Supporting diagnostic decisions using hybrid and complementary data mining applications: a pilot study in the pediatric emergency department

Lorenz Grigull[1] and Werner M. Lechner[2]

INTRODUCTION: This article demonstrates the capacity of a combination of different data mining (DM) methods to support diagnosis in pediatric emergency patients. By using a novel combination of these DM procedures, a computer-based diagnosis was created.

METHODS: A support vector machine (SVM), artificial neural networks (ANNs), fuzzy logics, and a voting algorithm were simultaneously used to allocate a patient to one of 18 diagnoses (e.g., pneumonia, appendicitis). Anonymized data sets of patients who presented in the emergency department (ED) of a pediatric care clinic were chosen. For each patient, 26 identical clinical and laboratory parameters were used (e.g., blood count, C-reactive protein) to finally develop the program.

RESULTS: The combination of four DM operations arrived at a correct diagnosis in 98% of the cases, retrospectively. A subgroup analysis showed that the highest diagnostic accuracy was for appendicitis (97% correct diagnoses) and idiopathic thrombocytopenic purpura or erythroblastopenia (100% correct diagnoses). During the prospective testing, 81% of the patients were correctly diagnosed by the system.

DISCUSSION: The combination of these DM methods was suitable for proposing a diagnosis using both laboratory and clinical parameters. We conclude that an optimized combination of different but complementary DM methods might serve to assist medical decisions in the ED.

Arriving at the correct diagnosis is a mandatory yet sometimes challenging task for clinicians. In particular, in the emergency department (ED), up to 15% of patients are misdiagnosed (1). Medical malpractice litigation accounts for a high incidence of malpractice payments, and pediatricians are often the targets of such litigations (2). Appendicitis and meningitis are frequent diagnoses encountered in such malpractice suits (2). In daily practice, medical doctors assemble clinical signs, symptoms, and laboratory results to suggest a diagnosis. The idea of computer-driven systems to assist or even replace "the human factor" has been evaluated for a long time. Very recently, an IBM-built computer system named Watson defeated two human rivals—not in medical diagnosis—but in the popular *Jeopardy!* show (http://www.wired.com/epicenter/2011/01/ibm-watson-jeopardy/). In the field of diagnostics, so-called clinical decision support systems (e.g., http://www.easydiagnosis.com and www.isabelhealthcare.com) boomed in the 1990s (3,4). Mathematically speaking, the systems are usually built using "yes" or "no" algorithms, a strategy that frequently fails in medicine as answers are rarely a clear-cut "yes" or "no." To overcome this problem, even more sophisticated mathematical procedures have been developed and evaluated in tailored medical scenarios (5–8). In this study, a combination of four different data mining (DM) methods using 26 parameters (14 clinical parameters, e.g., age, body temperature, and blood pressure; and 12 laboratory parameters, e.g., hemoglobin and leukocyte counts, C-reactive protein level) as input variables computed a distinct diagnosis for each patient. The growing interest in DM applications such as the ones used in this study is partly because of the fact that artificial neural networks (ANNs) are, in principle, able to learn from their data and, in addition, show distinctive skills in analyzing nonlinear data sets such as those frequently encountered in medicine (7,8).

A support vector machine (SVM), the second DM application used to program the diagnostic tool used in this study, is a supervised learning method that yields an appropriate discrimination program using known data inputs and outputs (training data set). It first generates an N-dimensional hyperplane separating the training data into 18 different half spaces and then classifying unknown *de novo* data by determining the half space they belong to (8). Among the supervised learning methods, SVM is considered to be one of the most accurate techniques. The third DM application used in our program is called fuzzy logic, which can be defined as a further development of classical logic. Fuzzy logic, as the name suggests, is valuable in cases where it is difficult to apply classical logic to model a system based on the knowledge available (9,10). For example, fuzzy logic has been applied in the classification of benign and malignant nuclei in cytological images (11) and to control ventilator support during mechanical ventilation (12). Of all of the DM applications available, these three hybrid, distinct, but completely different methods (SVM, fuzzy

[1]Department of Pediatric Haematology and Oncology, Medical University, Hannover, Germany; [2]Data Mining Consulting, Donauwörth, Germany.
Correspondence: Lorenz Grigull (grigull.lorenz@mh-hannover.de)

Received 20 May 2011; accepted 18 February 2012; advance online publication 28 March 2012. doi:10.1038/pr.2012.34

logic, and ANN) were chosen and combined in a novel program for this study. A voting algorithm was used to combine the diagnoses of the three different methods. This was necessary to enable one final diagnosis if the three DM applications come to different diagnoses. Whereas in the current literature, DM applications were mostly used to distinguish between only two choices (prognosis good or dismal; a complication will/will not occur), we combined four different DM solutions to arrive at one of 18 diagnoses for children who presented in the ED. Accordingly, this study was performed to investigate the reliability of computer-generated diagnoses in patients who were admitted to the ED of a university hospital.

RESULTS

In total, 692 patients (data sets) who were admitted via the ED were included for development and testing of the diagnostic tool. The mean age of the study population was 6.5 y old with a standard deviation of 2.5 y.

Figure 1 shows a screen shot displayed during a bootstrap training run. For each single training step, 18 patients who were randomly chosen from the training set of the 566 data records received a computerized diagnosis, and this diagnosis was compared to the reference diagnosis. The reference diagnosis was the one posed by the doctors and was cross-checked using medical definitions. Each DM method used in the study (SVM, fuzzy logic, and ANN) computed one diagnosis. All three diagnoses were combined into a final computed diagnosis using a voting algorithm.

This function of our program (**Figure 1**) was used to optimize the computer program, but it should not be used in daily practice using our clinical decision support system. **Figure 1** illustrates two points for each patient.

First, **Figure 1** shows whether or not the computerized and reference diagnoses are identical. Second, the four colored squares and the black circle illustrate whether or not the different DM applications led to the same or different diagnoses. In patient "431ZUN302," all four DM systems diagnosed "nephrotic syndrome." Such homogeneous results strongly support a correct computerized diagnosis. Accordingly, in **Figure 1**, 17 of 18 patients were diagnosed correctly.

In the bootstrap trial run shown in this figure, one patient received a wrong diagnosis by our program: discordant diagnoses were assigned to patient 7 (281WEG095). This patient was admitted because of fever and abdominal pain. Laparoscopic surgery was performed to rule out acute appendicitis. The appendix appeared normal, but the anesthesiologist noted putrid sputum, and basal pneumonia was later confirmed. Of note, the SVM had diagnosed "appendicitis," but the ANN had chosen "pneumonia." The doctors in the hospital diagnosed "appendicitis," too, but misdiagnosed basal pneumonia sometimes causing severe abdominal pain.

These trial runs were used to optimize the performance of the program. In the end, for a total of 566 training records, we arrived at a correct diagnosis in 549 children (97%, **Table 1**). To be able to measure the performance of the training algorithm, we calculated discrete receiver operating characteristic (ROC) functions for the three DM methods as well as the voting method (**Figure 2**)

The rate of correct diagnoses during the retrospective data analysis for each group of patients can be seen in **Table 1**. Here, 37 of 38 children with appendicitis (group 1) were diagnosed correctly (**Table 1**), but only 39 of 42 children with bronchitis or asthma were diagnosed correctly. The program achieved even better results in diagnostic groups 4, 5, 9, 10, 11, 15, and 17, where the system diagnosed all children using the voting algorithm. Only the groups of children diagnosed with "gastroenteritis" and "asthma, bronchitis" contained more than two patients who received a wrong diagnosis. The area under the ROC curve (AUC) was calculated as a general measure of the performance of each single DM method (**Figure 2a–c**). **Figure 2a** displays the results of the voter for each diagnosis in the training set. The voter made the correct diagnosis in 97% of the cases and therefore underscores the assumption that the combination of several mathematical systems works better than any single DM method

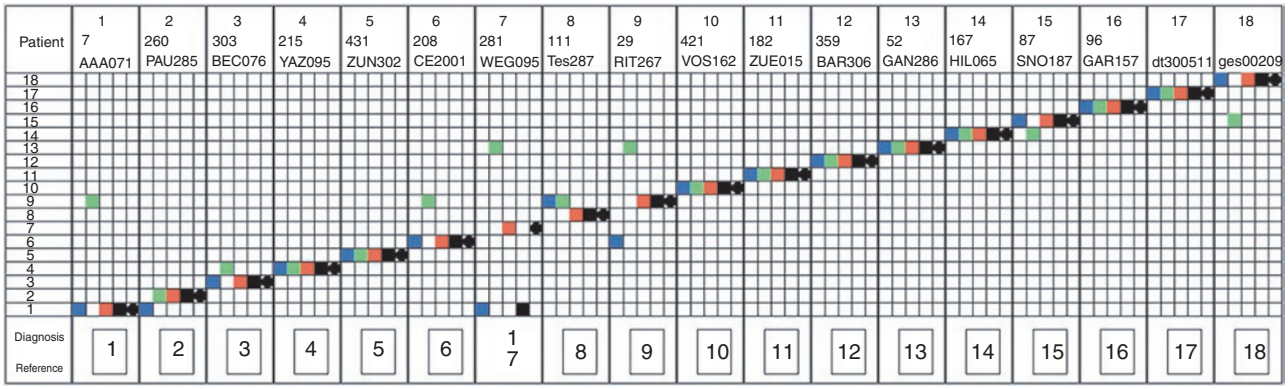


Figure 1. Result of one bootstrap run; 18 data sets were randomly chosen during a training procedure. The boxed numbers in the first line indicate the level of agreement between the reference and the computerized diagnosis, whereas two different subsequent numbers (**Figure 1**, patient 7; patient ID 281WEG 095) indicate different diagnoses between the computer and the reference. The diagnostic decision of the SVM is indicated with blue-filled square symbols, and the results of fuzzy logic in green-filled squares. The diagnosis of the ANN is denoted by a red-filled square symbol and the voter results in a black-filled square. The reference diagnosis is then indicated with a black-filled circle. ANN, artificial neural network; SVM, support vector machine.

Table 1. Details of patient diagnoses and diagnostic accuracy during retrospective and prospective testing with and without (w/o) voting function

Number of the different diagnostic groups	Diagnosis	Number of patients	Correct retrospective diagnoses with voter	Correct prospective diagnoses with voter	Correct prospective diagnoses w/o voter
1	Appendicitis (APP)	45	37/38	5/5	5/5
2	Abdominal disease, other than APP or enteritis (e.g., pancreatitis, hepatitis)	43	34/36	4/5	5/5
3	Gastroenteritis	45	35/38	4/5	5/5
4	Urinary tract infection	40	33/33	4/5	4/5
5	Nephrotic syndrome	29	22/22	2/5	5/5
6	Arthritis, coxitis	30	23/23	5/5	5/5
7	Pneumonia	54	45/47	4/5	5/5
8	Asthma, bronchitis	49	39/42	3/5	5/5
9	Other bacterial infection (e.g., sepsis, lymphadenitis)	56	49/49	4/5	5/5
10	Meningitis, encephalitis	31	24/24	4/5	5/5
11	Migraine, facial palsy, afebrile convulsions	48	41/41	3/5	5/5
12	Febrile convulsions	45	37/38	3/5	5/5
13	Vasculitis syndromes (HSP, HUS, SLE)	37	28/30	4/5	5/5
14	Malignant hematological disease (ALL, AML, NHL)	43	35/36	5/5	5/5
15	Benign hematological disease (ITP, TEC)	31	24/24	5/5	5/5
16	Diabetes mellitus, initial manifestation	28	20/21	5/5	5/5
17	H1N1 infection (bovine flu)	18	11/11	4/5	4/5
18	No diagnosis (healthy)	20	12/13	5/5	5/5
	Total	692	549/566 (97%)	73/90 (81%)	88/90 (98%)

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; HSP, Henoch–Schoenlein purpura; HUS, hemolytic uremic syndrome; ITP, idiopathic thrombocytopenic purpura; NHL, non–Hodgkin lymphoma; SLE, systemic lupus erythematosus; TEC, transient erythroblastopenia of childhood.

alone. The AUC values in **Figure 2** could be considered as a ranking of each single method, which varied between the prospective and retrospective tests of our system (**Figure 2a–c**).

The next question addressed in the study concerned the performance of the system during prospective testing. Records of ninety patients (five patients for each of the 18 diagnoses groups) were used exclusively for the prospective evaluation of the diagnostic tool.

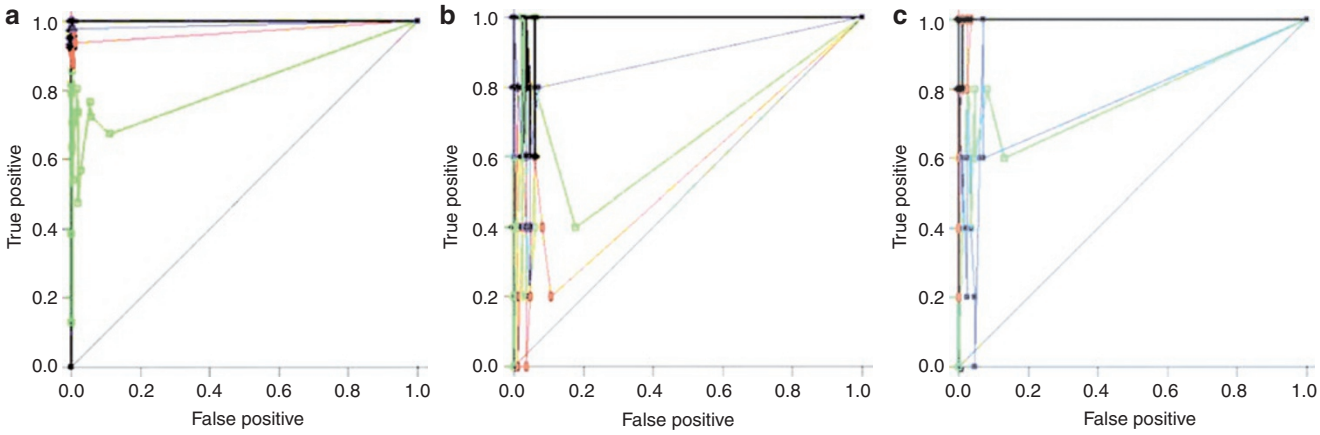
The prospective tests reached a rate of 81% correct diagnoses when the three DM methods were used together with the voting function (**Table 1**, column 5). During prospective testing, the voter diagnosed only two of five children with nephrotic syndrome (Diagnosis 5, **Figure 2b**) and three of five children of group 8 (patients with asthma and bronchitis). In contrast, for example, Diagnosis 14 (malignant hematological disease) and Diagnosis 15 (benign hematological disease) were diagnosed correctly by the voter in all patients during prospective testing. As a relevant measure of the performance during the prospective tests with and without the voter function, the ROC functions as well as the corresponding AUC values are shown in **Figure 2b** (prospective tests with voting function), illustrating the superiority of SVM as compared with the other DM methods used.

During prospective testing of the diagnostic tool, the correct diagnosis was made by at least one of the three DM methods in 98% (88 of 90 cases; **Figure 2c**, ROC curve on the right) of the children, and only two single data sets could not be recognized by any of the DM methods. **Figure 2c** illustrates this point.

The next part of the data analysis concerned the question of which diagnostic groups could be diagnosed the best through the use of the DM methods. Here, 31 of 32 children with appendicitis were diagnosed correctly using our program (**Table 1**). The program achieved even better results in diagnosing 20/20 children with "nephrotic syndrome," and 20/20 patients with "benign hematological disease" (idiopathic thrombocytopenic purpura and transient erythroblastopenia) were diagnosed correctly.

DISCUSSION

We challenged the provocative idea of generating a diagnosis in the "closed world" of a pediatric ED using a DM tool. Within an admittedly small choice of 18 diagnoses, the capability of modern mathematical methods was tested. In 2005, Mueller and coworkers reported the use of ANNs to predict whether extubation would succeed or not for preterm newborns (8).



Diagnosis	1	2	3	4	5	6
Results	27/38	34/36	35/38	33/33	22/22	23/23
Diagnosis	7	8	9	10	11	12
Results	45/47	39/42	49/49	24/24	41/41	37/38
Diagnosis	13	14	15	16	17	18
Results	28/30	35/36	24/24	20/21	11/11	12/13

Diagnosis	1	2	3	4	5	6
Results	5/5	4/5	4/5	4/5	2/5	5/5
Diagnosis	7	8	9	10	11	12
Results	4/5	3/5	4/5	4/5	3/5	3/5
Diagnosis	13	14	15	16	17	18
Results	4/5	5/5	5/5	5/5	4/5	5/5

Diagnosis	1	2	3	4	5	6
Results	5/5	5/5	5/5	4/5	5/5	5/5
Diagnosis	7	8	9	10	11	12
Results	5/5	5/5	5/5	5/5	5/5	5/5
Diagnosis	13	14	15	16	17	18
Results	5/5	5/5	5/5	5/5	4/5	5/5

Figure 2. (**a**) Retrospective emergency patients. In the table, the corresponding percentage of correct diagnoses is shown. For example, in diagnosis 1 (appendicitis), 37 of 38 patients received a correct computerized diagnosis. Blue line: AUC = 99% for SVM-ROC; green line: AUC = 82% for fuzzy-ROC; red line: AUC = 97% for neuro-ROC; black line: AUC = 100% for voter-ROC. (**b**) Prospective emergency patients with voter. Blue line: AUC = 88% for SVM-ROC; green line: AUC = 67% for fuzzy-ROC; red line: AUC = 58% for neuro-ROC; black line: AUC = 99% for voter-ROC. (**c**) Prospective emergency patients without voting. Blue line: AUC = 78% for SVM-ROC; green line: AUC = 78% for fuzzy-ROC; red line: AUC = 100% for neuro-ROC; black line: AUC = 100% for voter-ROC. AUC, area under the ROC curve; ROC, receiver operating characteristic; SVM, support vector machine.

According to their results, the ANNs proved to be more reliable than conventional statistics models (8). To establish the diagnosis from neuroimaging studies, Kloeppel and colleagues used SVM to classify magnetic resonance imaging scans of the brain (13). In their study, the computer classified Alzheimer's disease with a sensitivity of 95% (specificity 95%). In contrast, the radiologists classified only 80 to 90% of the cases correctly.

Misdiagnosis in the ED occurs in up to 15% of patients, and a broad range of diseases such as trauma, epileptic seizure, and Kawasaki syndrome are regularly misdiagnosed (14).

In this study, three different DM algorithms underscored their capability of allocating new patient data to the right diagnostic group. However, using the voter function in the case of different results between the DM applications resulted in 18% wrong diagnoses, illustrating the limitations of our approach. Nevertheless, the extent to which the thought-provoking suggestion of a smart computer application might result in a reduction in wrong diagnoses in the ED still remains unknown. In this case, a prospective study that will immediately and prospectively compare the rate of missed diagnoses with and without the diagnostic tool is planned.

Unlike systems that use the leading symptoms or main features of a disease to compute a diagnosis (15,16) and cross-check the items with medical textbooks, we developed a system that integrates both laboratory results and the chief complaints of the patients to arrive at a computer-borne diagnosis using modern mathematical techniques.

A comparison of DM applications was analyzed by Liew and colleagues, who compared two different DM methods: ANNs and decision trees, to model gallbladder disease in obese patients. They came to the conclusion that ANNs might be a useful tool for predicting the risk factors and prevalence of gallbladder disease and gallstone development in obese patients (17). Tsipouras and coworkers applied DM methods for the analysis of cardiac disease and electrocardiogram interpretations. According to one of their studies with different fuzzy logic applications, their program was able to classify cardiac arrhythmic beats (18). The major difference in our work is the implementation of three complementary DM applications in combination and the distinction of eighteen conditions (here: diagnoses), as compared with only two different decisions in the work of Tsipouras *et al.*

The system realized through the clinical decision support system called "ISABEL Healthcare" (http://www.isabelhealthcare.com) has gained a lot of interest in the field of computed clinical decision tools (4,16); ISABEL uses patient symptoms (pain, headache, fever) for a full-text search in textbooks and databases. Likewise, in the case of a complete and well-performed patient history and anamnesis, the system retrieves the correct diagnosis in up to 94% of cases (15).

Experienced clinicians might well question the value of a system diagnosing a patient having "other bacterial infections" or "abdominal disease, not classified." This acknowledged imprecision will be eliminated through the inclusion of larger patient numbers. However, the information "other bacterial infection" could be used to reduce excessive antibiotic therapy in nonbacterial infections. Another limitation of the current program version was the classification of a disease not represented by one of the 18 diagnostic groups. In this case, the program currently replies "classification into one group not possible. Consider

another diagnosis or uncommon disease presentation." This limitation will be eradicated during the expansion of the current pilot program version to more diagnostic groups.

The results presented still show limitations with respect to several aspects. First, the good result of 97% correct diagnoses was achieved only in the retrospective analysis but dropped to 81% correct diagnoses in the prospective analysis integrating the voting decision. This might be because of the fact that the construction, evaluation, and validation of the program were carried out on the same population sample ("in-sample analysis"), which may have led to over fitting of the model to the study population and therefore did not reflect the true validity of the questionnaire in daily clinical practice. The second limitation of the study was the relatively small sample size for each diagnostic group and the low number of different diagnoses. Increasing the number of patients and constructing the DM applications for different clinical scenarios, such as the detection of oncological or rare diseases, will remedy this.

To conclude, the art of making the right diagnosis should always remain in the hands of a doctor, but it is clear that less experienced doctors could particularly benefit from additional help and feedback. In this context, our study provides the first arguments in favor of using DM applications in the future.

METHODS

In the first step, data from all of the children admitted to our hospital via the ED in 2007 were scanned. Of this cohort of 1,880 children, 17 diseases and disease conditions ("diagnoses") representative for a pediatric ED and a healthy control group were selected, including important differential diagnoses or diagnostic categories (**Table 1**). Data from at least 18 patients per diagnosis were needed to enable the DM procedure. A total of 1,348 patients were excluded from the study. The majority of patients (n = 560) who were excluded had a preexisting and known diagnosis, for example, "cancer", and were admitted for chemotherapy or intravenous antibiotic treatment. In addition, 220 patients were admitted after trauma (e.g., head injury, fractures). A third large group who were not included comprised children admitted under the context of intoxication, icterus, or foreign body aspiration (n = 152). The remaining 532 children were distributed into 18 diagnostic groups, which resulted in an unequal distribution (e.g., appendicitis: n = 45). In five of the diagnostic groups, the number of patients in 2007 was too low to enable DM calculations (groups 5, 6, 13, 15, and 16). In this case, we screened all admissions in 2008 and 2009 and all children admitted via the ED. New children with diagnoses of nephritic syndrome, arthritis, vasculitis, or leukemia were subsequently included. This resulted in an uneven distribution of patients in the 18 diagnostic groups.

In the second phase of the study, we identified 26 parameters frequently investigated or measured in children in the ED whose medical situation could not be instantly diagnosed. Such a panel, including 14 clinical and 12 laboratory parameters (see data entry screen; **Figure 3**), was routinely used for sick children who presented at the ED in this study and formed the initial data for the construction of the diagnostic tool.

All patient data were then reviewed using laboratory results, clinical parameters, and details from patient medical histories. Protocols of surgical procedures and histopathological results were used where appropriate. Accordingly, a total of 692 patients were included in the study and stratified into one diagnostic group.

As the first step in the development of the program, the 692 data records were divided at random into three groups: 566 records were used for bootstrap training methods, 36 records (two per diagnostic group) were used for the validation procedure, and 90 records (five per diagnostic group) were used for prospective data tests (**Table 1**). The validation and test data sets were equally distributed between the 18 groups of patients (**Table 1**).

Although the data records referred to people, anonymized data records were used in the analysis. This study was approved by the institutional review board of the Medical University of Hannover. Informed consent was obtained from the patients or their legal guardians.

Mathematical Procedures

The fundamental idea of an SVM is the separation of different groups in an x-dimensional vector space using a mathematical kernel function to transform a patient's data points from the normal space up into an extended high-dimensional vector space, creating the optimal separation hyperplane. Any new patient can thus be allocated to predefined patient groups. Based on the work of Zadeh and coworkers, logical fuzzy variables are allowed to range continuously from 0.0 (false) up

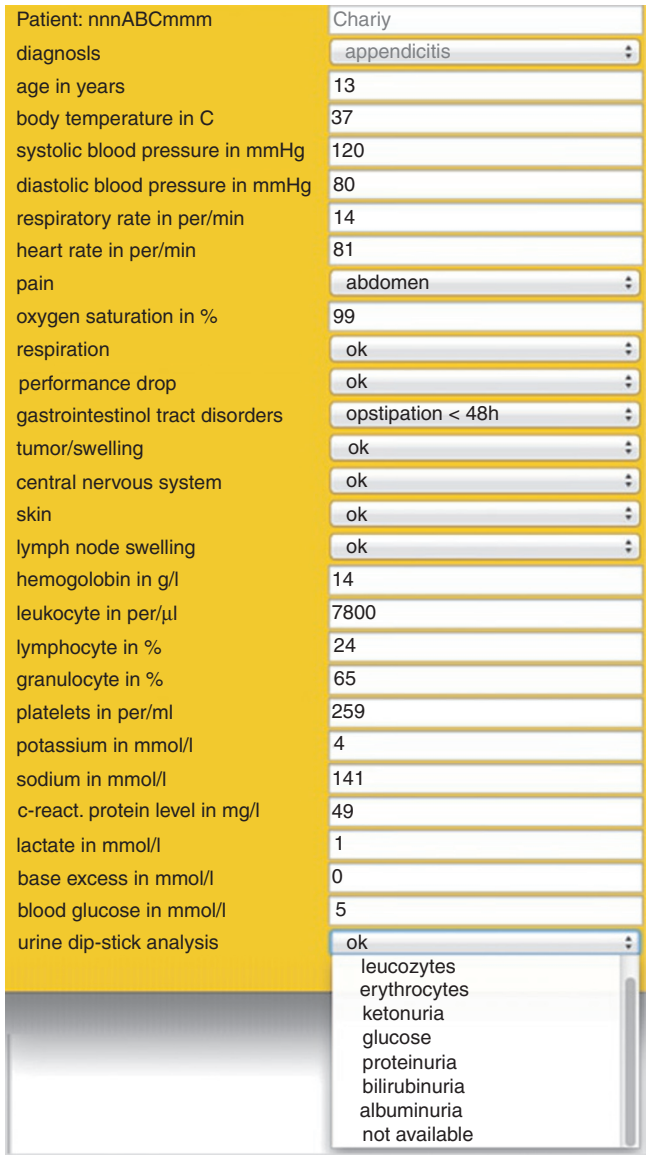


Figure 3. Data entry screen. On this screen shot, a patient with abdominal pain presented to the ED. The DM techniques came to the conclusion that the diagnosis was appendicitis, which was proven in the operation theater and later histologically. DM, data mining; ED, emergency department.

to 1.0 (true) and define the probability of a given statement (11). In all of the diagnostic groups, the data values for each patient correspond to a "fuzzy" variable ranging from 0 to 1 according to the Gaussian distribution of each value. In a new patient, these 26 fuzzy variables are then added for each diagnostic group, resulting in a specific value or probability for each of the 18 groups. In the fuzzy system, the new patient data set is then allocated to the diagnosis with the highest value.

Simulating simplified biological brains, ANNs work using forward and backward connected neurons and their dynamic electrical potentials. The artificial neuron fires an output signal if the total sum of the input signals exceeds the threshold. The output signals are then connected like synapses as inputs for other neurons in a large network of neurons. The numerical weight factors of each neuron input channel define the performance of the neural network. A backpropagation algorithm that takes into account the reference diagnoses can perform the calculation of these important weights. The recurrent Elman network was used to improve the performance of the neural network. This topology uses recurrent node connections and has demonstrated excellent performance and stability in many practical applications (8).

The neural network used in our study for medical diagnoses included 14,400 numeric weights distributed throughout three layers. The input layer included 100 parallel neurons, with each neuron gathering the 26 input signals, and there were 100 neurons in the hidden layer and 18 neurons in the output layer. Due to the Elman topology, the 100 output signals of the hidden layer were fed backward as additional input signals for the input layer neurons. The output layer performs the decision process. If, for example, the first neuron in the output layer delivers a 1.0 signal and all of the other 17 neurons switch down to −1.0, then this corresponds to the diagnosis with the number one (here: appendicitis). The relative deviation from this "ideal" output signal distribution allows a probability measure to be added to the diagnostic decision that corresponds to the scalar product of the output deviation vector.

Due to the different hybrid diagnostic methods, it is possible that the same patient (= one set of data) could receive three different diagnoses, one by each DM method, especially for unusual patient data. In this case, a voter algorithm calculates the final diagnostic decision by evaluating an optimized weighting function between the three diagnoses and their corresponding probabilities. The mathematical formula behind the voter function (d) of our system followed a specific calculation and therefore simply compared the probabilities (p) of the different DM applications (p_{svm}; p_{fuzzy}; p_{neuro}) for a given diagnosis z in order to calculate the final diagnosis. Accordingly, the term for the voter function is:

$$d=\left(\frac{p_{svm}\times z_{svm}+p_{fuzzy}\times z_{fuzzy}+p_{neuro}\times z_{neuro}}{p_{svm}+p_{fuzzy}+p_{neuro}}\right),$$

where z is the number of the diagnosis, and the index SVM indicates that the diagnosis is based on the calculation of the SVM. Accordingly, the values of neuro (= ANN) and fuzzy (= fuzzy logic) are given. The p values state the calculated probability of a diagnosis. Then, *d* is the optimal number that illustrates the closest relationship to one of the 18 diagnoses.

Example:

For one data set, the three DM applications come to the following diagnoses:

SVM → diagnosis zSVM = 1 with pSVM = 50% probability
Fuzzy → diagnosis zfuzzy = 2 with pfuzzy = 30% probability
Neuro → diagnosis zneuro = 8 with pneuro = 20% probability
Accordingly, the result for *d* follows the calculation:

$$d=\left(\frac{1\times 50+2\times 30+8\times 20}{50+30+20}\right)=\left(\frac{270}{100}\right)=2.7$$

In this example, the result of $d = 2.7$ is the closest to diagnosis 2, and therefore the voter will give diagnosis 2 as the output signal or diagnostic suggestion.

As key elements for successful data record training, the so-called bootstrap methods with the replacement of data records and ROC curves were applied. The AUC provides a measure of the accuracy of the training process, and the training procedure was stopped when the AUC (and as such the diagnostic accuracy of the system) reached a maximum for the actual voter algorithm.

The bootstrap run reflects the classical method of a stepwise adjustment of DM applications. For the development of our tool during each bootstrap run, 18 patient data sets (one from each group; $1 \times 18 = 18$) were randomly selected from the training group ($n = 566$). According to the result of the bootstrap run (→ correct diagnoses in the selection of 18 patients), the mathematical formulae of our system were slightly changed to optimize the tool. Then, after each bootstrap run, the (slightly changed) mathematical algorithm (e.g., the numerical weights in the ANN) was tested using the validation database. After this, the next bootstrap run followed using 18 different, randomly selected patient data sets from the training group. This two-step approach was repeated ~1,000 times until the results of the validation test were optimal. After optimal tuning of the system was reached, the parameters were stored and the system proceeded to the final step of prospective testing, which was realized by diagnosing the unknown data (five per diagnostic group; $5 \times 18 = 90$) set ("prospective test").

The prospective tests were useful for estimating how the diagnostic tool would perform using data records that were completely new to the system and simulated a daily life situation in the ED. The sophisticated bootstrap applied and the validation of the procedure emphasized the generalizing effect inherent in DM methods. While a low training error is necessary for practical purposes, a good performance of the system only depends on a high degree of the generalization capability. All three single diagnoses during a bootstrap run were displayed by the system to visualize alternative interpretations of the patient records (**Figure 1**).

ACKNOWLEDGMENTS

Clemens Betzel is gratefully acknowledged for a critical review of the manuscript.

STATEMENT OF FINANCIAL SUPPORT

This project was partly supported by a scientific grant from "Elternverein krebskranke Kinder Hannover e.V."

REFERENCES

1. Burroughs TE, Waterman AD, Gallagher TH, et al. Patient concerns about medical errors in emergency departments. Acad Emerg Med 2005;12:57–64.
2. Selbst SM, Friedman MJ, Singh SB. Epidemiology and etiology of malpractice lawsuits involving children in US emergency departments and urgent care centers. Pediatr Emerg Care 2005;21:165–9.
3. Budde T, Haude M, Höpp HW, et al. A prognostic computer model to individually predict post-procedural complications in interventional cardiology: the INTERVENT Project. Eur Heart J 1999;20:354–63.
4. Ramnarayan P, Cronje N, Brown R, et al. Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. Emerg Med J 2007;24:619–24.
5. Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996;347:1146–50.
6. Griffin MP, Lake DE, O'Shea TM, Moorman JR. Heart rate characteristics and clinical signs in neonatal sepsis. Pediatr Res 2007;61:222–7.
7. Chien CW, Lee YC, Ma T, et al. The application of artificial neural networks and decision tree model in predicting post-operative complication for gastric cancer patients. Hepatogastroenterology 2008;55:1140–5.
8. Mueller M, Wagner CL, Annibale DJ, Hulsey TC, Knapp RG, Almeida JS. Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modeling. Pediatr Res 2004;56:11–8.
9. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995.
10. Zadeh LA. A note on prototype theory and fuzzy sets. Cognition 1982;12:291–7.

11. Hafiane A, Bunyak F, Palaniappan K. Fuzzy Clustering and Active Contours for Histopathology Image Segmentation and Nuclei Detection. Lect Notes Comput Sci 2008;5259:903–14.
12. Schäublin J, Derighetti M, Feigenwinter P, Petersen-Felix S, Zbinden AM. Fuzzy logic control of mechanical ventilation during anaesthesia. Br J Anaesth 1996;77:636–41.
13. Klöppel S, Stonnington CM, Barnes J, et al. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain 2008;131(Pt 11):2969–74.
14. Anderson MS, Todd JK, Glodé MP. Delayed diagnosis of Kawasaki syndrome: an analysis of the problem. Pediatrics 2005;115:e428–33.
15. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23:Suppl 1:37–40.
16. Britto J. ISABEL at the helm. A web-based diagnosis system speeds clinical decisions for pediatric physicians. Health Manag Technol 2004;25:28–9.
17. Liew PL, Lee YC, Lin YC, et al. Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients. Dig Liver Dis 2007;39:356–62.
18. Tsipouras MG, Exarchos TP, Fotiadis DI, et al. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans Inf Technol Biomed 2008;12:447–58.