Upper Gastrointestinal Tract Bleeding:
Assessing the Diagnostic Contributions of the History and Clinical Findings

CHRISTIAN OHMANN, PHD, KLAUS THON, MD, HARTMUT STÖLTZING, MD, QIN YANG, WILFRIED LORENZ, MD

Various strategies can be used in the diagnosis of upper gastrointestinal tract bleeding. This study investigates the relevance of anamnestic and clinical findings for the diagnosis of the bleeding source. The authors introduced a computer-aided diagnostic system using Bayes' theorem and compared it with clinicians' predictions using anamnestic and clinical findings only. There was no difference in the overall accuracy rates, but a difference was observed in the diagnostic behaviors of the two "systems." In addition, the discriminatory ability of the computer-aided system, the sharpness of the predictions obtained, and the reliability of the posterior probabilities were analyzed. It is concluded that the clinician and the computer-aided system are not able to discriminate well between the disease categories. Derived classification matrices and probability-based measures show the reasons for the inadequacy of diagnostic information obtainable from the clinical history and physical findings. *Key words*: computer-aided diagnosis; Bayes' theorem; probabilistic diagnosis; discriminatory ability; reliability; clinical accuracy; upper gastrointestinal tract bleeding. **(Med Decis Making 6:208–215, 1986)**

Patients admitted to the hospital with acute upper gastrointestinal tract hemorrhage present many problems, one of which is the need for early diagnosis of the source of hemorrhage. Different diagnostic strategies can be used.[21] The diagnosis may be based on the history and clinical findings, on upper gastrointestinal radiography, or on endoscopic findings. Several prospective trials showed that endoscopy is more accurate than radiography.[15] However, there is a higher potential risk in using endoscopy compared with radiography.[7] Clinical history and examination are thought to be inferior to both in diagnostic accuracy, but carry no risk.[4]

These results raise the question whether the history and clinical findings are necessary in the diagnostic decision making process. The answer depends primarily on the amount of diagnostic information provided by these data. If, as has been suggested, little diagnostic information is obtained, this process of careful questioning has little clinical relevance. However, if useful diagnostic information could thus be obtained, the patient could be spared the risk and discomfort of endoscopy or radiography.

Studies performed to measure the diagnostic relevance of the history and clinical findings have been rare, and cover only some aspects of the problem.[4, 19, 24] We investigated the diagnostic predictions of experienced clinicians and of a successful computer-aided model.[3] The analysis of these predictions, which were compared with proven final diagnoses, was not restricted to the common but inadequate concept of discriminatory ability, e.g., measured by accuracy or predictive value. Additional criteria such as sharpness of the diagnostic predictions and reliability of the probabilities were considered.[9, 13, 14]

Patients and Methods

PATIENTS

We investigated 457 consecutive patients admitted on an emergency basis for acute upper gastrointestinal tract bleeding to the Marburg Surgical Clinic between January 1978 and February 1983. The criterion for acute upper gastrointestinal tract bleeding was either hematemesis or melena as defined in the O.M.G.E. International Upper Gastro-Intestinal Bleeding Survey.[18] As soon as each patient was admitted to the hospital, a detailed history was taken and a careful physical examination was performed. All data were documented on a computer questionnaire especially designed for the purpose. The protocol contained 35 history variables and nine clinical investigations which

Received June 25, 1985, from the Department of Theoretical Surgery and Surgery Clinic, Centre for Operative Medicine I, University of Marburg, Marburg, West Germany. Accepted for publication after revision October 15, 1985. Supported by grant from Deutsche Forschungsgemeinschaft (Oh 39/2-1). Presented in part at the Royal College of Physicians Computer Workshop, Paris, France, 1983, and at the annual meeting of the German Society for Medical Documentation and Statistics (GMDS), Heidelberg, Germany, 1983.

Address correspondence and reprint requests to Dr. Ohmann: Department of Theoretical Surgery, Centre for Operative Medicine I, University of Marburg, Baldingerstraße, D-3550 Marburg, West Germany.

were expected to discriminate well between the possible diseases (table 1[19]). In order to estimate the conditional probabilities adequately, four disease categories were formed: gastric ulcer, duodenal ulcer, esophageal varices, and a group containing all other possible bleeding sources.

FINAL DIAGNOSIS

Endoscopy was performed on each patient, almost always within four hours of admission. About 50% of the patients had a second or third endoscopic examination during the first ten days after admission, and 15% of the patients were operated on. The final diagnosis of the bleeding source was based on the findings at the emergency endoscopy and on histologic and x-ray findings, findings at operation, and all further endoscopic findings.

When the data did not yield a clear diagnosis, two clinicians from the endoscopy unit were called to agree upon the final diagnosis. In 82% of the patients a unique bleeding source could be identified, but there were problems in diagnosing the bleeding sources in patients who had multiple lesions (18%). Patients having one lesion with signs of bleeding and another lesion without signs of bleeding were assigned to the former diagnostic category.[6] In the remaining cases the two clinicians were asked to define the major bleeding source and the patients were assigned to the appropriate diagnostic categories.

COMPUTER-AIDED DIAGNOSIS

The computer-aided diagnosis was performed with the "Independence Bayes" model, which assumes the conditional independence of the symptoms within every disease category and uses Bayes' theorem to calculate the posterior probabilities.[10, 17] An *a priori* probability of $P(D) = 0.25$ for every disease category D was chosen, which agrees approximately with our admission rates. The conditional probabilities P(S/D) were estimated by dividing the number of patients with disease D and symptom S by the number of patients with disease D. For each patient the disease D with the highest posterior probability was taken as the computer prediction.

To achieve an unbiased estimate of the actual error rates of the computer-aided diagnostic system, the patients were divided into two groups, a training set and a test set.[27] The training set included all patients admitted to the hospital between January 1978 and December 1981 ($n = 362$) and was used to estimate the conditional probabilities P(S/D). The performance of the computer-aided system was tested in a separate validation sample (test set) of all patients admitted to the hospital between January 1982 and February 1983 ($n = 95$). All calculations were done on a Hewlett-Packard desk-top computer (HP 9815A).

CLINICIANS' PREDICTIONS

In addition to the computer-aided prediction, a diagnostic prediction from the clinician, using the history and physical findings only, was noted prospectively on the computer questionnaire for every patient in the test set. The same clinician took the history, performed the physical examination, and filled in the questionnaire for any given patient. In a six-month pilot period from July 1981 to December 1984 the four participating clinicians from the endoscopy unit were able to familiarize themselves with this type of prediction. For five patients in the test set no diagnostic prediction was made by the clinician, hence 90 diagnostic predictions by the clinicians could be analyzed.[20]

Table 1 • Features of the History and Physical Examination Used in the Diagnosis of Upper Gastrointestinal Tract Bleeding

History
Source of referral
Gender
Place of origin
Marital status
Occupation
Blood group
Age
Height
Weight (kg)
Hematemesis (yes, no)
Hematemesis, since when
Retching
Melena, (yes, no)
Melena, since when
Ulcer complications
Ulcer evidence
Ulcer operation
Nausa
Vomiting
Regurgitation
Heartburn
Dysphagia
Pain
Duration of symptoms
Appetite
Weight (change)
Bowel habit
Bowel movements (per day)
Drugs, (yes, no)
Drugs, date
Smoking
Alcohol use
Past history, which
Past history, date
Past history, data quality
Clinical examination
Mental state
General appearance
Skin
Abdomen
Rectal examination
Pulse rhythm
Pulse rate
Systolic blood pressure
Diastolic blood pressure

Table 2 • The Forced Classification Matrix for the Diagnostic Predictions of the Clinician in the Test Set (n = 95)*

Clinician's Prediction	Final Diagnosis: Gastric Ulcer	Duodenal Ulcer	Varices	Other	Total
Gastric ulcer	10	2	0	10	22
Duodenal ulcer	6	14	2	7	29
Varices	1	1	18	1	21
Other	1	1	3	13	18
TOTAL	18	18	23	31	90

*Five of the clinicians' predictions were missing.

Table 3 • The Forced Classification Matrix for the Diagnostic Predictions of the Computer in the Test Set (n = 95)

Computer Prediction	Final Diagnosis: Gastric Ulcer	Duodenal Ulcer	Varices	Other	Total
Gastric ulcer	10	4	2	2	18
Duodenal ulcer	7	8	0	7	22
Varices	1	1	19	2	23
Other	3	6	3	20	32
TOTAL	21	19	24	31	95

Results

CLINICIANS' PREDICTIONS VERSUS FINAL DIAGNOSES

Table 2 shows the forced classification matrix for the diagnostic prediction of the clinician.[9] The predictions were accurate in 55 of 90 patients (61%). Accuracies in the different disease categories were 14 of 18 (78%) in the duodenal ulcer group, 78% in the varices group, 56% in the gastric ulcer group, and 42% in the diagnostic category "other." Of 21 predictions of varices as the bleeding source 18 were correct, which gives a predictive value of 86%.[9] The predictive value for the diagnostic category "other" was 72%; for duodenal ulcer, 48%; and for gastric ulcer, 45%.

COMPUTER PREDICTION: CLASSIFICATION MATRIX

The forced classification matrix in table 3 shows accurate predictions for 57 of 95 patients (60%). The computer prediction was accurate in 19 of 24 cases (79%) in the varices group, 65% in the disease category "other," 48% in the gastric ulcer group, and 42% in the duodenal ulcer group. Predictive values ranged from 19 of 23 cases (83%) in the varices group, to 63% for "other," 56% for gastric ulcer, and 36% for duodenal ulcer.

CLINICIAN VERSUS COMPUTER

Although there was very little difference between the overall accuracies of the clinicians' predictions (61%) and the computer's predictions (60%), there were marked differences with regard to two disease categories (tables 2 and 3). For duodenal ulcer the clinician was 36% more accurate than the computer. In the diagnostic category "other" the opposite was true, with a difference of 33% in the accuracy rates. The predictive values showed only moderate differences of up to 12% between the clinicians and the computer.

Since our two systems were tested on the same cases, paired-comparison techniques are appropriate to test for differences in performance.[12] Table 4 shows that in addition to 40 patients correctly diagnosed by both systems, 15 cases were correctly diagnosed by the clinician and not by the computer and 16 the other way around. This gives a nonsignificant result in the McNemar test, which means that the null hypothesis of equal nonerror rates cannot be rejected. On the other hand there is a difference in the diagnostic behaviors of the two systems, which can be documented by the high frequency of 31 of 90 cases (34%) in the heteronomous cells of table 4. The null hypothesis of a non-agreement coefficient equals zero between the clinician and the computer is tested by an inversion $\overline{\Phi}$ of Pearson's phi-coefficient Φ (table 4).[16]

Table 4 • Paired Comparison of the Clinicians' Predictions and the Computer Predictions in the Test Set (n = 95)

Computer Prediction	Clinicians' Prediction: Correct	Incorrect	Missing	Total
Correct	40	16	1	57
Incorrect	15	19	4	38
TOTAL	55	35	5	95

$$\overline{\Phi} = 1 - \Phi = 1 - \frac{40 \cdot 19 - 16 \cdot 15}{\sqrt{56 \cdot 34 \cdot 55 \cdot 35}}$$

Using the chi-square distribution with 1 degree of freedom, a significant result ($p < 0.001$) is obtained. Thus, the alternative hypothesis of non-agreement between the systems has to be accepted.

COMPUTER: DERIVED CLASSIFICATION MATRICES

All previous measurements of performance were based on the forced classification matrix, in which all patients are allocated to a disease.[9, 20] However, when studying discriminatory ability, it is also interesting to look at the assigned probabilities. This can be done only for the computer-aided system.

For further consideration of the data, diseases with low probabilities could be omitted. This is illustrated in table 5, where those diseases D, with a posterior probability ($P(D/S) < 0.10$ were excluded. The exclusion matrix shows that the diagnosis "varices" can be

well distinguished from the other diagnostic categories.

In 18 of 21 cases of gastric ulcer (86%), 89% of cases of duodenal ulcer, and 84% of cases in the disease category "other," the diagnosis "varices" could be excluded. For the 24 patients who had varices, the bleeding source "gastric ulcer" could be excluded 16 times (67%), duodenal ulcer could be excluded 18 times (75%), and "other" could be excluded 15 times (63%). The discrimination of the computer-aided system between patients who had ulcers and all patients with "other" sources of hemorrhage was moderate. The discriminatory ability to separate duodenal ulcer patients from gastric ulcer patients was bad. This can be seen in the low exclusion rates of 7 of 21 (33%) duodenal ulcers in gastric ulcer patients and of 7 of 19 (37%) gastric ulcers in duodenal ulcer patients.

In table 6 the patients for whom a confident diagnosis was made are separated from patients for whom the diagnosis was not conclusive.[9] In 60 of 95 (63%) computer-aided predictions the largest posterior probability (P(D/S) did not exceed 0.8. Defining sharpness of a diagnostic system as the ability to assign high probability values to one disease, our system could not be described as sharp in the presence of so many doubtful cases.[14] On examination of the sharp diagnoses only, it is interesting that the diagnostic accuracy was 24 of 35 (69%), which is hardly different from the overall accuracy of 60%.

Table 5 • Exclusion Matrix of the Computer-aided System in the Test Set (n = 95)*

Excluded Disease	Final Diagnosis: Gastric Ulcer	Duodenal Ulcer	Varices	Other
Gastric ulcer	3	7	16	14
Duodenal ulcer	7	4	18	15
Varices	18	17	4	26
Other	11	8	15	5
Maximum possible value	21	19	24	31

*Diseases D with p(D/S) < 0.1 are excluded.

Table 6 • Classification Matrix with Doubt of the Computer-aided System in the Test Set (n = 95)*

Computer Prediction	Final Diagnosis: Gastric Ulcer	Duodenal Ulcer	Varices	Other
Gastric ulcer	1	1	—	1
Duodenal ulcer	—	2	—	2
Varices	3	1	13	1
Other	1	1	—	8
Doubt	16	14	11	19
TOTAL	21	19	24	31

*For patients with the largest probability p(D/S) not exceeding 0.80 the computer-aided prediction was classified as doubt.

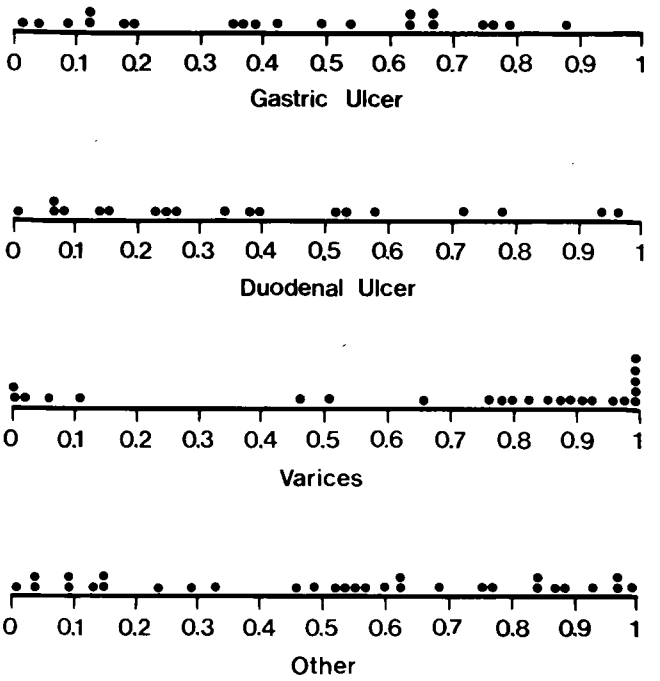

FIGURE 1. Dot diagrams of the probabilities assigned to the actual disease categories in the test set (n = 95). Each dot represents a patient.

COMPUTER: PROBABILITY-BASED MEASURES

In addition to the classification matrices used to measure the performance of a diagnostic system, several other measures which are continuous functions of the assigned probabilities should be used.[13, 14, 20] The dot diagram in figure 1 provides a first impression of the distributions of the probabilities assigned to the actual diseases. The overall average probability for the actual diseases was 0.52 in the test set (table 7), with marked differences between the four diagnostic categories (fig. 1). The varices group especially had a different distribution, with a small peak near 0 and a high peak near 1, compared with the approximately uniform distributions in the other three diagnostic categories.

Two other criteria reflect other aspects of the degrees of discrimination between the diagnostic categories (table 7). These criteria are based on scores that describe the discrepancy between the actual disease D and the posterior probabilities assigned to the four disease categories. One of the most popular scoring methods in nonmedical applications is the *quadratic score* or Brier score:

$$\frac{1}{N} \cdot \sum_i \left[(1 - P_{i\,d(i)})^2 + \sum_{j \neq d(i)} P_{ij}^2 \right]$$

where N is the number of patients, P_{ij} the posterior probability for D_j in patient i, and d(i) the index of the

Table 7 • Discriminatory Ability and Reliability of the Computer-aided System

	Training Set (n = 362)		Test Set (n = 95)	
Criterion*	Observed	Expected†	Observed	Expected
Non-error rate	0.67	0.78	0.60	0.73
Average probability for the actual disease	0.61	0.69	0.52	0.63
Quadratic	0.46	0.31	0.59	0.37
ϵ-modified logarithmic‡	−0.81	−0.49	−1.00	−0.58

*Criteria are defined in the text.
†Calculated under the null hypothesis of perfect reliability of the probabilities.
‡ϵ = 0.01.

actual disease of patient i.[13] If the assigned probability to the actual disease is 1.00, then patient i clearly contributes nothing to the quadratic score. On the other hand, if some other disease is assigned a probability of 1, the term of the i^{th} patient becomes 2. Hence the lower limit is 0 and the upper limit is 2. In our case the quadratic score was 0.59 in the test set (table 7). Utilizing the quadratic score, there is little difference between using our system and using an uninformative indifferent system, where each disease is assigned a proability of 0.25 throughout, which leads to a quadratic score of 0.75.[14]

The *ϵ-modified logarithmic score*:

$$\frac{1}{N} \cdot \sum_{i} \left[\ln w\,(P_{id(i)}) + \epsilon \cdot \sum_{j \neq d(i)} \ln\,(w(P_{ij})/\epsilon) \right]$$

where N is the number of patients, P_{ij} the posterior probability for D_j in patient i, d(i) the index of the actual disease of patient i, $\epsilon > 0$ and $w(P_{ij}) = (1 - \epsilon) \cdot P_{ij} + \epsilon$, penalizes especially low probabilities for the actual disease.[14] The ϵ-modified logarithmic score is approximately equal to:

$$\frac{1}{N} \cdot \sum_{i} \ln\,(P_{id(i)} + \epsilon)$$

where N is the number of patients, $P_{id(i)}$ the posterior probability for the actual disease and $\epsilon > 0$. Using an $\epsilon = 0.01$ produces a theoretical minimum of −4.56 and a derived maximum of 0. Our computer-aided diagnostic system produces an ϵ-modified logarithmic score of −1.00, which is again not very different from the score of −1.26 of the indifferent system, where each disease is assigned a probability of 0.25 (table 7). A comparison between the two samples in table 7 shows that the criteria calculated in the training set are superior to the same criteria calculated in the test set.

COMPUTER: RELIABILITY* OF THE PROBABILITIES

One important aspect of a good performance in probabilistic diagnosis is the reliability of the posterior probabilities, which is quite distinct from the question of discrimination.[11, 13, 14] The posterior probability P that a patient has disease D giving a symptom vector S is called reliable when in a sample of adequate size of patients all having the same symptom vector S, about P% do actually have the disease D. Usually it is not possible to collect enough cases with identical symptoms and verify that within sampling fluctuations, the assigned diagnostic probabilities can be trusted. One method of overcoming these difficulties is to consider the test set as a whole and hypothesize that whenever an event is assigned a probability P it will occur with frequency P. Using perfect reliability as the null hypothesis, departures from this perfect state of affairs can be measured and tested.[13, 14]

In table 7 the expected values of the diagnostic scores are calculated under the null hypothesis of perfect reliability. If we use the difference between the observed and the expected values as a reliability measure, we can see that the observed non-error rate is 13% lower than the expected rate, which has to be calculated as the average maximum probability.[13] The observed average probability for the actual disease is only 52% and therefore 11% smaller than expected. Regarding these two reliability measures as normally distributed, the null hypothesis of perfect reliability must be rejected ($p < 0.01$, $p < 0.001$).[13, 14, 20] In addition, the expected values of the quadratic score and the ϵ-modified logarithmic score do suggest better results than could be observed in the study. The training set shows the same trend for all reliability measures as the test set.

There are many ways in which a system may deviate from reliable performance. In order to measure whether a system favors a particular disease (size bias), a comparison of the observed and expected frequencies for every disease is necessary. The expected frequency in a disease category D is calculated as the average sum of the posterior probabilities for the disease D.[13] Table 8 shows that there is an overassignment in the duodenal ulcer group, with 23.7 expected instead of 19 observed cases. In the varices group and in the "other disease" class there were small underassignments, with 21.2 and 28.5 expected cases compared with 24 and 31 observed cases, respectively. This gives a nonsignificant test result using approximate standard normal test statistics.[13] Another possibility for the measurement of the reliability of the posterior probabilities is to divide the probabilities into intervals and compare the expected and observed frequencies in each subgroup, using a chi-square goodness-of-fit test for every disease.[1] In table 8 this is done, using four equidistant probability intervals. The common trend in all

*"Reliability" as used in the European literature cited here corresponds broadly to "calibration" in recent North American literature.—*Ed.*

four disease categories is a higher expected than observed value in the interval 0.76 to 1.00 and a smaller expected than observed value in the interval 0.00 to 0.25. Only the results in the varices group and those in the "other disease" category are significant ($p < 0.05$).

Discussion

The clinicians and the computer-aided system were not able to discriminate adequately between the four given disease categories, as could be seen in the accuracy rates of 61% and 60%. The results of our computer-aided diagnostic system are comparable to the results in a multicenter trial with an accuracy of 59% and to our earlier results with accuracy rates of 65% to 69%.[4, 24] We could not achieve the excellent results of computer-aided diagnostic systems used for other diagnostic problems such as the acute abdomen.[3, 25] These results suggest that there is little relevant diagnostic information in the history and physical findings; nevertheless, some points must be further discussed before any definite conclusions can be reached.

Regarding the poor performance of the clinicians, it is important to note that no inexperienced doctor took part in this study. All doctors were experienced members of the endoscopic unit and had had a minimum of two years of regular training in the diagnosis of upper gastrointestinal tract bleeding. It may be argued that neither experienced doctors nor successful computer-aided models can produce good results if the correct questions are not posed and the wrong physical examinations are performed. The variables collected in our study contained all clinical attributes which were thought to be important in diagnostic terms. The computer questionnaire was based on the protocol of the O.M.G.E. International Upper Gastro-Intestional Bleeding Survey, expanded and clarified to a detailed protocol by our senior clinician.[4, 18] Therefore it is unlikely that any important diagnostic variables have been omitted.

The quality of the data is thought to be high, for two reasons. Before starting our trial in 1978, we discussed terminology in detail; all terms used in describing upper gastrointestional tract bleeding were carefully defined.[4, 18] In addition, there was a prospective trial of collection of the data using a computer questionnaire, performed by experienced clinicians. Nevertheless, for 19% of the patients more than 20% of the data was missing. The main part of this data loss probably relates to the poor condition of some patients at the time of admission, so that neither detailed histories nor careful physical examinations could be obtained. A comparison of the computer-aided system's performances for patients with and without missing data reduces the accuracy rate by about 9% for diagnostic predictions based on missing data.

Table 8 • Comparison of the Observed and Expected Frequencies (Goodness of Fit) for Every Disease in Four Intervals of Probabilities in the Test Set ($n = 95$)

Probability for the Actual Disease	Final Diagnosis: Gastric Ulcer Obs*	Gastric Ulcer Exp†	Duodenal Ulcer Obs	Duodenal Ulcer Exp	Varices Obs	Varices Exp	Other Obs	Other Exp
0.76–1.00	4	5.7	3	7.9	16	16.4	10	12.9
0.51–0.75	5	6.6	5	7.2	2	2.4	8	7.0
0.26–0.50	5	4.2	4	5.1	1	0.9	4	5.0
0.00–0.25	7	5.0	7	3.5	5	1.5	9	3.6
TOTAL	21	21.5	19	23.7	24	21.2	31	28.5

*Obs = observed frequency.

†Exp = expected frequency = sum of probabilities for the actual disease. The calculation was done seperately for every combination of the disease categories and the intervals of probabilities.

In about 20% of our emergency cases the patients have multiple lesions in the upper gastrointestinal tract. Most of these patients have only one bleeding source and one or two accompanying lesions. A bias is introduced if these patients are assigned to one of the four disease categories. The accuracy of the computer prediction is about 10% higher for patients with a single lesion compared with patients with multiple lesions, which underlines the problems of using one-disease models.[27] The contributions of missing data and multiple diseases to the error rate are moderate and only partly explain the poor results.

Computer-aided diagnostic systems using Bayes' theorem are very popular.[17, 25] Nevertheless, the question arises whether the appropriate model was used in our study. The simplifying assumption of independence of symptoms is a matter of great controversy.[22] Comparisons of different diagnostic techniques showed, however, that the independence model is a good discriminator even when the assumptions are strictly unjustified.[2, 23] This does not imply that the independence model, using all the data from the history and physical examination, is the best choice of all possible statistical models. However, the results in the literature suggest that differences in diagnostic accuracies due to the choice of the model are often small compared with the influences of other factors such as the type, the quality and the completeness of the data collected.[2, 22, 23] If medical decision making methods are to stand any chance of success, they must be simple to use and comprehensible to the clinician, conditions that are well satisfied by the "independent Bayes" model. For better understanding of the underlying structure of the diagnostic problem from the statistical viewpoint, it would be interesting to use only a few important diagnostic variables instead of looking at all signs, symptoms, and diagnostic tests. This point is currently under investigation by the application of a stepwise linear logistic model[22] and the independence model together with different variable-selection procedures.[8]

In this study we were not restricted to the simple determination of diagnostic accuracy but tried to analyze the reasons for the disappointing results. The diagnostic predictions of the clinicians were different from the computer predictions (table 4). This means that computer-aided diagnostic systems, which have been used since 1978 in our Surgical Clinic, have probably had no substantial influence on the clinicians' views of the diagnostic process. Since the clinicians were not forced to assign probabilities to the different disease categories, a definite answer to this question cannot be given. The impression that clinicians are now coming to regard clinical diagnosis as a process of statistical or probabilistic nature seems to be rather overly optimistic.[5, 11, 26] One main problem that prevents a change from the traditional view of the diagnostic process as an intuitive art, based upon personal experience and textbook knowledge, to a probabilistic and statistical diagnosis is that calculated posterior probabilities of computer-aided models cannot be trusted. In our study, the independence model produces figures that are not real probabilities and thus cannot help the clinicians to estimate probabilities. At the worst, it may engender a false sense of certainty and mislead the clinician in his decision making process.[9, 11, 13, 14] Assuming perfect reliability of the probabilities of the independence model, departures from this perfect state of affairs have been measured in our study. Significant differences between observed and expected values for the non-error rate, the average probability for the actual disease, the quadratic criterion, and the ϵ-modified logarithmic criterion indicate that the discriminatory performance is less than would be expected from the predictions themselves (table 7).[12–14] The probabilistic predictions are overconfident, which may be related to the fact that in the independence model related information is considered as unconnected evidence.[13, 22] The overconfident predictions are symmetrically distributed throughout the diagnostic categories, which means that no particular disease is fovoured by the computer-aided system (table 8).

The nonreliability of the probabilities produced by the computer-aided system leads to difficulties in interpreting derived classification matrices and probability-based measures of performance (tables 5–7).[13, 14] Even when the probabilities of the independence model could be trusted, the different performance measures (expected values) give disappointing results concerning the discriminatory ability (table 7). The main reason for this is that the computer-aided model is not able to assign high probability values to one disease, as could be seen in the average maximum probability of 73% (= expected non-error rate) and in the other performance measures (table 7).[13, 14] Computer-aided systems that have good discriminatory ability must necessarily produce sharp predictions, i.e., predictions that assign nearly 100% to one disease.[14] The many non-sharp predictions in our study indicate that little diagnostic information is provided by the clinical history and physical examination. Only the bleeding source esophageal varices could be well discriminated from other sources. The separation of duodenal ulcer patients from gastric ulcer patients was bad using this model (table 5).

One reason for the disappointing results is that in upper gastrointestinal tract bleeding clinical signs and symptoms that normally could point to a particular diagnosis may be dominated by the effects of the blood loss, especially in dramatic cases with severe hemorrhage. On the other hand, the history and physical findings occasionally suggest a diagnosis that is not the bleeding source. Jaundice and ascites, for example, indicate esophageal varices, but this may be misleading since bleeding in a patient who has liver disease with esophageal varices may be the result of peptic ulceration or gastric erosions.[4] The various interactions between elements of the history and the clinical findings, the effects of the bleeding, and the underlying lesion limit the ability of both the clinician and the computer-aided system to correctly identify the source of hemorrhage. It appears that the initial clinical features are more helpful in determining prognosis than diagnosis. Several studies have shown that the short-term prognosis, i.e., whether the bleeding would continue or subside, could be predicted with sufficient accuracy using clinical signs and symptoms on admission and computer-aided prognostic systems.[4, 18]

In summary, it is concluded that at present there seems to be no combination of symptoms and signs that reliably points to a particular diagnosis, even when sophisticated computer-aided systems are used. If an accurate diagnosis of the source of bleeding is required at an early stage, high-technology investigations such as endoscopy must be employed.[18]

The authors thank Dr. Madeleine Ennis and Marlene Verfürth for assistance in the preparation of this report.

References

1. Cox DR: The analysis of binary data. London, Methuen, 1970, pp 90–95
2. Croft JD: Mathematical models in medical diagnosis. Ann Biomed Engineering 2:69–89, 1974
3. De Dombal FT, Leapper DJ, Staniland JR, et al: Computer-aided diagnosis of acute abdominal pain. Br Med J 2:9–13, 1972
4. De Dombal FT, Morgan AG, Staniland JR, et al: Clinical features—computer analysis, *in*: Dykes PW, Keighley MRB (eds): Gastrointestinal Hemorrhage. Bristol, John Wright, 1981, pp 155–165
5. Diamond GA: Computer diagnosis: revolution or revelation. Int J Cardiol 2:219–220, 1982
6. Forrest JAH, Finlayson NDC, Shearman DJC: Endoscopy in gastrointestional bleeding. Lancet II:394–397, 1974
7. Gilbert DA, Silverstein FE, Tedesco FJ, et al: National ASGE survey on upper gastrointestinal bleeding. Complications on endoscopy. Dig Dis Sci 26:55–59, 1981
8. Habbema JDF, Hermanns J: Selection of variables in discriminant

analysis by F-statistic and error rate. Technometrics 19:487–493, 1977
9. Habbema JDF, Hilden J, Bjerregaard B: The measurement of performance in probabilistic diagnosis: I. The problem, descriptive tools, and measures based on classification matrices. Meth Inform Med 17:217–226, 1978
10. Hall GH: The clinical application of Bayes' theorem. Lancet II:555–557, 1967
11. Healy MJR: Computer-aided diagnosis—an overview of some theoretical problems, *in*: De Dombal FT, Gremy F (eds): Decision Making and Medical Care. Amsterdam, North-Holland Publishing Company, 1976, pp 357–359
12. Hilden J, Habbema JDF: A comparative evaluation of two systems in probabilistic jaundice diagnosis, *in*: Alperovitch A, De Dombal FT, Gremy F (eds): Evaluation of Efficacy of Medial Action. Amsterdam, North-Holland Publishing Company, 1979, pp 123–132
13. Hilden J, Habbema JDF, Bjerregaard B: The measurement of performance in probabilistic diagnosis: II. Trustworthiness of the exact values of the diagnostic probabilities. Meth Inform Med 17:227–237, 1978
14. Hilden J, Habbema JDF, Bjerregaard B: The measurement of performance in probabilistic diagnosis: III. Methods based on continuous functions of the diagnostic probabilities. Meth Inform Med 17:238–246, 1978
15. Kinard HB, Powell DW, Sandler RS, et al: A current approach to acute upper gastrointestinal bleeding. J Clin Gastroenterol 3:231–240, 1981
16. Lienert GA: Verteilungsfreie Methoden in der Biostatistik. Meisenheim am Glan: Verlag Anton Hain, 1973, pp 528–533
17. Lusted LB: Introduction to Medical Decision Making. Springfield, Ill., Charles C Thomas, 1968, pp 1–69
18. Morgan AG, Clamp SE: O.M.G.E. International Upper Gastro-Intestinal Bleeding Survey 1978–1982. Scand J Gastroenterol 19 (suppl 95):41–58, 1984
19. Ohmann C, Thon K, Stöltzing H, et al: Klinische und computerunterstützte Diagnose bei oberer Gastrointestinalblutung. Deutsch Med Wochenschr 108:1484–1486, 1983
20. Ohmann C, Thon K, Stöltzing H, et al: Computerunterstützte Diagnose bei der oberen Gastrointestinalblutung, *in*: Köhler CO, Tautu P, Wagner G (eds): Der Beitrag der Informationsverarbeitung zum Fortschritt in der Medizin (Proceedings der 28. Jahrestagung der GMDS). Berlin, Heidelberg, New York, Tokyo, Springer Verlag, 1984, pp 251–258
21. Rohde H, Troidl H, Lorenz W, et al: Neue Ansätze zur Frage: Hat die Notfallendoskopie für den Chirurgen Bedeutung? Med Klin 73:773–780, 1978
22. Spiegelhalter DJ: Statistical aids in clinical decision-making. Statistician 31:19–36, 1982
23. Titterington DM, Murray GD, Murray LS, et al: Comparison of discrimination techniques applied to a complex data set of head injured patients. JR Statist Soc A 144:145–175, 1981
24. Thon K, Ohmann C, Rohde H, et al: Einführung der computerunterstützten Diagnose bei der oberen Gastrointestinalblutung. Langenbeck's Arch Chir suppl: 231–235, 1982
25. Wardle A, Wardle L: Computer-aided diagnosis—a review of research. Meth Inform Med 17:15–28, 1978
26. Wagner HN: Bayes' theorem: an idea whose time has come? Am J Cardiol 49:875–877, 1982
27. Zentgraf R, Victor N: Some problems arising in the statistical treatment of diagnosis. Meth Inform Med 17:10–15, 1978