. The results obtained were compared to those from the C4.5 model. Both C4.5 and ANN were built using an open source machine learning software, WEKA [24] and were validated using the five-fold cross validation method.

Performance evaluation and experimental trials

For consistency, the CCA samples were referred to as "positive" while the non-CCA samples were labeled as "negative". The confusion matrix and defined performance measures for a binary class problem, used to measure performance of the classification models were: sensitivity (SEN), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV) and accuracy (ACC). These measures were defined as follows: the terms TP and TN denoted correct classification for positive and negative samples while FP and FN defined incorrect classification for positive and negative samples, respectively. The diagnostic values were determined as SEN = TP / (TP + FN); SPEC = TN / (TN + FP); PPV = TP / (TP + FP); NPV = TN / (TN + FN); and ACC = (TP + TN) / (TP + FN + FP + TN).

To make sure that the training data was chosen randomly to avoid bias, C4.5 model construction in this study was replicated ten times. This was done by training and validating the classification model according to the five-fold cross validation method. Performance of each model was then evaluated. These procedures were subsequently repeated ten times. A 95% confidence interval for all performances was then constructed and used for comparison and evaluation purposes.

Statistical analysis

Statistical analysis was performed using the SPSS 14.0 for Windows Evaluation software (SPSS Inc., Chicago, USA). Mann–Whitney U test was used to compare the serum level of marker of the CCA and the control groups. *P*-values <0.05 were considered statistically significant.

Please cite this article as: Pattanapairoj S, et al, Improve discrimination power of serum markers for diagnosis of cholangiocarcinoma using data mining-based approach, Clin Biochem (2015), http://dx.doi.org/10.1016/j.clinbiochem.2015.03.022