Evaluation of a Neural Network Classifier for Pancreatic Masses Based on CT Findings

Mitsuru Ikeda, MD[1], Shigeki Ito, MD[2],
Takeo Ishigaki, MD[3], and Kazunobu Yamauchi, MD[1]

[1]Department of Medical Information and Medical Records,
Nagoya University Hospital.
[2]Department of Radiology, Nagoya Memorial Hospital.
[3]Department of Radiology, Nagoya University, School of Medicine.

(Address)
Department of Medical Information and Medical Records
Nagoya University Hospital
65, Tsurumai-cho, Showa-ku, Nagoya 466, JAPAN.
TEL: +81-52-744-2666; FAX: +81-52-744-1356
Email: a40495a@nucc.cc.nagoya-u.ac.jp

Abstract

We have investigated a neural network classifier based on CT findings extracted by a radiologist for the differential diagnosis between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis, and compared its classification performance with that of Bayesian analysis, Hayashi's quantification method II, and radiologists. The three computerized classification methods were designed to classify categorized CT findings extracted by a radiologist, and were trained and tested on 71 cases. There was comparable performance of the neural network, the Bayesian analysis, Hayashi's quantification method II, and the radiologists, in classifying pancreatic carcinoma and inflammatory mass.

Key words: Computer aided diagnosis .
Neural network.
Receiver operating characteristic curve (ROC).
Radiology and radiologists.
Pancreas.

1. Introduction

Using computed tomography (CT), radiologists often find differential diagnoses of pancreatic masses difficult. Especially, the distinction between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis bases on CT findings is still not certain, although a large number of studies have been made on the CT findings of these two disease entities (1-5). Thus, a computerized second opinion would be especially helpful for clinicians in the differential diagnosis between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis based on CT findings.

On the other hand, artificial neural networks are solving problems that previous technologies have been unable to resolve satisfactorily (6), and are beginning to find applications in many fields, including the field of radiological diagnosis (6-24). Therefore, we have applied a neural network classification based on CT findings extracted by a radiologist to separating the mass-forming pancreatitis from the ductal adenocarcinoma in the pancreatic mass differential diagnosis.

In addition, we compared the classification performance of the neural network with those of the other two classification methods; these are the probabilistic method using Bayes' formula (Bayesian analysis) (25-28), the discriminant method using Hayashi's quantification method II (29). We also compared the classification performance of these three computerized classification methods with that of radiologists.

According to Boone et al (8, 9), there are two steps in the radiological diagnosis: the first is a pattern recognition task in which the radiologist views the image and compiles a list of any abnormalities present (that is, "radiographic findings") (9); and the second step is to evaluate the radiographic findings, using cognitive processes developed through medical training and experience, to arrive at a differential diagnosis (8). The aim of this paper is to evaluate the performance of the three computerized classification methods in this second step. Moreover, in this paper, we are not concerned with the between-observer and within-observer variations, owing to the

subjective nature of the input values that the computerized classification methods require from radiologists (16).

2. Materials and Methods

2.1 Database

The subjects used for this study were retrospectively obtained from CT files with mass lesions in the pancreas; these were composed of 32 patients (28 males and 4 females; age 40.8 years mean; range 35-75 years) whose mass lesions had been diagnosed as a pancreatic inflammatory mass, and 76 patients (50 males and 26 females, age 60.3 years mean; range 38-81 years) whose mass lesions had been diagnosed as pancreatic ductal adenocarcinoma. Here, the diagnoses of 93 of these patients (67 with ductal adenocarcinoma and 26 with inflammatory mass) had been established at pathological examination. CT was performed on Toshiba 900S, Toshiba TCT80A (Toshiba, Tokyo), or Hitachi CTW600 (Hitachi, Tokyo) scanners. All patients underwent a CT examination consisting of plain CT, dynamic CT, and high-dose enhancement CT, and the pancreas was scanned with 5-mm contiguous sections in the enhancement CT.

We compiled a comprehensive list of CT findings pertinent to the interpretation of the pancreatic mass, based on practical clinical experience and review of the literature (9). From this, a checklist-style form listing CT findings was drawn up (9). Initially, this list consisted of 32 features with two or three categorized findings for each. These features could be grouped into the following three categories: features related to the pancreatic mass, features related to the pancreatic parenchyma, and features of secondary abnormalities about arteries, veins, bile duct, lymph nodes, etc. The features related to the pancreatic mass and the pancreatic parenchyma are shown in Table 1 and Table 2.

An experienced radiologist (with more than 10 years experience) interpreted all CT images of 108 cases, extracted the above-mentioned 32 features from each case, and

entered them on the data form; from this data form, the database used for this study was constructed. These 32 features are large, compared to the number of training cases for the computerized classification methods. Therefore, we restricted to the features related to the pancreatic mass and the pancreatic parenchyma. Further, every feature was evaluated with χ^2 test for its ability to discriminate between the pancreatic carcinoma and the inflammatory mass. Eighteen of these 32 features had significantly different categorized findings between the pancreatic carcinoma and the inflammatory mass (at a significance level of 0.05); so, we selected the database consisting of these 11 features related to the pancreatic mass and the pancreatic parenchyma as the training and testing data for the three computerized classification methods (Table 1, Table 2).

For a training and testing set for the computerized classification methods, 39 cases with pancreatic carcinoma were randomly selected from the entire database, and all 32 cases with inflammatory mass in the entire database were used; for comparison of the three computerized classifiers, this training and testing data set is the same for all the three computerized classification methods.

2.2 Neural Networks

We used the commercially available software (NeuralWorks Professional II; NeuralWare, Pittsburgh) on a workstation (Sun 4, Fujitsu, Tokyo), to design, train, and test a variety of artificial neural networks. Three-layer, feed-forward networks with a back-propagation algorithm were employed in this study (16). However, there is no generally accepted rule for determining the topology and the hierarchy of the network (9, 19). Hence, we tested numerous different architectures with a different number of hidden nodes and combinations of input parameters for optimizing the neural network architectures. Only the results of the successful neural network architecture are reported here; this neural network consisted of 15 input units, 15 hidden units, and one output unit, which represented the diagnosis (0 = pancreatic carcinoma, 1 = inflammatory mass).

2.3 Probabilistic Method Using Bayes' Formula (Bayesian Analysis)

The Bayes theorem is a rather simple method of combining CT findings to estimate the probability that a mass lesion in the pancreas is adenocarcinoma (27). In applying Bayes' formula, it is necessary to develop a probability matrix (26). We obtained this probability matrix for the 11 features given in the section 2.1 from a training set and applied Bayes' formula to patterns within a testing set in order to determine the probability for pancreatic carcinoma; that is, the probabilities in this probability matrix were estimated by the relative frequencies calculated from a training data set. Here, a rate of 0.5 was used for the respective incidence of the two possible diseases (that is, pancreatic carcinoma and inflammatory mass). The software for these calculations was written by us.

2.4 Hayashi's Quantification Method II

The Hayashi's quantification method II is a linear discriminant analysis technique for the system described by the qualitative variables (29), and implemented in a statistical package such as SPSS (Statistical Package for the Social Sciences) (30, 31). By using the SPSS package (30, 31), we obtained the discriminant function for a training set and applied it to patterns within a testing set; the 11 features given in the section 2.1 were used as independent variables for this discriminant function

2.5 Evaluation of Performance

To predict the ability of the three computerized classification methods to generalize from the training cases and make diagnoses concerning cases that had not been included in the training, we employed the leave-one-out method (16). In this method, all but one of the above-mentioned 71 cases was used to train the neural network, to determine the probability matrix for the Bayesian analysis, and to obtain the discriminant function for the Hayashi's quantification method II. The single case that

was left out was then used to test the neural network, the Bayesian system, and the discriminant function. This procedure was repeated so that each of the 71 cases was used once as a testing case (16).

To evaluate the three computerized classification methods, we also conducted another experiment, in which training was performed on the above-mentioned 71 cases, and testing was performed on the group of the remaining 37 cases with pancreatic carcinoma and the same 32 cases with inflammatory mass as in the training set. That is, for cases with pancreatic carcinoma, this experiment is a cross-validation study [training the classifier with one group of subjects, and testing it in another separate group (12)], and for the cases with inflammatory mass, this is a re-substitution experiment [training the classifier with one group of subjects, and testing it in the same group used to train it (12)]. Thus, we call this experiment a "semi-cross-validation" method in what follows. Further, for reference, we performed re-substitution experiments, in which testing was performed on the group used to train the classifier (12).

Four radiologists agreed to participate as readers for this study. As a group they had an average of 10 years (median 7 years) experience as diagnostic radiologists. They were not informed of the purpose of this study, and were blind to individual clinical information. They were told that the subjects were the CT images with the mass lesion in the pancreas which was either pancreatic ductal adenocarcinoma or pancreatic inflammatory mass.

To evaluate the performance of radiologists in distinguishing between pancreatic carcinoma and inflammatory mass, these four radiologists read the CT images and reported inflammatory/carcinoma mass judgments on the following discrete five-point scale: 1 = definitely inflammatory mass, 2 = probably inflammatory mass, 3 = possibly carcinoma, 4 = probably carcinoma, and 5 = definitely carcinoma. For these five points, 0, 0.25, 0.5, 0.75, and 1 were assigned, respectively, to the probability judgments of the radiologists regarding the presence of the pancreatic carcinoma. They

interpreted the CT images of 30 cases with pancreatic carcinoma and 29 cases with inflammatory mass; these cases were extracted from the entire database.

Receiver operating characteristic (ROC) analysis (32-34) was employed to evaluate the performance of the three computerized classification methods and radiologists in distinguishing between pancreatic carcinoma and inflammatory mass. The area under the ROC curve (A_z) (32-34) was used as an index of performance. Gurney pointed out that "no one has yet validated the predictive accuracy of neural networks with true tests of predictive accuracy such as the Brier index (35)." So, we also calculated the Brier scores for assessing the performance of the three computerized classification methods and radiologists (24, 35-37). Here, using the Sanders decomposition, the Brier score was partitioned into two components, resolution (often referred to as discrimination) and calibration (24, 38); we also calculated this discrimination as $\frac{1}{N}\sum n_i o_i(1-o_i)$, where o_i = the observed rate of pancreatic carcinoma in category i , n_i = the number of patients in category i, N = the total sample size, and the categories correspond to the five based on predicted 0 to 20%, 20 to 40%, 40 to 60%, 60 to 80%, and 80 to 100% probabilities of pancreatic carcinoma (24, 38).

For calculating the correlation coefficients among the judgments of the computerized classification methods and radiologists and for calculating the Brier scores, the outputs of the neural network and the discriminant function for Hayashi's quantification method II were converted to a scale ranging from 0 to 1, with 0 being the maximum output value, 1 the minimum for the neural network, 0 as the minimum discriminant function value, and 1 for the maximum of the Hayashi's quantification method II; we considered these normalized outputs as the probability with which the pancreatic mass on CT would be pancreatic carcinoma. As to the judgments of the four radiologists, the above-mentioned probabilities assigned to the discrete five-point scale were used for these calculations.

3. Results

3.1 Performance of Computerized Classification Methods

Fig. 1 compares the performance of the three computerized classification methods as tested with the leave-one-out method, in classifying pancreatic carcinoma and inflammatory mass. The area under the ROC curve and the Brier score for the three computerized classification methods are summarized in Table 3. The performance of the neural network was lower than that of the other two computerized classification methods, but the differences among the A_Z values obtained by the neural network, the Bayesian analysis, and the Hayashi's quantification method II were not statistically significant.

In the leave-one-out experiment, there was a mild correlation between the outputs of the neural network and the probability estimates by the Bayesian system (its Pearson correlation coefficient (r) is 0.729). There were also mild correlations between the outputs of the neural network and the discriminant coefficients by the Hayashi's quantification method II (r = 0.803) and between the probability estimates by the Bayesian system and the discriminant coefficients by the Hayashi's quantification method II (r = 0.788).

Fig. 2 compares the performance of the three computerized classification methods as tested with the "semi-cross-validation" experiment, in classifying pancreatic carcinoma and inflammatory mass. (The area under the ROC curve and the Brier score are summarized in Table 3.) In this experiment, the performance of the neural network was better than that of the other two computerized classification methods, but the differences among the A_Z values obtained by the neural network, the Bayesian analysis, and the Hayashi's quantification method II were not statistically significant. The difference in the Brier score between the neural network and the Hayashi's quantification method II was statistically significant ($p < 0.01$).

3.2 Comparison between Computerized Classification Methods and Radiologists

Fig. 3 compares the performance of the radiologists and the neural network as tested with the leave-one-out method, in classifying pancreatic carcinoma and inflammatory mass. The area under the ROC curve and the Brier score for the four radiologists are also shown in Table 3. The performance of the neural network was almost equivalent to that of the radiologists, and there was no statistically significant difference between the A_Z value obtained by the neural network and the ones by the radiologists.

There were mild to weak correlations between the judgments of the three computerized classification methods and those of the four radiologists (Table 4). Two-way analysis of variance with no repeated measures showed that the three computerized classification methods had a statistically significant difference with the correlation coefficients ($p < 0.005$); the outputs of the probabilistic method using Bayes' formula were closer correlated with the radiologists' judgments than those of the other two computerized classification methods, and the correlation between the outputs of the neural network and the radiologists' judgments was the weakest among them; however, these estimated means of the correlation coefficients failed to achieve statistical significance.

4. Discussion

As indicated by the leave-one-out method, the neural network's performance was worse than that of the other two computerized classification methods, as to both A_Z and Brier score. However, in the "semi-cross-validation" experiment, the neural network's performance proved to be better than that of the other two; especially, according to the Brier score, the neural network's performance exceeded that of the Hayashi's quantification method II with a rather high level of statistical significance. The question now arises: "which among the three computerized classification methods showed the best performance?"

The difference in performance between the leave-one-out method and the "semi-cross-validation" experiment originates from the shrinkage; here, the shrinkage refers to the difference between the observed classification rate and the true rate (35). Because the testing set in the "semi-cross-validation" method is more similar to the training set than in the leave-one-out method, the shrinkage in the "semi-cross-validation" method can be considered to be larger than that in the leave-one-out method. In fact, the neural network performance in the "semi-cross-validation" method was better than in the leave-one-out method. However, the other two computerized classification methods' performance in the "semi-cross-validation" method was inferior to that in the leave-one-out method. Moreover, the Spiegelhalter's Z scores are not compatible with the above considerations. Therefore, these results indicate that the estimation of the true classification rate in the computerized classification methods is difficult, as pointed out by Gurney (35), although the leave-one-out method makes the most efficient use of the data set that is available and results in a complete cross-validation method (21). Further, these results suggest that the neural network has more ability to classify the training set than the other two classification methods; this is also suggested from the result that the neural network performance was the best among the three computerized classification methods in the re-substitution experiment. In other words, the present results indicate that the neural network is apt to be over-trained, as previously reported (20).

From the above considerations, it seems reasonable to conclude that the performances of the three computerized classification methods, in classifying pancreatic carcinoma and inflammatory mass, were comparable. Gurney et al pointed out that neural networks offered no advantage over the less sophisticated Bayesian system in the prediction of probabilities of malignancy in solitary pulmonary nodules (24). Also, in this study, the neural network classifier offered no significant advantage over the simple Bayesian method and the linear discriminant analysis method. Here, the probabilistic method using simply the Bayes' formula requires the most simple computation among the three computerized classification methods using multiple

variables. Therefore, the Bayesian analysis should be given more credit, although the estimation of disease incidences is somewhat problematic.

In the leave-one-out method, the neural network that used image features related to the pancreatic mass and the pancreatic parenchyma (extracted by an experienced radiologist) performed at a slightly higher than the average level achieved by four radiologists, as to both A_Z and Brier score. Here, the neural network in this study used only the limited features related to the pancreatic mass and the pancreatic parenchyma, and did not use the features of secondary abnormalities that are very useful for the differential diagnosis between the pancreatic carcinoma and mass-forming pancreatitis (such as liver metastasis, lymph node swelling, etc.). From these considerations, our results indicate that, when a radiologist with less experience finds differential diagnosis of a pancreatic mass difficult, a three-layer neural network classifier by entering CT findings given in a checklist-style form can be used as a computer-aided diagnostic tool for the differential diagnosis between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis.

Moreover, in the leave-one-out method, the Bayesian analysis and the Hayashi's quantification method II performed at a higher level than the radiologists; in particular, the difference in A_Z between the Bayesian analysis and radiologist C, the Hayashi's quantification method II and radiologist C, and the Hayashi's quantification method II and the combined data by pooling four radiologists was statistically significant ($p < 0.05$). Therefore, the Bayesian analysis and the Hayashi's quantification method II can be also applied to distinguish between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis, by entering CT findings given in a checklist-style form.

As mentioned before, the three computerized classification methods were comparable with each other, in their performance in classifying pancreatic carcinoma and inflammatory mass. However, there were only mild correlations among the probabilistic judgments of these three methods. Further, there were rather weak correlations between the judgments of the three computerized classification methods

and those of the four radiologists. These results drive us to the question whether, as a computerized second opinion, the judgments given by the computerized classifier might be closely correlated with those of the experienced radiologist. This seems very important and interesting, and deserves further consideration.

It has been pointed out that the neural network has little or no capability to explain the underlying rules (39). The other two computerized classification methods discussed in this paper also have little such explanation capability with respect to the underlying rules, although the classification mechanism in the Bayesian analysis is relatively comprehensible. Therefore, the analysis of the radiographic features by the classification methods using multiple variables offers little in the way of broadening our knowledge of radiological diagnosis. However, computerized classification methods have the potential for solving differential diagnostic problems which are not readily resolvable by the experienced radiologist. So, these methods may be the limits of performance in differential diagnosis in terms of the radiological modality. From this viewpoint, we can say that the limit of the probability that a randomly selected pair of cases with pancreatic carcinoma and mass-forming pancreatitis would be ranked so that the pancreatic mass on the CT of the patient with pancreatic carcinoma is assigned a higher probability than that of the patient with pancreatitis is about 0.92 (based on Bayesian analysis performance in the leave-one-out method).

To evaluate the probabilistic judgments, we calculated the two indexes of the area under the ROC curve and the Brier score. Here, it is open to question whether the output of the neural network and the discriminant function for Hayashi's quantification method II can be regarded as a "probability estimate" or not; however, this question is beyond the scope of this paper. Although there were some inconsistencies in the relations between these two indexes, there was a rather close correlation between them (r = - 0.888). So, there seems to be a functional relationship between the area under the ROC curve and the Brier score; this deserves further consideration.

5. Conclusions

In the differential diagnosis between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis, the performance of the neural network classifier by entering CT findings given in a checklist-style form was comparable to that of the Bayesian analysis and the Hayashi's quantification method II. Further, these three computerized classification methods offered almost the same performance as radiologists in classifying pancreatic carcinoma and inflammatory mass based on CT findings.

Acknowledgment: This study was supported in part by Grant-in-Aid for Scientific Research (B) 05454601 from the Ministry of Education, Science and Culture in Japan.

References

1. Neff C.C.; Simeone J.F.; Wittenberg J.; Mueller P.R.; Ferrucci J.T. Inflammatory pancreatic masses: problems in differentiating focal pancreatitis from carcinoma. Radiology 150(1):35-38; 1984.
2. Freeny P.C.; Marks W.M.; Ryan J.A.; Traverso L.W. Pancreatic ductal carcinoma: diagnosis and staging with dynamic CT. Radiology 166(1):125-133; 1988.
3. Luetmer P.H.; Stephens D.H.; Ward E.M. Chronic pancreatitis: reassessment with current CT. Radiology 171(2):353-357; 1989.
4. DelMaschio A.; Vanzulli A.; Sironi S.; Castrucci M.; Mellone R.; Staudacher C.; Carlucci M.; Zerbi A.; Parolini D.; Faravelli A.; Cantaboni A.; Garancini P.; Carlo V.D. Pancreatic cancer versus chronic pancreatitis: diagnosis with CA 19-9 assessment, US, CT, and CT-guided fine-needle biopsy. Radiology 178(1):95-99; 1991.
5. Schulte S.J.; Baron R.L.; Freeny P.C.; Patten R.M.; Gorell H.A.; Maclin M.L. Root of the superior mesenteric artery in pancreatitis and pancreatic carcinoma: evaluation with CT. Radiology 180(3):659-662; 1991.
6. Maclin P.S.; Dempsey J. Using an artificial neural network to diagnose hepatic masses. J. Med. Syst. 16(5):215-225; 1992.
7. Asada N.; Doi K.; MacMahon H.; Montner S.M.; Giger M.L.; Abe C.; Wu Y. Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: pilot study. Radiology 177(3):857-860; 1990.
8. Boone J.M.; Gross G.W.; Greco-Hunt V. Neural networks in radiologic diagnosis: I. Introduction and illustration. Invest. Radiol. 25:1012-1016; 1990.
9. Gross G.W.; Boone J.M.; Greco-Hunt V.; Greenberg B. Neural networks in radiologic diagnosis: II. Interpretation of neonatal chest radiographs. Invest. Radiol. 25:1017-1023; 1990.

10. Fujita H.; Katafuchi T.; Uehara T.; Nishimura T. Application of artificial neural network to computer-aided diagnosis of coronary artery disease in myocardial SPECT bull's-eye images. J. Nucl. Med. 33(2):272-276; 1992.
11. Miller A.S.; Blott B.H.; Hames T.K. Review of neural network applications in medical imaging and signal processing. Med. Biol. Eng. Comput. 30:449-464; 1992.
12. Kippenhan J.S.; Barker W.W.; Pascal S.; Nagel J.; Duara R. Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects. J. Nucl. Med. 33(8):1459-1467; 1992.
13. Scott R. Artificial intelligence: its use in medical diagnosis. J. Nucl. Med. 34(3):510-514; 1993.
14. Tourassi G.D.; Floyd C.E.; Sostman H.D.; Coleman R.E. Acute pulmonary embolism: artificial neural network approach for diagnosis. Radiology 189(3):555-558; 1993.
15. Scott J.A.; Palmer E.L. Neural network analysis of ventilation-perfusion lung scans. Radiology 186(3):661-664; 1993.
16. Wu Y.; Giger M.L.; Doi K.; Vyborny C.J.; Schmidt R.A.; Metz C.E. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 187(1):81-87; 1993.
17. Siebler M.; Rose G.; Sitzer M.; Bender A.; Steinmetz H. Real-time identification of cerebral microemboli with US feature detection by a neural network. Radiology 192(3):739-742; 1994.
18. Kippenhan J.S.; Barker W.W.; Nagel J.; Grady C.; Duara R. Neural-network classification of normal and Alzheimer's disease subjects using high-resolution and low-resolution PET cameras. J. Nucl. Med. 35(1):7-15; 1994.
19. Porenta G.; Dorffner G.; Kundrat S.; Petta P.; Duit-Schedlmayer J.; Sochor H. Automated interpretation of planar thallium-201-dipyridamole stress-redistribution

scintigrams using artificial neural networks. J. Nucl. Med. 35(12):2041-2047; 1994.

20. Chan K.H.; Johnson K.A.; Becker J.A.; Satlin A.; Mendelson J.; Garada B.; Holman B.L. A neural network classifier for cerebral perfusion imaging. J. Nucl. Med. 35(5):771-774; 1994.
21. Gross G.W.; Boone J.M.; Bishop D.M. Pediatric skeletal age: determination with neural networks. Radiology 195(3):689-695; 1995.
22. Tourassi G.D.; Floyd C.E.; Sostman H.D.; Coleman R.E. Artificial neural network for diagnosis of acute pulmonary embolism: effect of case and observer selection. Radiology 194(3):889-893; 1995.
23. Baker J.A.; Kornguth P.J.; Lo J.Y.; Williford M.E.; Floyd C.E. Jr. Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 196(3):817-822; 1995.
24. Gurney J.W.; Swensen S.J. Solitary pulmonary nodules: Determining the likelihood of malignancy with neural network analysis. Radiology 196(3):823-829; 1995.
25. Ledley R.S.; Lusted L.B. Reasoning foundations of medical diagnosis. Science 130(3366):9-21; 1959.
26. Lodwick G.S.; Haun C.L.; Smith W.E.; Keller R.F.; Robertson E.D. Computer diagnosis of primary bone tumors: a preliminary report. Radiology 80(2):273-275; 1963.
27. Gurney J.W. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis: Part I. Theory. Radiology 186(2):405-413; 1993.
28. Gurney J.W.; Lyddon D.M.; McKay J.A. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis: Part II. Application. Radiology 186(2):415-422; 1993.

29. Hayashi C. On the predication of phenomenon from qualitative data and quantification of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics 3(2):69-98; 1952.
30. Miyake I.; Yamamoto K. SPSS toukei pakkeiji I. kisohen. Tokyo: Touyoukeizaishinpousya; 1976.
31. Miyake I.; Nakano K.; Mizuno K.; Yamamoto K. SPSS toukei pakkeiji II. kaisekihen. Tokyo: Touyoukeizaishinpousya; 1976.
32. Swets J.A.; Pickett R.M. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press; 1982.
33. Hanley J.A.; McNeil B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29-36; 1982.
34. Hanley J.A.; Mcneil B.J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148(3):839-843; 1983.
35. Gurney J.W. Neural networks at the crossroads: caution ahead. Radiology 193(1):27-30; 1994.
36. Spiegelhalter D.J. Probabilistic prediction in patient management and clinical trials. Stat. Med. 5:421-433; 1986.
37. Redelmeier D.A.; Bloch D.A.; Hickam D.H. Assessing predictive accuracy: how to compare Brier scores. J Clin Epidemiol 44(11):1141-1146; 1991.
38. McClish D.K.; Powell S.H. How well can physicians estimate mortality in a medical intensive care unit? Med. Decis. Making. 9(2):125-132; 1989.
39. Bergeron B.P.; Shiffman R.S.; Rouse R.L. Data qualification: Logic analysis applied toward neural network training. Comput. Biol. Med. 24(2):157-164; 1994.

Figure Legends

1. Fig. 1.

ROC curves comparing performance of the neural network, the Bayesian analysis, and the Hayashi quantification II method, as tested with the leave-one-out method.

2. Fig. 2.

ROC curves comparing the performance of the neural network, the Bayesian analysis, and the Hayashi quantification II method, as tested with the "semi-cross-validation" experiment. Here, the "semi-cross-validation" experiment means the experiment in which testing was performed on the group consisting of malignant cases not used in training (37 cases with pancreatic carcinoma) and the same benign cases used in training (32 cases with inflammatory mass).

3. Fig. 3.

ROC curves comparing performance of the four radiologists and the neural network as tested with the leave-one-out method.

4. Table 1.

Features related to pancreatic mass

5. Table 2.

Features related to pancreatic parenchyma

6. Table 3.

Area under ROC curve and Brier score for various classification methods

7. Table 4.

Correlation coefficients between computerized classification methods and radiologists

Table 1

Feature	Categorized Findings for Each Feature	χ^2 test[a]
location of mass	pancreas head, pancreas body, pancreas tail	NS
size of mass	≤ 2 cm, 2~4 cm, ≥ 4 cm	NS
density of mass on plain CT	low density, iso density, high density	NS
density of mass on dynamic CT	low density, iso density, high density	$p < .001$
density of mass on high-dose CE CT	low density, iso density, high density	$p < .001$
margin of mass on high-dose CE CT	well-defined and smooth, well-defined and irregular, undefined	$p < .05$
ring enhancement of mass on high-dose CE CT	absence, presence	$p < .01$
density homogeneity of mass on high-dose CE CT	homogenous, slightly inhomogenous, inhomogenous	$p < .01$
central low density in mass on high-dose CE CT	absence, presence	$p < .001$
main pancreatic duct dilatation in mass	absence, presence	$p < .001$
calcification in mass	absence, presence	$p < .001$
central cyst formation in mass	absence, presence (size: < 1 cm), presence (size: ≥ 1 cm)	$p < .001$
peripheral cyst formation in mass	absence, presence (size: < 1 cm), presence (size: ≥ 1 cm)	NS

CE, contrast enhancement; CT, computed tomography; NS, not significant at the .05 level.

[a] χ^2 test for comparisons of categorized findings between the pancreatic carcinoma and inflammatory mass.

Table 2

Feature	Categorized Findings for Each Feature	χ^2 test[a]
main pancreatic duct dilatation	absence, presence (size: < 50%), presence (size: ≥ 50%)	NS
location of dilated main pancreatic duct	absence, distal to mass, proximal to mass or diffuse	$p < .001$
atrophy of pancreatic parenchyma distal to mass	absence, presence	$p < .01$
calcification in pancreas	absence, presence	NS
cyst formation in pancreas	absence, presence (size: < 1 cm), presence (size: ≥ 1 cm)	NS

NS, not significant at the .05 level.

[a] χ^2 test for comparisons of category values between the pancreatic carcinoma and inflammatory mass.

Table 3

	Area under ROC Curve (A_Z)	Standard Deviation of A_Z	Brier Score	Spiegelhalter's Z Score	Resolution in Brier Score
Neural Network (leave-one-out)	0.866	0.043	0.170	2.664	0.160
Neural Network (“semi-cross-validation”)	0.916	0.048	0.0809	- 0.0935	0.0695
Neural Network (re-substitution)	0.998	0.002	0.0186	- 2.031	0.0141
Bayesian Analysis (leave-one-out)	0.922	0.036	0.120	5.847	0.103
Bayesian Analysis (“semi-cross-validation”)	0.902	0.041	0.153	7.100	0.130
Bayesian Analysis (re-substitution)	0.949	0.027	0.0902	3.459	0.0867
Hayashi’s Quantification Method II (leave-one-out)	0.919	0.032	0.139	- 2.639	0.136
Hayashi’s Quantification Method II (“semi-cross-validation”)	0.893	0.042	0.167	- 0.857	0.131
Hayashi’s Quantification Method II (re-substitution)	0.978	0.015	0.0977	- 3.371	0.0547
Radiologist A	0.910	0.039	0.131	0.408	0.126
Radiologist B	0.799	0.064	0.190	1.145	0.161
Radiologist C	0.753	0.068	0.215	2.440	0.187
Radiologist D	0.886	0.058	0.155	10.589	0.104
Pooling Four Radiologists	0.835	0.028			

Table 4

Computerized Classification Methods	Radiologists' Comparison	Correlation Coefficients	95 % Confidence Interval for Estimated Correlation Coefficient (r) between Computerized Classification Method and Radiologists
Neural Network	A	0.507	
	B	0.424	
	C	0.419	
	D	0.553	
			$0.332 < r < 0.620$
Bayesian Analysis	A	0.677	
	B	0.574	
	C	0.513	
	D	0.814	
			$0.500 < r < 0.789$
Hayashi's Quantification Method II	A	0.623	
	B	0.608	
	C	0.475	
	D	0.727	
			$0.464 < r < 0.752$

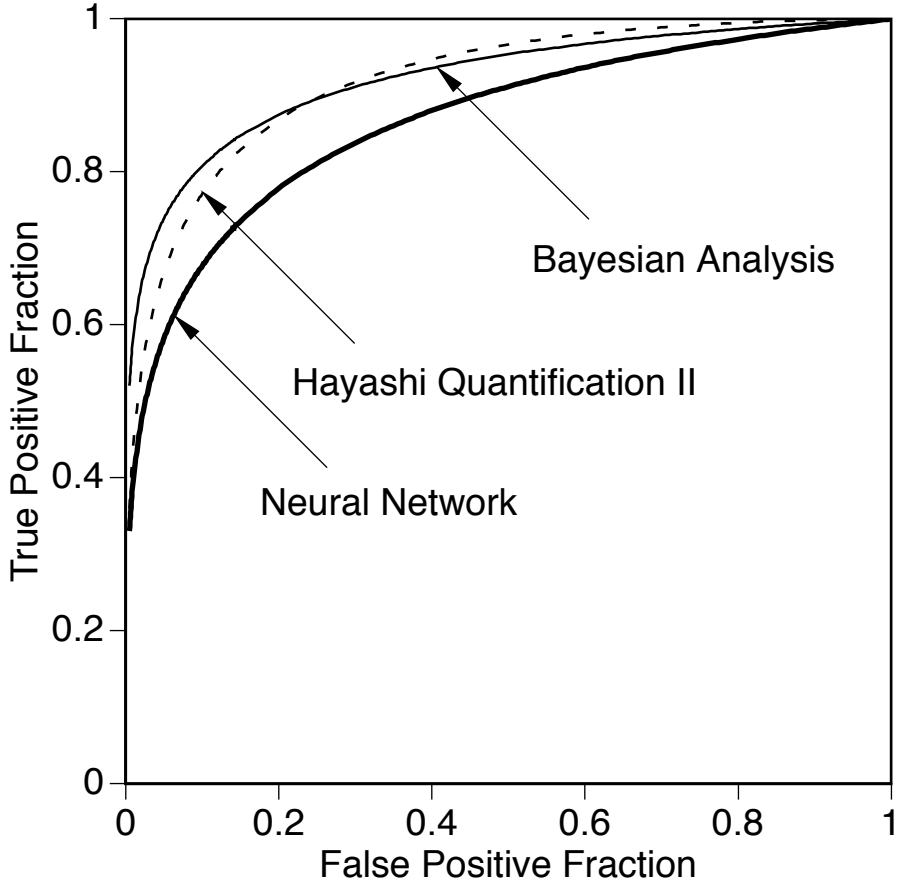
1
0.8
0.6
0.4
0.2
0
True Positive Fraction
0
0.2
0.4
0.6
0.8
1
False Positive Fraction
Bayesian Analysis
Hayashi Quantification II
Neural Network

Fig. 1

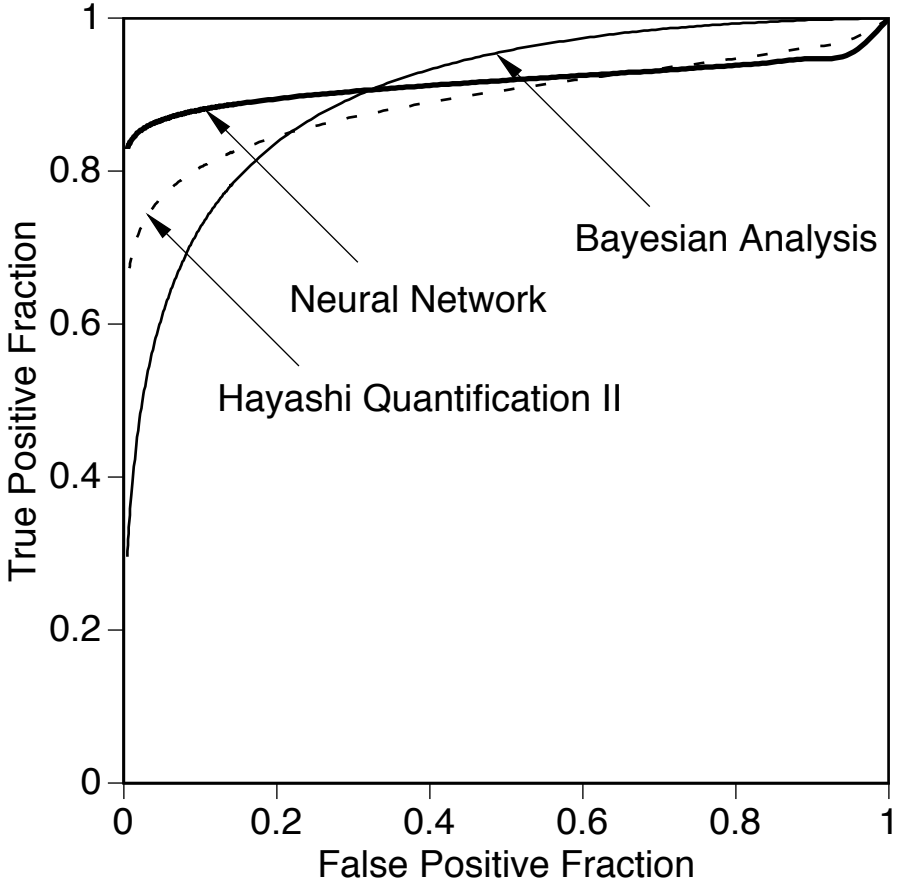
1
0.8
0.6
0.4
0.2
0
True Positive Fraction
0
0.2
0.4
0.6
0.8
1
False Positive Fraction
Bayesian Analysis
Neural Network
Hayashi Quantification II

Fig. 2

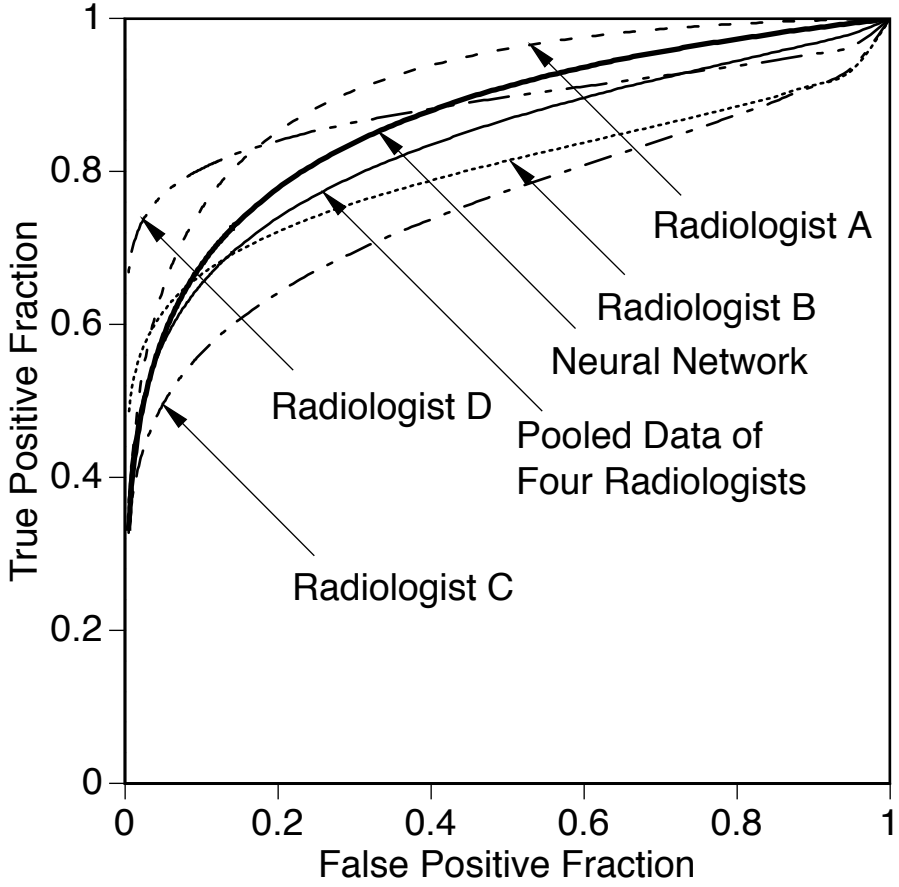
1
0.8
0.6
0.4
0.2
0
True Positive Fraction
0
0.2
0.4
0.6
0.8
1
False Positive Fraction
Radiologist A
Radiologist B
Neural Network
Pooled Data of
Four Radiologists
Radiologist D
Radiologist C

Fig. 3

Summary

We have investigated a neural network classifier based on CT findings extracted by a radiologist for the differential diagnosis between the pancreatic ductal adenocarcinoma and mass-forming pancreatitis, and compared its classification performance with that of Bayesian analysis, Hayashi's quantification method II, and radiologists.

The subjects were retrospectively obtained from CT files with mass lesions in the pancreas; these were 32 patients whose mass lesions had been diagnosed as a pancreatic inflammatory mass, and 76 patients whose mass lesions had been diagnosed as pancreatic ductal adenocarcinoma. An experienced radiologist interpreted these CT images, extracted initially the 32 features from each case, and entered them on the database used for this study. From these data, 18 of these 32 features had significantly different categorized findings between the pancreatic carcinoma and the inflammatory mass ($p < 0.05$), by the χ^2 test. So, we selected the database consisting of these 11 features related to the pancreatic mass and the pancreatic parenchyma as the training and testing data for the three computerized classification methods. For a training and testing set for the computerized classification methods, 39 cases with pancreatic carcinoma were randomly selected from the entire database, and all 32 cases with inflammatory mass in the entire database were used. To predict the ability of the three computerized classification methods, we employed the leave-one-out method, and also conducted the "semi-cross-validation" experiment, in which training was performed on the above-mentioned 71 cases, and testing was performed on the group of the remaining 37 cases with pancreatic carcinoma and the same 32 cases with inflammatory mass as in the training set. The area under the receiver operating characteristic (ROC) curve (A_Z) was used as an index of performance. Further, we also calculated the Brier scores.

In the leave-one-out experiment, A_Z was 0.866 (Brier score, 0.170) for the neural network, 0.922 (Brier score, 0.120) for the Bayesian analysis, 0.919 (Brier score, 0.139)

for the Hayashi's quantification method II, and 0.835 when pooling four radiologists. So, the performance of the neural network was lower than that of the other two computerized classification methods, but the differences among the A_Z values obtained by the neural network, the Bayesian analysis, and the Hayashi's quantification method II were not statistically significant. On the other hand, in the "semi-cross-validation" experiment, A_Z was 0.916 (Brier score, 0.0809) for the neural network, 0.902 (Brier score, 0.153) for the Bayesian analysis, 0.893 (Brier score, 0.167) for the Hayashi's quantification method II, and the differences among these A_Z values were not statistically significant.

Therefore, there was comparable performance of the neural network, the Bayesian analysis, Hayashi's quantification method II, and the radiologists, in classifying pancreatic carcinoma and inflammatory mass.

Mitsuru Ikeda, Shigeki Ito, Takeo Ishigaki, and Kazunobu Yamauchi