Artificial Neural Network Approach in Laboratory Test Reporting

Learning Algorithms

Ferhat Demirci, MD,[1,2] Pinar Akan, MD,[2,3] Tuncay Kume, MD,[3] Ali Riza Sisman, MD,[3] Zubeyde Erbayraktar, MD, PhD,[3] and Suleyman Sevinc, PhD[4]

From the [1]Clinical Biochemistry Laboratory, Dr Suat Seren Chest Disease and Thoracic Surgery Training and Research Hospital, Izmir, Turkey; and [2]Department of Neurosciences, The Institute of Health Sciences, [3]Department of Biochemistry Faculty of Medicine, and [4]Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, Izmir, Turkey.

Key Words: Neural networks (computer); Machine learning; Biochemistry; Clinical laboratory; Information systems; Autoverification

Am J Clin Pathol August 2016;146:227-237

DOI: 10.1093/AJCP/AQW104

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

ABSTRACT

Objectives: *In the field of laboratory medicine, minimizing errors and establishing standardization is only possible by predefined processes. The aim of this study was to build an experimental decision algorithm model open to improvement that would efficiently and rapidly evaluate the results of biochemical tests with critical values by evaluating multiple factors concurrently.*

Methods: *The experimental model was built by Weka software (Weka, Waikato, New Zealand) based on the artificial neural network method. Data were received from Dokuz Eylül University Central Laboratory. "Training sets" were developed for our experimental model to teach the evaluation criteria. After training the system, "test sets" developed for different conditions were used to statistically assess the validity of the model.*

Results: *After developing the decision algorithm with three iterations of training, no result was verified that was refused by the laboratory specialist. The sensitivity of the model was 91% and specificity was 100%. The estimated κ score was 0.950.*

Conclusions: *This is the first study based on an artificial neural network to build an experimental assessment and decision algorithm model. By integrating our trained algorithm model into a laboratory information system, it may be possible to reduce employees' workload without compromising patient safety.*

The accurate and timely reporting of laboratory tests, especially those with critical values, is as important a quality requirement as their proper execution.[1] The methods used to report test results should be rapid, reliable, and standard, with no variation between specialists.[2] The evaluation process can be standardized with validated and generally accepted algorithms defining the evaluation stages. Within the quality control system, these stages facilitate the identification of errors in patients' test results.[3]

Few studies to date have defined verification algorithm models for hematologic tests or biochemical tests such as glucose, sodium, and potassium. In these studies, the evaluation criteria were defined using logic-based, simple rule-based approaches (if-then-is).[4,5] These simple rule-based approaches have limited evaluation criteria and are not open to continued development. Today, the artificial neural network (ANN) approach, which enables more biomarkers to be used in the diagnosis of a disease, is by its very nature open to continued development.[6] The ANN approach allows the evaluation and cross-examination of large data sets involving highly nonlinear mathematical calculations. Based on our search of the literature, a verification algorithm model for laboratory result reporting has not been created using the ANN approach.

The aim of this study was to develop a decision algorithm model to report biochemistry test results using the ANN approach. Our model was designed to be integrated into laboratory information systems to improve efficiency and quality without compromising patient safety.

© American Society for Clinical Pathology, 2016. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com

Materials and Methods

Study Design

This study was conducted in a central laboratory of a university hospital. To create a decision algorithm model, first a training set occurred. A set including sample results of the patients was evaluated as approvable or rejectable by each biochemistry specialist without any time limitations. The consistency between the specialists' decisions and the reference control rules was revealed. The set with the highest κ value for agreement with the specialists' results was used as a training set for machine learning.

Then the new decisions were taken from the ANN model using different test sets (test set 1, test set 2). To evaluate the performance of the model, the decisions of the ANN model and biochemistry specialists were compared in two test sets. Also, the set of patients' results was reevaluated based on the reference control rules. Although the main contribution of the current study is intended to focus on the merits of the ANN approach as the automated predictive model, a decision tree model, based on the C4.5 algorithm,[7] was also built to further enhance and strengthen the study.

Finally, we also compared the ANN predictions and C4.5 predictions with those of the specialists' decisions. Test result data used in the study were generated by the Abbott Architect C16200 (Abbott, Lake Bluff, IL) and the Beckman Coulter (Brea, CA) AU 5800 analytical systems.

Participants

The laboratory test results of 252,847 samples, analyzed in the Dokuz Eylül University (DEU) Hospital Central Laboratory between April 2013 and April 2014, were evaluated. Results of zero or negative values, results not having numerical values in the laboratory information system, and results from samples with total request numbers less than 50% of the number required (24 parameters) were excluded from the study. Furthermore, samples from weekend days appeared to come mostly from intensive care units and emergency services. Since these samples contained a higher proportion of critical values, they were excluded to maintain the homogeneity of the study samples. We thus attempted to avoid biased learning by the ANN model. After applying the exclusion criteria, 100,621 samples were excluded, and the study was conducted on the remaining data pool consisting of 152,226 samples. From this data pool, the samples were randomly selected from the first week of each month to account for monthly and seasonal variance for evaluation by the specialists. In **Figure 1**, the Standards for Reporting Diagnostic Accuracy diagram is shown to report the flow of the participants throughout the study. For the training set, the number of data points was heuristically set to the square of the minimum number of attributes. The minimum required number for a training set was 576, whereas we included 1,847 data points in the training set of our study. Critical and noncritical result distribution in the training set is presented in **Figure 2**.

Test Parameters Used to Evaluate Sample Results and to Create a Training Set

In accordance with the related guideline (Auto 10-A) recommendations, when selecting tests for this study, we decided that it was appropriate to begin with a subset of tests (those that may have critical values) as opposed to the entire battery of tests routinely performed in the laboratory.[3]

Selected test parameters are listed in **Table 1**.[1,2,9,10,13] The parameters were chosen due to their high numbers of test requests, clear and established reference ranges, frequent delta check monitoring, auto analyzers' "flag" warning features, and remarkable monitoring of abnormal test results.[1,10] Critical value is defined as the test result obtained from a clinical specimen that is likely to indicate an acute risk to the health of the patient and a required medical intervention.[10] The critical tests (sodium, potassium, calcium, magnesium glucose, and uric acid) and their values used in this study were determined by consensus of DEU's central laboratory and clinicians. The other test results, which may be related to the critical test levels and have high request frequency (50% of the total requests), were also evaluated.

Also, eight different trials were done to determine the optimal test parameters (eg, alanine aminotransferase [ALT], aspartate aminotransferase [AST], or ALT/AST) and training set. To create different input types, we included results at critical values as well as a proportion of rejected samples corresponding to the rate of repeated and rejected samples during routine laboratory operation (∼9.8%). The training set with the highest κ value for agreement with the specialists' results was used for the remainder of the study **Table 2**.

Creation of Reference Control Rules

Reference control rules used in this study were defined based on the literature. First, the randomly selected 1,500 results were entered into the FileMaker 8.5 Pro (FileMaker, Santa Clara, CA) software and evaluated by seven different biochemistry specialists in a virtual laboratory information system. The specialists were asked to either approve or reject the test results, and reasons for rejection were noted. These results were evaluated also based on the reference control rules as approvable or rejectable. The consistency between the seven specialists' decisions and the reference control rules was revealed. In cases where there was a difference between the reference control rules and the

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

DOI: 10.1093/ajcp/aqw104
© American Society for Clinical Pathology

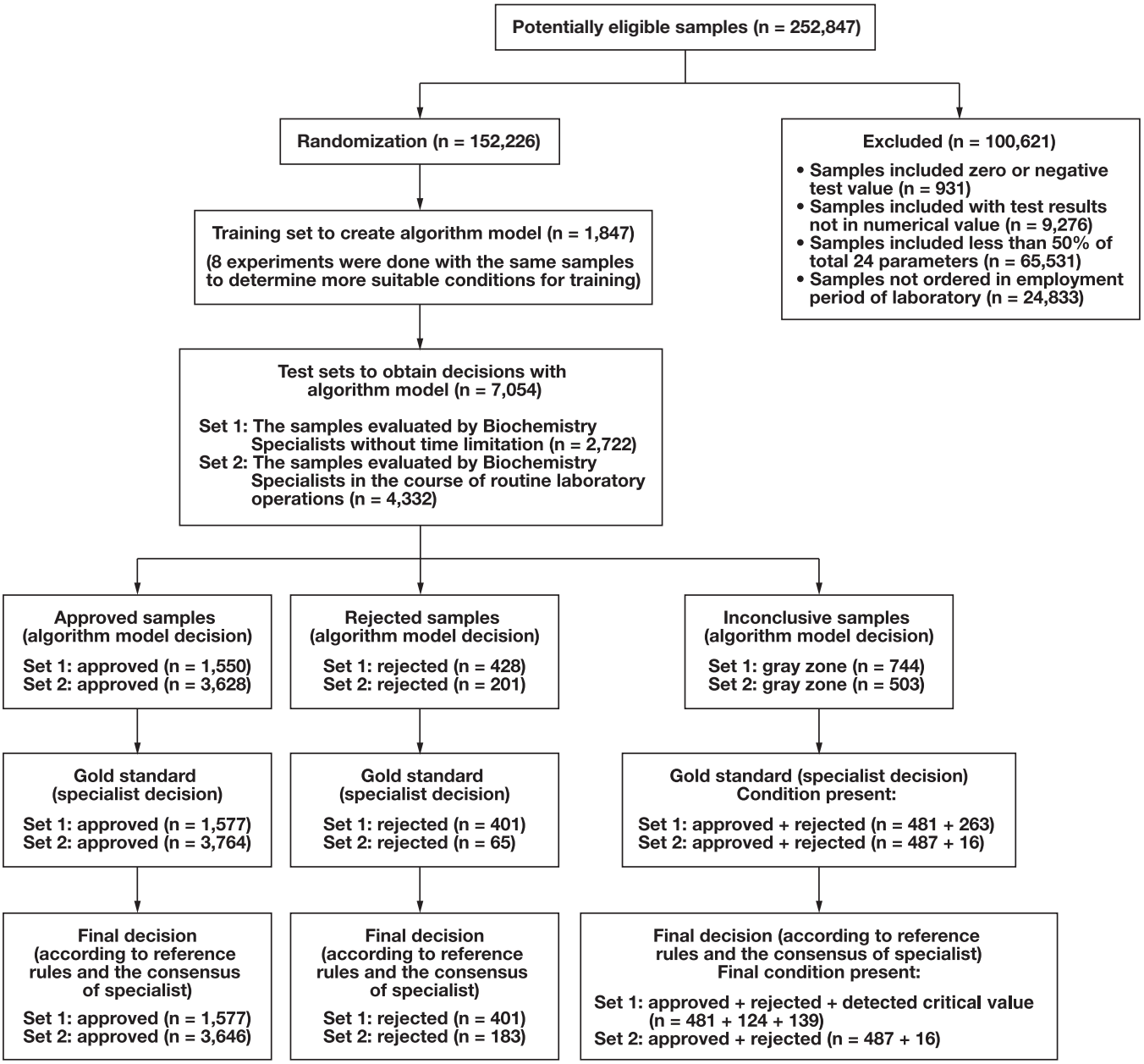


Figure 1 Standards for Reporting Diagnostic Accuracy diagram to report flow of the participants through the study.

decisions of at least four of the specialists, all specialists were asked to reevaluate and give a final decision. The reference control rules were revised based on the final decision. Finally, another 1,847 results were evaluated as either approvable or rejectable, and these results were used as a training set for the ANN model.

The reference control rules defined in this study can be summarized as follows.

Index Values

If samples had hemolytic, icteric, and/or lipemic index values over a level that may have resulted in interference, their results were considered erroneous and not reported.[8,11,12] The interference limits of the index values for the tests selected for our experimental model were taken from related companies.

Critical Value Control

Consistency Check

To detect whether preanalytic error sources had an effect on parameters reflecting patients' clinical status, we took the following parameters into account.

Delta Control

Test variance is dependent on biological variance (within-subject CV_I%) and analytical variance (CV_A%). Analytical

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

© American Society for Clinical Pathology

DOI: 10.1093/ajcp/aqw104

variance is the rate of change between the first and second test results during a specified time period (%): [(Second Test Result – First Test Result) × 100/Second Test Result].[13]

Reference change values (RCVs) can be expressed as follows:

$RCV = 2^{1/2} \times Z \times (CV_A^2 + CV_I^2)^{1/2}$ ($Z = 2.58$ for 99% confidence interval). If variance over the RCV rate occurs within the specified time frame and is not attributed to any clinical status, it is assessed as an inconsistency.[13]

Possibility of Analysis From an Intravenous Catheterized Arm[14]

Sodium (Na) more than 160 mmol/L, chloride (Cl) more than 110 mmol/L, potassium (K) less than 3.5 mmol/L, glucose less than 60 mg/dL and/or positive (+) delta in Na and Cl, and negative (–) delta in glucose and K are considered rule violations, and the results are not approved due to the possibility of the sample being taken from an arm receiving intravenous physiologic saline.

Na less than 136 mmol/L, Cl less than 98 mmol/L, K more than 5.5 mmol/L, glucose more than 110 mg/dL and/or (–) delta in Na and Cl, and (+) delta in glucose and K are considered rule violations, and the results are not approved due

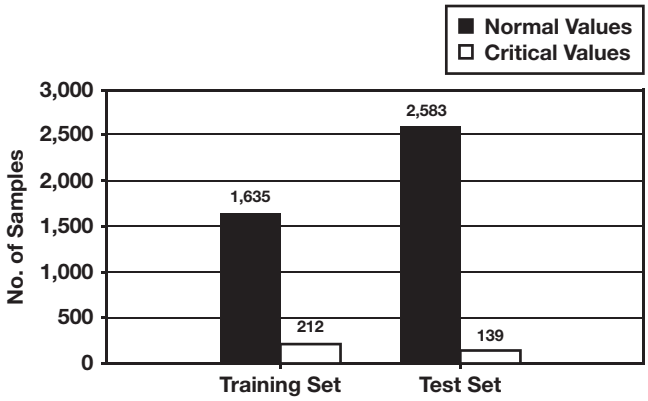


Figure 2 Features of the training set and test set 1. The training set comprised 1,847 patient samples; test set 1 comprised 2,722 patient samples.

Table 1
Parameters Used to Create a Decision Algorithm Model (Artificial Neural Network)[a]

Critical Tests[b]	Related Tests[c]	Delta Values[d]	Other Variables[e]
Sodium (Na^+)	Chloride (Cl)	Delta Na^+	Patient age
Potassium (K^+)	Serum urea nitrogen	Delta K^+	Hemolytic index
Calcium (Ca^{++})	Creatinine	Delta Ca^{++}	Lipemic index
Magnesium (Mg^+)	Aspartate aminotransferase	Delta Mg^+	Icteric index
Glucose	Alanine aminotransferase	Delta glucose	
Uric acid	γ-Glutamyl transferase	Delta uric acid	
	Alkaline phosphatase	Delta Cl^-	

[a]In our study, a total of 24 parameters obtained from the laboratory information system were used to create a decision algorithm model. The parameters are shown in four groups in the table.
[b]Tests that have critical values defined previously in our hospital. Critical value is defined as the test result obtained from a clinical specimen that is likely to indicate an acute risk to the health of the patient and a required medical intervention.[10]
[c]Related tests (they have no defined critical values, but their levels can change relative to critical test levels[9]).
[d]Delta values are defined as the ratio of test variance that is calculated using the biological variance and analytical variance value of the test. Analytical variance is the rate of change between the first and second test results during the specified time period (%): [(Second Test Result – First Test Result) × 100 / Second Test Result].[13] The detailed calculation formula is shown in the Materials and Methods section.
[e]Other variables that have been widely used to evaluate the suitability of clinical specimens for test reporting in the clinical laboratory.[1,2]

Table 2
Features of Experiments Run With Different Conditions to Obtain a More Suitable Training Set

Experiment No.	Are There Instances With Critical Values?	Is There Population Selection?	Are There Test Ratios Like AST/ALT?[a]	Negative Cutoff[b]	Positive Cutoff[b]	κ Score
1	–	–	–	–0.3650	0.6833	0.780
2	–	+	+	–0.7470	0.6400	0.738
3	+	–	+	–0.2133	0.3958	0.701
4	–	–	+	–0.4060	0.5755	0.693
5	+	–	–	0.0705	0.7948	0.649
6	–	+	–	0.0735	0.5998	0.605
7	+	+	+	–0.0001[c]	0.0001[c]	0.430
8	+	+	–	–0.0001[c]	0.0001[c]	0.321

ALT, alanine aminotransferase; AST, aspartate aminotransferase; +, positive; –, negative.
[a]Were tests such as ALT and AST taken separately, or were test ratios like AST/ALT used? –, used the tests only; +, used the test ratios.
[b]The area between the negative and positive cutoff points represents the gray zone.
[c]To fit within the cutoff points, the κ values were calculated using zero as the cutoff point.

DOI: 10.1093/ajcp/aqw104
© American Society for Clinical Pathology

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

to the possibility of the sample being taken from an arm receiving an intravenous dextrose.

Control of Samples With Fibrin

Pipetting errors due to fibrin in a sample are easily recognized by the immediate fall in all analytical results. Na less than 136 mmol/L, K less than 3.5 mmol/L, calcium (Ca) less than 8.4 mg/dL, glucose less than 70 mg/dL, and (–) deltas for most of these tests are accepted as rule violations, and the decision is made not to approve the results.

EDTA Contamination Control

Ca less than 4 mg/dL, K more than 10 mmol/L, (–) delta for Ca, and (+) delta for K indicate rule violations, and the decision is given to not report the results due to the possibility of EDTA in the sample.[15]

Control of Samples With Delayed Analysis

Glucose less than 40 mg/dL, K more than 6 mmol/L, and a hemolytic index less than 50 suggest that a sample's analysis was delayed and is considered a rejection criterion.[15]

Control of Samples Incompatible With Life

Some samples, especially those from the intensive care unit, yield results that biochemistry specialists consider incompatible with life, such as Na of 185 mmol/L, K of 8.5 mmol/L, and Ca of 15 mg/dL. Certain special rules were determined as rejection criteria based on the specialists' judgment (data not shown).

Limit Control

Since most patients admitted to the hospital have health problems, the test results of their samples generally fall outside the reference intervals and are at a certain limit.[16] In the current study, the limit control points were defined according to the reference intervals and total allowable error rates of each test[17] and are summarized in **Table 3**.

The operation of reference control rules is summarized in **Figure 3**.

Table 3
Test Limit Control Points

Test	Reference Range[a]	% TEa[b]	Control Range
Sodium, mmol/L	136-146	6.10	127.70-154.91
Potassium, mmol/L	3.5-5.5	9	3.19-6.00
Chloride, mmol/L	98-107	9	89.18-116.63
Calcium, mg/dL	8.6-10.2	11	7.65-11.32
Magnesium, mmol/L	0.7-1.05	16	0.59-1.22
Uric acid, mg/dL	2.4-7	17	1.99-8.19
Glucose, mg/dL	70-110	10	63.00-121.00
Aspartate aminotransferase, U/L	5-34	20	4.00-40.80
Alanine aminotransferase, U/L	0-55	20	0.00-66.00
γ-Glutamyl transferase, U/L	12-64	22	9.36-78.08
Alkaline phosphatase, U/L	40-150	30	28.00-195.00
Serum urea nitrogen, mg/dL	8.9-20.6	9	8.10-22.45
Creatinine, mg/dL	0.6-1.3	22	0.47-1.59

[a]Reference ranges recommended by the respective companies.
[b]Total allowable error (www.westgard.com/biodatabase1.htm).

Algorithm Model Creation Using the ANN Approach

The selected data were processed with the Weka 3.6.2 (Weka, Waikato, New Zealand) software, which implements a number of machine-learning algorithms. We used the Weka software due to its ease of use, easily accessible training documents, and compatibility with various operating systems.

The Weka software created rules based on the training set and calculated a function result called a "predicted value" as output. The desired outputs are a single result; however, the mathematical formula cannot always meet this condition. To avoid this, the outputs were organized according to specific cutoff points. The cutoff points we used in the model were based on the prediction points of a receiver operating characteristic (ROC) curve analysis. Due to the double-peak distribution of the predicted values produced, the data were sorted in ascending order to determine decision cutoff values. Values at the 25th and 75th percentiles were accepted as decision prediction points (**Figure 4** and Table 2). For positive values, predicted values over the 25th percentile were determined as approvable (+1); for negative results, those under the 75th percentile were determined as rejectable (–1). Values in the middle (the gray zone) were considered rejectable or requiring reassessment.

In the experiments, two kind of sets were used: a training set and a test set. The training set comprised test results required to create the learning model, while the test set comprised results with which to assess the model. Data in the training set labeled whether the results were evaluated as approvable or rejectable; data in the test set were used to demonstrate whether the model appropriately evaluated results. First, the outcomes of ANN model were compared with those of specialists' common evaluation by using test set 1. By analyzing the outcomes of test set 1, we considered that the accuracy of the algorithm model was highly acceptable. Then we applied the algorithm model to the test results of the routine laboratory operation (RLO). Of 152,226 test results of the RLO, we continued to study 4,332 test results (test set 2).

In our study, the accuracy, precision, sensitivity (true-positive rate), and specificity (true-negative rate) of the evaluation algorithm model were determined, accepting the final decisions of the specialists as the gold standard. Other studies in the literature have accepted multiple experts' joint decisions on test results as the gold standard, as in our study.[2,16,18]

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

© American Society for Clinical Pathology

DOI: 10.1093/ajcp/aqw104

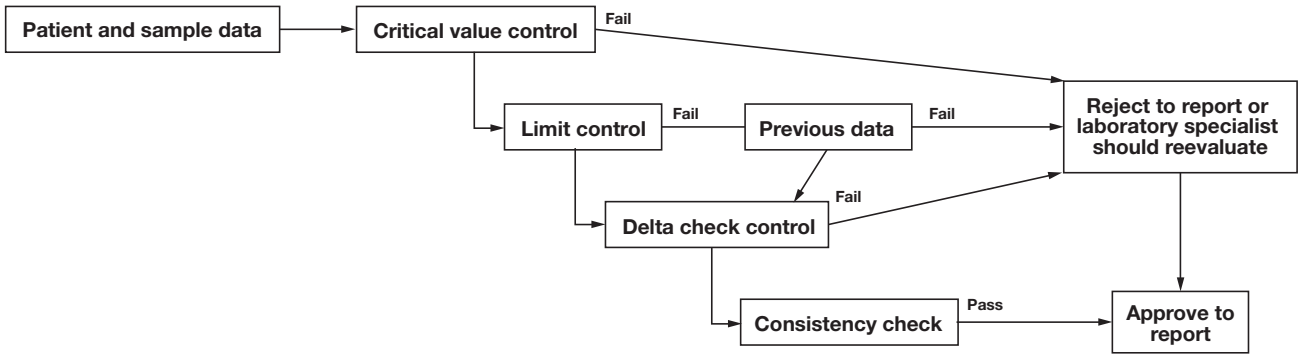


Figure 3 Sequences of reference control rules created according to the literature and common opinions of biochemistry specialists.

Furthermore, to gauge the validity of our model, we compared the outcomes with those of another machine-learning algorithm, the C4.5 decision tree model, in test set 2.

Details of the characteristics of the ANN model are provided as supplementary information (all supplemental materials can be found at *American Journal of Clinical Pathology* online).

Ethics Approval

The study was approved by the DEU Central Laboratory administration and the DEU Faculty of Medicine Non-Invasive Research Ethics Committee (Resolution Number 2013/05-04, dated February 14, 2013).

Statistical Analysis

The optimal cutoff points of predictive values in the ANN outcomes were determined by ROC curve analysis. The κ analysis and McNemar test were used to assess the agreement between the decisions produced by the algorithm created by the Weka software and the decisions given by specialists based on the reference control rules. Interrater agreement between the seven clinical chemists was investigated using the κ analysis. The sensitivity and specificity values of the ANN model were determined according to the specialists' decisions. *P* values less than .05 were accepted as statistically significant.

Results

The results of the specialists' evaluation according to the reference control rules had κ values between 0.49 and 0.91. Detailed analysis revealed that the κ values were between 0.81 and 1.00 (perfect agreement[19]) for five of seven specialists, between 0.61 and 0.80 (significant agreement[19]) for one of seven, and between 0.41 and 0.60 (moderate agreement[19]) for one of seven **Table 4**. The interrater κ scores are shown in **Table 5**. Six of the specialists showed at least significant agreement with each other.

The sensitivity of the specialists' decisions ranged from 38.1% to 88.3%, and their specificity ranged from 89.6% to 98.8%. Since there was a statistically significant difference between the predictions of the ANN model and biochemists' decisions ($P < .05$), the results were reassessed and ANN trials were repeated for different conditions.

At the optimal cutoff point determined in our ROC curve analysis, the sensitivity was 87.9%, specificity was 100%, and the area under the curve was 0.981. The positive and negative cutoff points were determined as 0.8878 (values at the 25th percentile and higher were accepted as +1) and –0.9410 (values at the 75th percentile and lower were accepted as –1), respectively.

Outcomes of the test set 1 are shown in **Table 6** and **Table 7**. The κ value of the set was 0.941, so the ANN model was accepted as valid. Afterward, the model was applied to the test results in the RLO (test set 2). Sex distribution of the specimens was 54.8% from female patients and 45.2% from male patients. In test set 2, of 4,332 test results, 503 results fell in the gray zone and 3,829 results could be evaluated at the calculated cutoff points. The predictions of the ANN model in this experiment are shown in **Table 8**. Initially, the model sensitivity was 27.8% and the specificity was 99.7%. Reassessment of the samples that had received opposing decisions from the ANN model revealed that the nine samples with false-positive results had been rejected due to hemolysis, despite hemolytic index values of 0 as measured by an autoanalyzer. When these results were reassessed by the specialists, they were approved **Table 9**. Investigation of the results that the specialists verified but were rejected by the ANN model revealed that 127 of 145 were approved by the specialists despite a high hemolytic index. When these samples were reassessed, the specialists agreed that those samples should be rejected as proposed by the ANN model. At the end of this process, there were 18 patients' results that could not be reported accurately.

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

DOI: 10.1093/ajcp/aqw104
© American Society for Clinical Pathology

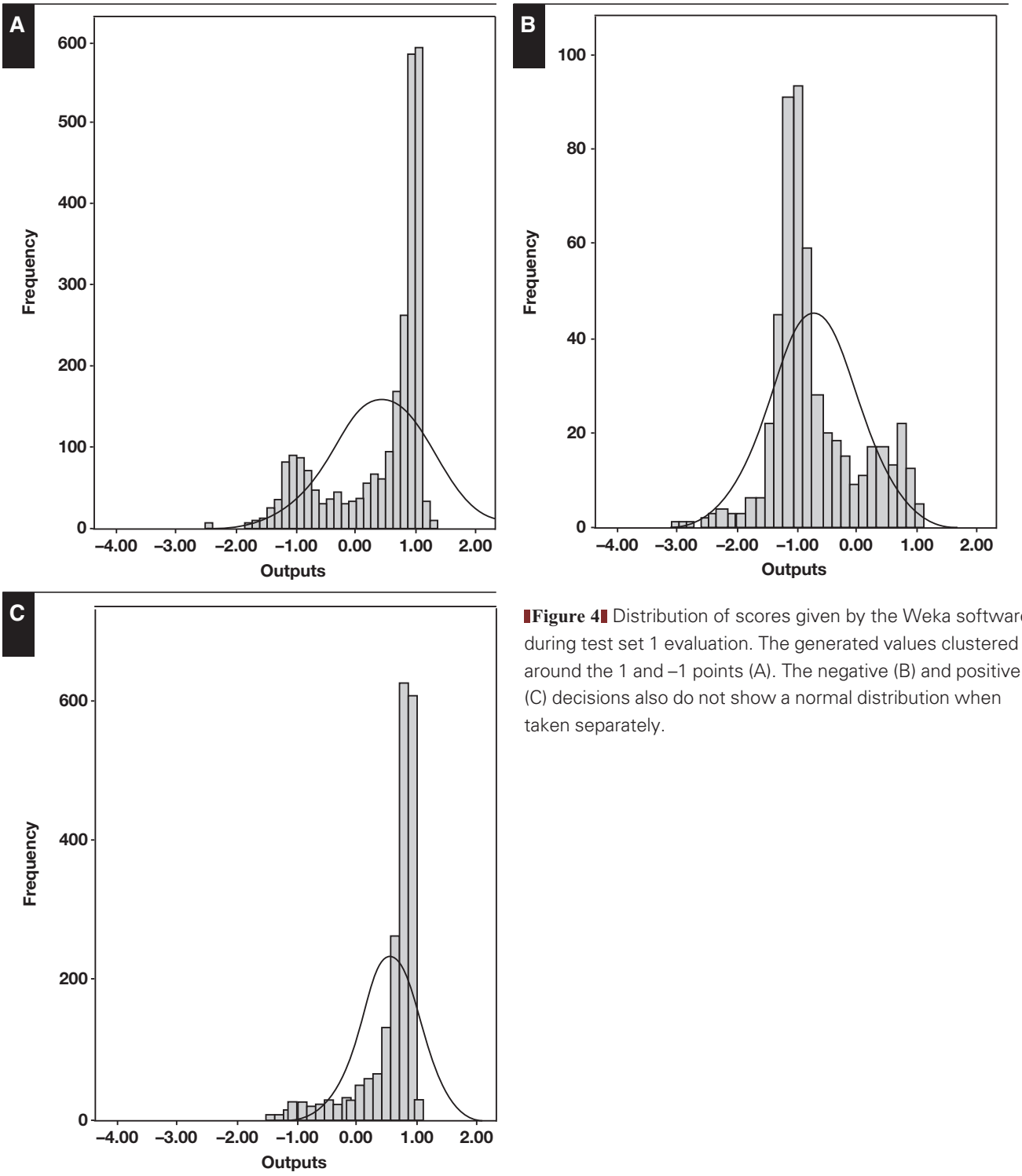


Figure 4 Distribution of scores given by the Weka software during test set 1 evaluation. The generated values clustered around the 1 and –1 points (A). The negative (B) and positive (C) decisions also do not show a normal distribution when taken separately.

The Weka software was trained by repeated application of data sets, including results from different brands of the analyzer. Following this training, patients' results in the RLO yielded no results that were rejected by specialists but approved by the ANN model, whereas there were 18 (0.5%) results approved by specialists but rejected by our model **Table 10**. Our first κ score was 0.406. After reevaluation by the specialists, the κ score increased to 0.95 and the model sensitivity reached a high of 91%. As the false-positive rate was 0%, the model specificity reached a high of 100%.

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

© American Society for Clinical Pathology

DOI: 10.1093/ajcp/aqw104

Table 4
Evaluation of the Biochemists' Decisions Relative to the Reference Control Rules (n = 1,500)[a]

Characteristic	CC 1	CC 2	CC 3	CC 4	CC 5	CC 6	CC 7	Reference Control Rules
Validated (V)	1,102 (73.5)	1,108 (73.9)	1,120 (74.7)	1,324 (88.3)	1,118 (74.5)	1,108 (73.9)	1,098 (73.2)	1,067 (71.1)
Rejected (R)	398 (26.5)	392 (26.1)	380 (25.3)	176 (11.7)	382 (25.5)	392 (26.1)	402 (26.8)	433 (28.9)
Correct validation (CV)	1,057 (95.9)	1,055 (95.2)	1,056 (94.2)	1,065 (80.4)	1,033 (92.4)	1,038 (93.7)	1,028 (93.6)	1,067 (100)
Correct rejection (CR)	338 (84.9)	304 (77.5)	317 (83.4)	160 (90.9)	288 (75.6)	282 (72.1)	275 (68.4)	433 (100)
Incorrect validation (IV)	45 (4.1)	53 (4.8)	64 (5.8)	259 (19.6)	85 (7.6)	70 (6.3)	70 (6.4)	0 (0)
Incorrect rejection (IR)	60 (15.1)	88 (22.5)	63 (16.6)	16 (9.1)	93 (24.4)	109 (27.9)	127 (31.6)	0 (0)
Sensitivity, %	88.3	85.1	83.2	38.1	77.2	80.1	79.7	100
Specificity, %	94.8	92.6	94.7	98.8	92.3	91.4	89.6	100
κ Score	0.91	0.89	0.87	0.49	0.80	0.83	0.82	1.00

CC, clinical chemist.
[a]Values are presented as number (%) unless otherwise indicated. Sensitivity: CR / CR + IV (True Positive / True Positive + False Negative).
Specificity: CV / CV + IR (True Negative / True Negative + False Positive).
In our study, *true positive* means that problems with the sample are determined correctly and the results are rejected.

Table 5
Interrater Agreement Between the Seven Clinical Chemists

	κ Score						
Chemist	CC 1	CC 2	CC 3	CC 4	CC 5	CC 6	CC 7
CC 1	1.00	0.98	0.97	0.60	0.86	0.86	0.84
CC 2	0.98	1.00	0.95	0.58	0.85	0.85	0.83
CC 3	0.97	0.95	1.00	0.57	0.85	0.85	0.82
CC 4	0.60	0.58	0.57	1.00	0.50	0.51	0.50
CC 5	0.86	0.85	0.85	0.50	1.00	0.76	0.74
CC 6	0.86	0.85	0.85	0.51	0.76	1.00	0.95
CC 7	0.84	0.83	0.82	0.50	0.75	0.95	1.00

CC, clinical chemist.

Table 6
Comparison of All Biochemistry Specialists' Evaluation and the Outputs of the Decision Algorithm Model (Artificial Neural Network) in Test Set 1

Artificial Neural Network	Clinical Biochemists, No. (%)[a]		
	Rejected	Validated	Total, No. (%)
Rejected	395 (98.5)	33 (2.1)	428 (21.6)
Validated	6 (1.5)	1,544 (97.9)	1,550 (78.4)
Total	401 (100.0)	1,577 (100.0)	1,978 (100.0)

[a]The samples evaluated by biochemistry specialists without time limitations.

Table 7
Sensitivity and Specificity of a Decision Algorithm Model (Artificial Neural Network) for Test Set 1 Results

Characteristic	%	95% Confidence Interval, %
Sensitivity	92.2	89.2-94.5
Specificity	99.6	99.1-99.8
True positive	98.5	96.6-99.3
False positive	1.5	0.6-3.3
True negative	97.9	97.0-98.5
False negative	2.1	1.5-3.0

Table 8
Comparison of Biochemistry Specialists' Evaluation and the Outputs of the Decision Algorithm Model (Artificial Neural Network) in Test Set 2 (Routine Laboratory Results)[a]

Artificial Neural Network	Clinical Biochemist, No. (%)[a]		
	Rejected	Validated	Total, No. (%)
Rejected	56 (86.2)	145 (3.9)	201 (5.2)
Validated	9 (13.8)	3,619 (96.1)	3,628 (78.4)
Total	65 (100)	3,764 (100)	3,829 (100)

[a]The samples evaluated by biochemistry specialists in the course of routine laboratory operations.

When the C4.5 decision tree algorithm model was applied to the data used in our test set 2, none of the patients' results fell in the gray zone; the sensitivity and specificity were 48.9% (43.9%-53.9%) and 99.9% (99.8%-100%), respectively; and the κ value was 0.633.

Discussion

Systems using ANNs are machine-learning algorithms and therefore can learn and test what they have learned. During the learning phase, the data can be divided into groups; one group is used as the basis of learning, while the others can be used for validation to reach the most accurate mathematical formula.[20] These systems have not yet been used for the evaluation of medical test results. According to our literature search, this study represents the first development of an ANN-based test result evaluation model that can be integrated into laboratory information systems to assist biochemistry specialists. Our algorithm model learns through the use of different training sets, and the final decisions produced through this process are highly compatible with decisions made by biochemistry specialists.

The development of this type of algorithm model can involve the use of various approaches, including ANN,

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

DOI: 10.1093/ajcp/aqw104
© American Society for Clinical Pathology

Table 9
Properties of Samples Given False-Positive Decisions by the Decision Algorithm Model (ANN) in Test Set 2

Clinical Biochemist	Conclusion (–1: Rejected)	Reason for Rejection	ANN Predicted Value	Error	ANN Conclusion (+1: Validated)	HEM_INDX	ICT_INDX	LIP_INDX
Assistant 3	–1	Hemolysis	1.02	2.02	+1	0	0	0
Specialist 1	–1	Hemolysis	1.08	2.08	+1	0	0	0
Specialist 1	–1	Hemolysis	1.03	2.03	+1	0	0	0
Specialist 1	–1	Hemolysis	1.05	2.05	+1	0	0	0
Specialist 1	–1	Hemolysis	1.02	2.02	+1	0	0	0
Specialist 1	–1	Hemolysis	0.91	1.91	+1	0	0	0
Assistant 3	–1	Hemolysis	1.04	2.04	+1	0	0	0
Assistant 3	–1	Hemolysis	1.08	2.08	+1	0	0	0
Assistant 3	–1	Hemolysis	0.96	1.96	+1	0	0	0

ANN, artificial neural network; HEM_INDX, hemolytic index; ICT_INDX, icteric index; LIP_INDX, lipemic index.

Table 10
Comparison of Biochemists' Evaluations and Algorithm Model (Artificial Neural Network) Decisions for Corrected Routine Results in Test Set 2[a]

Artificial Neural Network	Clinical Biochemist, No. (%)[a]		
	Rejected	Validated	Total, No. (%)
Rejected	183 (100)	18 (0.5)	201 (5.3)
Validated	0 (0)	3,628 (99.5)	3,628 (94.7)
Total	183 (100)	3,646 (100)	3,829 (100)

[a]Clinical biochemists reevaluated the samples in test set 2.

logistic regression, decision trees, and support vector machines. We chose to employ the ANN approach in this study because it presents advantages over other approaches in several respects. ANN is a complex collection of weighted logistic regression (sigmoid) boxes. The logistic regression method can offer a probability value for each decision outcome, which requires some interpretation, and it may not be intuitively obvious.[21]

ANN classifies by optimizing a highly nonlinear total loss function, thus allowing simulation of highly complex and nonlinear decision boundaries. In contrast, decision trees use comparison operators on numerical attributes, such as less than or greater than, and they typically construct decision boundaries by approximation using these operators. As a consequence of this, ANNs can generate smoother decision surfaces than decision trees.[7,22-24] In our study, the results about the predictions of the ANN model and C4.5 decision tree algorithm model were supported by this information. Our study showed that the sensitivity and specificity of the ANN model were higher than the decision tree model.

There are studies in the literature describing the use of Boolean logic to report medical test results. In these studies, between 13 and 4,500 Boolean logic–based verification rules were installed as "middleware" and enabled an autoverification rate of approximately 80% to 93% through a laboratory information system.[16,25-28] They have been shown to decrease the overall laboratory workload by up to 70%.[25] In another study, 13 Boolean logic–based autoverification rules for hormone and tumor markers were installed on a laboratory information system without middleware, and an autoverification rate of 80% was achieved.[27] Torke et al[28] installed autoverification rules for clinical chemical and urine analysis results into a laboratory information system. They reported autoverification of 62% of clinical chemical test panels, 73% of results from a single analyte, and 43% of urine analyses. In addition, they calculated that the system provided an annual workforce gain equivalent to 5.5 full-time employees. These fabricated rules are also known to decrease the proportion of patients requiring manual evaluation and shorten turnaround times.[5] Limited evaluation and cross-check capacities as well as not being open to continued development can be considered the disadvantages of these studies.

Oosterhuis et al[2] developed a statistical method of autoverification using 33 parameters and reported an autoverification rate of 86.6% of the total samples with a sensitivity of 77.9% and a specificity of 88.6%. Their lower sensitivity and specificity values compared with those in our study may be attributable to many samples being caught in "delta control" evaluation, as most of the study's samples came from inpatients with serious medical issues. To increase the sensitivity and specificity of the model we developed, diagnostic frameworks related to delta control changes were evaluated by the specialists while creating the training sets, which allowed the ANN to learn the specialists' approach. After implementing at least three different "learning sets" with the Weka software, the sensitivity and specificity of the resulting predictive values increased to 91% and 100%, respectively.

The current study also compared the decisions given on the test results by a virtual laboratory information system and seven different biochemistry specialists based on reference control rules. The incorrect approval rates of the specialists ranged from 4.1% to 19.6%, and the incorrect rejection rates varied from 9.1% to 31.6%. These rates are consistent with

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

© American Society for Clinical Pathology

DOI: 10.1093/ajcp/aqw104

those found in the literature.[2] The legal liabilities that may arise from incorrect approval may cause greater avoidance of incorrect approval compared with incorrect rejection. Incorrect rejections do not have legal or professional consequences, as they prevent results from reaching patients, although they do affect statistical analyses.

The κ analysis revealed agreement scores of 0.49 to 0.91 with the seven specialists' reference decisions. Although five of the specialists showed perfect agreement ($\kappa = 0.81 - 1.00$), two had lower agreement scores. If these results had been obtained from specialists in different centers, the agreement discrepancies may have been attributable to differences in the style and quality of training. However, in this study, it may be due to varying experience with routine biochemistry, as the specialists worked in the same center but different laboratories.

In our study, when the ANN was used to retrospectively reassess results previously assessed by specialists during routine laboratory operation, our model approved nine (13.8%) results that had been rejected by the specialists and rejected 145 (3.9%) results that had been approved by the specialists. The test results with these contradictory decisions were evaluated again by the specialists. In assessments made without time constraints, the specialists changed their reporting decisions, thus increasing the agreement between their decisions and the model's verification decisions (false-negative rate was 0.5%, false-positive rate was 0.0%).

In a study including 11 hospitals, a test panel called the "basal metabolic panel" was created and included the tests we selected for our study. The test panel's turnaround time was measured as approximately 60 minutes. It was reported that an average of 55,000 samples per year come to the emergency services of these centers.[29] If a biochemistry specialist spends an average of 1 minute evaluating each sample result, an average of 916.7 hours would be spent annually on evaluation alone. If one considers that emergency services provide 24-hour uninterrupted service, a biochemical specialist processing samples from emergency services spends 38.2 (8.8%) of 365 days just evaluating results. Using our system in such a laboratory could lighten the workload by 767.6 hours (31.9 days) annually.

Our proposed model has some limitations. The lack of integration of analytic calibration data as well as internal quality control results into our system is an important problem. However, to avoid this problem, the specialists were asked only to evaluate patient results that had already passed technical verification. In the future, the Boolean and ANN models may be combined into a hybrid laboratory information management system to monitor the quality control process. On the other hand, the autoverification of forensic tests such as blood alcohol may cause legal problems.[16] Therefore, a separate manual confirmation process may be necessary when patient results requiring judicial follow-up are verified in our system.

On the other hand, in our study, using a consensus of specialists' decisions for evaluating the test results and consequently using these results as a training set for the creation of an ANN model might cause a classification bias in our study. However, similar studies in the literature have accepted multiple experts' joint decisions on test results as the gold standard, as in our study.[2,16,18]

In 2010, a consensus was reached regarding the reporting of hematologic test results, and all guidelines were combined under one heading.[30] Similarly, in the field of clinical biochemistry, some sort of consensus could be reached wherein the most effective methods for result reporting are chosen. Of these methods, the strongest candidate is likely a method based on ANN. With the learning ability that ANN possesses, it can be used as a strong and flexible tool for making correct decisions.

In conclusion, our study may form a basis for later studies that investigate the establishment of patient safety-oriented, new-generation laboratory evaluation systems for reporting medical test results.

Corresponding author: Ferhat Demirci, MD, Clinical Biochemistry Laboratory, Dr Suat Seren Chest Disease and Thoracic Surgery Training and Research Hospital, Yenisehir, 35110 Izmir, Turkey; drferhat5505@hotmail.com.

We thank Canan Coker, MD, Sezer Uysal, MD, Murat Ormen, and Ozlem Gursoy Calan, MD, for the academic support in test reporting. Also, we acknowledge Aylin Demirci, MD, and Oktay Yildirim, MSc, for the critical reading of the manuscript and technical contributions.

References

1. Dighe AS, Rao A, Coakley AB, et al. Analysis of laboratory critical value reporting at a large academic medical center. *Am J Clin Pathol*. 2006;125:758-764. doi:10.1309/R53XVC2U5CH6TNG8.
2. Oosterhuis WP, Ulenkate HJ, Goldschmidt HM. Evaluation of LabRespond, a new automated validation system for clinical laboratory test results. *Clin Chem*. 2000;46:1811-1817.
3. Neeley W. *Autoverification of Clinical Laboratory Test Results: Approved Guideline*. Wayne, PA: Clinical and Laboratory Standards Institute; 2006.
4. Niland JC, Stiller T, Neat J, et al. Improving patient safety via automated laboratory-based adverse event grading. *J Am Med Inform Assoc*. 2012;19:111-115. doi:10.1136/amiajnl-2011-000513.
5. Smith N, Rosenfeld D, Watman R. Hematology autovalidation system. *Lab Hematol*. 1999;5:52-55.
6. Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. 2001;7:673-679. doi:10.1038/89044.

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

DOI: 10.1093/ajcp/aqw104
© American Society for Clinical Pathology

7. Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo, CA: Elsevier; 2014.
8. Dimeski G. Interference testing. *Clin Biochem Rev*. 2008;29:S43.
9. Wu AH. *Tietz Clinical Guide to Laboratory Tests*. St Louis, MO: Saunders/Elsevier; 2006.
10. Wagar EA, Friedberg RC, Souers R, et al. Critical values comparison: a College of American Pathologists Q-Probes survey of 163 clinical laboratories. *Arch Pathol Lab Med*. 2007;131:1769-1775. doi:10.1043/1543-2165(2007)131[1769:CVCACO]2.0.CO;2.
11. Guder WG, da Fonseca-Wollheim F, Heil W, et al. The haemolytic, icteric and lipemic sample recommendations regarding their recognition and prevention of clinically relevant interferences: recommendations of the working group on preanalytical variables of the German Society for Clinical Chemistry and the German Society for Laboratory Medicine. *Lab Med*. 2000;24:357-364. doi:10.1515/labm.2000.24.8.357.
12. Vermeer HJ, Thomassen E, de Jonge N. Automated processing of serum indices used for interference detection by the laboratory information system. *Clin Chem*. 2005;51:244-247. doi:10.1373/clinchem.2004.036301.
13. Fraser CG, Stevenson HP, Kennedy IMG. Biological variation data are necessary prerequisites for objective autoverification of clinical laboratory data. *Accredit Qual Assur*. 2002;7:455-460. doi:10.1007/s00769-002-0526-3.
14. Lobo DN, Stanga Z, Simpson JA, et al. Dilution and redistribution effects of rapid 2-litre infusions of 0.9% (w/v) saline and 5% (w/v) dextrose on haematological parameters and serum biochemistry in normal subjects: a double-blind crossover study. *Clin Sci*. 2001;101:173-179.
15. Wang S. Autoverification of clinical chemistry results. *Clevel Clin*. 2011. http://www.aacc.org/resourcecenters/archivedprograms/expert_access/2011/november/pages/default.aspx. Accessed March 15, 2015.
16. Shih M-C, Chang H-M, Tien N, et al. Building and validating an autoverification system in the clinical chemistry laboratory. *Lab Med*. 2011;42:668-673. doi:10.1309/LM5AM4IIXC4OIETD.
17. Jones JB. A strategic informatics approach to autoverification. *Clin Lab Med*. 2013;33:161-181. doi:10.1016/j.cll.2012.11.004.
18. Krasowski MD, Davis SR, Drees D, et al. Autoverification in a core clinical chemistry laboratory at an academic medical center. *J Pathol Inform*. 2014;5:13. doi:10.4103/2153-3539.129450.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174. doi:10.2307/2529310.
20. Wikipedia. Backpropagation. 2015. http://en.wikipedia.org/wiki/Backpropagation. Accessed January 25, 2015.
21. Bruin J. How do I interpret odds ratios in logistic regression? *UCLA Stat Consult Gr*. 2011. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/citingats.htm. Accessed January 30, 2016.
22. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*. 2002;35:352-359. doi:10.1016/S1532-0464(03)00034-0.
23. Lisboa PJG. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw*. 2002;15:11-39. doi:10.1016/S0893-6080(01)00111-3.
24. Mitchell TM. *Artificial neural networks. In: Machine Learning*. New York, NY: McGraw-Hill; 1997:81-126. doi:10.1145/242224.242229.
25. Valdiguie PM, Rogari E, Philippe H. VALAB: expert system for validation of biochemical data. *Clin Chem*. 1992;38:83-87.
26. Valdiguié PM, Rogari E, Corberand JX, et al. The performance of the knowledge-based system VALAB revisited: an evaluation after five years. *Eur J Clin Chem Clin Biochem*. 1996;34:371-376.
27. Dorizzi RM, Caruso B, Meneghelli S, et al. The DNSev™ expert system in the auto-verification of tumour markers and hormones results. *Accredit Qual Assur*. 2006;11:303-307. doi:10.1007/s00769-006-0160-6.
28. Torke N, Boral L, Nguyen T, et al. Process improvement and operational efficiency through test result autoverification. *Clin Chem*. 2005;51:2406-2408. doi:10.1373/clinchem.2005.054395.
29. Holland LL, Smith LL, Blick KE. Reduce emergency department turnaround time outliers can reduce emergency department patient length of stay. *Am J Clin Pathol*. 2005;124:672-674. doi:10.1309/E9QPVQ6G2FBVMJ3B.
30. International Society for Laboratory Hematology consensus guidelines: rules. 2009. http://www.islh.org/2010/index.php?page=consensus_rules. Accessed November 20, 2015.
31. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323:533-536. doi:10.1038/323533a0.
32. MacKay, DJC. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press; 2003.

Downloaded from https://academic.oup.com/ajcp/article/146/2/227/1731226 by Dokuz Eylul University Library (DEU) user on 02 September 2021

© American Society for Clinical Pathology

DOI: 10.1093/ajcp/aqw104