Clinical Chemistry 46:11
1811–1817 (2000)

Laboratory Management

Evaluation of LabRespond, a New Automated Validation System for Clinical Laboratory Test Results

Wytze P. Oosterhuis,[1*] Herman J.L.M. Ulenkate,[2] and Henk M.J. Goldschmidt[1]

Background: **Manual validation of laboratory test results is time-consuming, creating a demand for expert systems to automate this process. We have started to set up the program "LabRespond", which covers five validation levels: administrative, technical, sample, patient, and clinical validation. We present the evaluation of a prototype of an automated patient validation system based on statistical methods, in contrast to the commercially available program "VALAB", a rule-based automated validation system.**

Methods: **In the present study, 163 willfully altered, erroneous test results out of 5421 were submitted for validation to LabRespond, VALAB, and to a group of clinical chemists (n = 9) who validated these test results manually. The test results rejected by three or more clinical chemists (n = 281) served as a secondary reference standard.**

Results: **The error recovery rates of clinical chemists ranged from 23.9% to 71.2%. The recovery rates of LabRespond and VALAB were 77.9% and 71.8%, respectively (difference not significant). The false-positive rates were 82.7% for LabRespond, 83.6% for VALAB, and 27.8–86.7% for clinical chemists. Using the consensus of three or more clinical chemists as the secondary reference standard, we found error recovery rates of 64.8% for LabRespond and 72.2% for VALAB ($P = 0.06$). Compared with VALAB, LabRespond detected more ($P = 0.003$) erroneous test results of the type that were changed from abnormal to normal.**

Conclusions: **The statistical plausibility check used by LabRespond offers a promising automated validation method with a higher error recovery rate than the clinical chemists participating in this study, and a performance comparable to VALAB.**

Despite increases in automation in clinical laboratories, human intervention is still required for some phases of the analytical and postanalytical process, including validation of test results. The aim of validation is to prevent reporting of erroneous test results to clinicians. The frequency of errors in laboratories varies in the literature, e.g., 0.47%, 1:1000, or 1:100, at least in part because of differences in the categorization of the errors *(1, 2)*. In an automated environment, most of the errors are of human origin *(1)*. The validation of a large number of laboratory test results often is performed by visual inspection. In some laboratories, this inspection is limited to abnormal or extremely abnormal results. In this nonautomated validation process, test results are screened in the context of other test results and the patient information available. Results that have a low plausibility in the eyes of the validator will generate a subsequent action, such as retesting or a call to the clinician.

Manual test validation is a time-consuming process with large interindividual variation. In addition, it slows down the response of the laboratory to the clinic. For that reason, critical care units often receive results that are only partly validated. It has been shown that it is possible to automate this validation process, e.g., by use of the commercially available program VALAB (Validation Assistée pour LABoratoire) *(3–7)*. VALAB is a rule-based expert system, installed on a personal computer, that reviews clinical laboratory reports coming from the laboratory information system, taking the biological quantities for which it has been designed into account. The Dutch Working Group of Clinical Chemometrics has developed a prototype of the validation program "LabRespond" *(8)*, which applies several validation steps. LabRespond is a total quality management approach on five validation levels: administrative, technical, sample, patient, and clin-

[1] Department of Clinical Chemistry and Hematology, St. Elisabeth Hospital, Hilvarenbeekseweg 60, 5022 GC Tilburg, The Netherlands.
[2] Diagnostic Center SSDZ, Department of Clinical Chemistry, PO Box 5011, 2600 GA Delft, The Netherlands.
*Address correspondence to this author at: Department of Clinical Chemistry and Hematology, St. Elisabeth Hospital, PO Box 90.151, 5000 LC Tilburg, The Netherlands. Fax 31-13-5352390; e-mail oosthuis@knmg.nl.
Received May 9, 2000; accepted August 22, 2000.

ical validation. The checks at each validation level are described in the *Appendix*. In contrast to VALAB, LabRespond is not a rule-based expert system, and only a minimum of validation rules had to be implemented. The fourth level, patient validation, is based on a statistical calculation of the plausibility of each individual test result using historical data of test results. The statistical plausibility check is outlined, allowing a multivariate context check in two dimensions: time and other test results. In the present study, we evaluated our new validation method by comparing LabRespond with visual inspection of individual clinical chemists and with VALAB.

Materials and Methods

LabRespond: description, design, and configurations

Calculation of test plausibility based on historical data. One of the five levels, the patient level, is evaluated here and will be described in brief (see the *Appendix* for a more detailed description with examples).

Determination of observed frequencies. All test results of 33 different laboratory tests for a 3-month period (Table 1) were downloaded from the laboratory automation system. All of these results were reported to the physicians and were validated in the usual way (manually). The test results were divided into seven classes that were defined separately for each test. The observed frequencies were expressed as fractions. The limits of each class were set at fixed percentiles. The test result plausibility of each class was derived from the actual frequencies. Using the frequencies of occurrence of combinations of two test results, we constructed 7 × 7 frequency matrices. Only test combinations showing a statistically significant correlation ($P < 0.05$) were included. These were combinations of two different tests (e.g., hemoglobin and mean corpuscular volume) as well as combinations of a test result with its previous historical result (delta check).

Calculation of expected frequencies. For all combinations of two test results (i.e., for each cell of the 7 × 7 matrices), the expected frequency of occurrence was calculated by multiplying the observed fractions of both test results in the corresponding classes with the total number of test results. Subsequently, the ratio of the observed frequency and the expected frequency was calculated of each cell of the 7 × 7 matrices and used as the parameter in the calculation of the test plausibility. The expected frequency was based on the assumption of independence of both results. A ratio of 1 means that the frequency of occurrence of the combination of test results is just as probable as expected. A ratio >1 means the combination is more probable, and <1 means the combination is less probable than expected. Zero means that the combination of test results was not found within the historical data. All extreme values [outside percentiles <1% (class 1) and >99% (class 7)] were always rejected and reported as "out of range".

Delta check. For each test, a time window was defined, ranging from 7 days to 3 months (see Table A-2 in the *Appendix*). The time windows were a compromise: if a window is set too wide, the delta check can be performed in more cases, allowing a more sensitive validation process in those cases where a historical test result is available. On the other hand, information is lost by widening the time window because the delta check loses sensitivity because the correlation between the actual and the historical test result weakens with a longer time interval. Test results outside this window were excluded from the calculation of test plausibilities. The calculations of test plausibilities were performed separately on men and women to compensate for gender-dependent differences in test results.

Sensitivity. The sensitivity of the validation process can be predefined by changing the threshold plausibility for validation. We calibrated the validation rate of LabRespond to be equal to that of VALAB for optimal comparison. This means that of all test results evaluated by both systems, 86.6% were accepted and 13.4% were rejected. This validation rate was a result of the routine performance of VALAB, which was not altered for this study.

Table 1. Tests included in this study.

Electrolytes	Potassium, sodium, calcium, phosphate
Metabolites	Glucose, creatinine, urea, uric acid, bilirubin, direct bilirubin, cholesterol, HDL-cholesterol, triglycerides, albumin
Enzymes	Alkaline phosphatase, creatine kinase, aspartate aminotransferase, alanine aminotransferase, γ-glutamyltransferase, lactate dehydrogenase, amylase
Proteins	Total protein, albumin, C-reactive protein
Hormones	Free T_4,[a] human chorionic gonadotropin
Hematology	Leukocyte count, hemoglobin, mean cell volume, erythrocyte count, hematocrit, mean cell hemoglobin, mean cell hemoglobin concentration, red cell distribution width, platelet count
Miscellaneous	HbA1c, vitamin B_{12}, ferritin

[a] T_4, thyroxine; Hb, hemoglobin.

Experimental design

In the present study, we collected from 2 consecutive days the hematology and clinical chemistry laboratory test results (n = 5421) of 33 different laboratory tests (Table 1). These 33 tests covered the larger part of ~70% of the routine laboratory production in our hospital. These test results (n = 5421) were already validated on the usual manner by visual inspection, and all results had been reported to the clinic (615 reports). We therefore considered these test results as a validated set. All patient laboratory data were retrieved from the Department of Clinical Chemistry, De Heel Hospital, Zaandam, The Netherlands. In this validated set, an independent clinical chemist introduced 163 minor or major errors at a maxi-

mum rate of 1 error per report. The results were changed without knowledge of how the output of LabRespond (the calculated plausibility of the erroneous result) would be influenced. Different types of errors were introduced to test the sensitivities of the different validation methods in various ways. The types of errors introduced were chosen in such a way that the most serious errors were introduced in the highest frequency. Most errors were made from normal to abnormal (n = 114). Together with the errors of the type from abnormal to normal (n = 33), these were considered the most serious. The remaining errors were made from normal to normal or abnormal to abnormal (n = 16). The introduced errors were to be recovered by clinical chemists and both expert validation systems, LabRespond and VALAB. Both systems can operate on a personal computer, with the software running under MS-DOS.

The records entered in both expert systems contained the following data: gender of the patient, date of the report, test results and when available, and the previous result of each test with the date. The 615 cumulative reports were examined by nine independent clinical chemists from five different countries (five from The Netherlands, one from Italy, one from the US, one from the United Kingdom, and one from Germany) who were very familiar with manual validation. The clinical chemists were instructed to mark all results they regarded doubtful. The cumulative reports, including those with introduced errors, were printed in the usual layout, except for the following adaptations: the reports did not contain information of the patient's name, the clinician's name, the department, the type of patient (in- or outpatient), complementary data about the therapy or medication, and no results of tests outside those presented in Table 1. The reports were printed with the patient's birth date and gender; requested results of the tests presented in Table 1, including the date; and previous test results when available, including the date, reference limits, and remarks on sample integrity. For every error that was introduced, the recovery by each of the clinical chemists and each expert system was scored. Test results that were rejected by three or more clinical chemists (n = 9) were considered as the consensus of the clinical chemist group and served as a secondary reference standard (some test results without introduced errors did fall into this category). This was analogous to an evaluation study of VALAB, and the number of three was taken for optimal comparison of results *(6)*.

The error recovery rate of the clinical chemist group was compared with the recovery rates of both LabRespond and VALAB. The error recovery performance is expressed in sensitivity and specificity, and presented in a ROC curve. In addition, we compared LabRespond with VALAB for the detection of the type of introduced errors.

STATISTICS

We assumed that the sensitivity of VALAB to detect the introduced errors was 75% *(4)*. The null hypothesis was that there was no difference in error recovery rate between two different validation methods, LabRespond and VALAB. Rejection of the null hypothesis was required to conclude that the methods were different. Using an α level of 0.05 (two-sided) and a power of 0.80, we calculated that at least 155 errors should be introduced to be able to detect a difference of 10% *(9)*. For testing on significance of differences between proportions, we used the χ^2 test. $P \leq 0.05$ was considered statistically significant. All calculations were performed with SPSS software (SPSS Inc.).

Results

The results obtained from examination of the laboratory reports by clinical chemists and by both expert systems, LabRespond and VALAB, are presented in Table 2. The total percentage of validated test results by VALAB was 86.6%, and LabRespond was calibrated to the same validation rate for optimal comparison. All clinical chemists had higher validation rates (89.8–98.4%), leading to higher specificities (91.7–99.5%) compared with both expert systems (88.6% for LabRespond, and 88.4% for VALAB). The error recovery rates of LabRespond (77.9%; Fig. 1, ●) and VALAB (71.8%; Fig. 1, ■) were not statisti-

Table 2. Results of the error recovery study of clinical chemists and both expert systems (percentages relative to the number of introduced errors).

		Clinical chemist[a]									Expert system	
		1	2	3	4	5	6	7	8	9	VALAB	LabRespond
Validated, %	TN + FN[b]	89.8	92.6	92.8	96.9	97.0	97.2	98.2	98.4	98.4	86.6	86.6
Rejected, %	TP + FP	10.2	7.4	7.2	3.1	3.0	2.8	1.8	1.6	1.6	13.4	13.4
Correct rejected, %	TP	2.1	1.0	2.0	1.6	1.2	1.0	1.3	0.8	0.7	2.2	2.3
Incorrect validated, %	FN	0.9	2.0	1.0	1.4	1.8	2.0	1.7	2.2	2.3	0.8	0.7
Incorrect rejected, %	FP	8.1	6.5	5.2	1.5	1.8	1.8	0.5	0.8	0.8	11.2	11.0
Correct validated, %	TN	88.9	90.5	91.8	95.5	95.2	95.2	96.5	96.2	96.1	85.8	85.9
Sensitivity (error recovery), %		71.2	31.9	68.1	52.8	39.9	33.7	43.6	28.2	23.9	71.8	77.9
Specificity, %		91.7	93.3	94.7	98.4	98.2	98.2	99.5	99.2	99.1	88.4	88.6

[a] Clinical chemist; data are numbered 1–9.
[b] TN, true negative; FN, false negative; TP, true positive; FP, false positive.

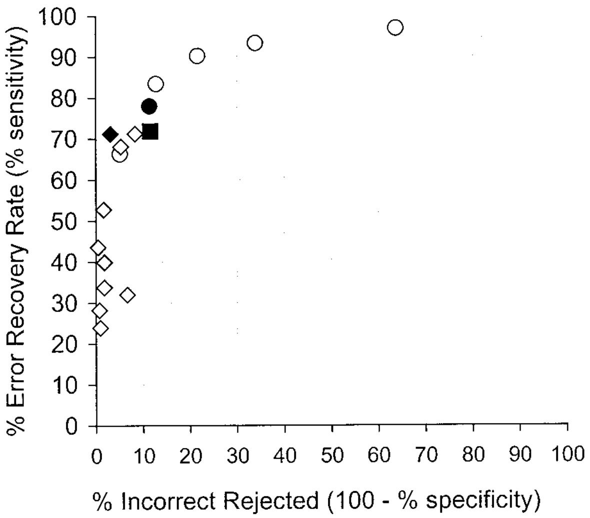

Fig. 1. ROC curve representing LabRespond at different threshold plausibility levels for validation (○); LabRespond at the same validation rate as VALAB (●); VALAB (■); individual clinical chemists (◇); and secondary reference standard (consensus of the clinical chemists, e.g., results rejected by three or more of nine clinical chemists; ◆).

cally significantly different (P = 0.20). The false-positive rates [defined as FP/(FP + TP), where FP is the number of false positives and TP is the number of true positives] were 82.7% (LabRespond) and 83.6% (VALAB), respectively. The false-positive rates of the clinical chemists were 27.8–86.7%. The error recovery rates of eight of nine clinical chemists (23.9–68.1%) were significantly lower than that of LabRespond. VALAB had a significantly higher error recovery rate than seven of nine clinical chemists (23.9–52.8%). We compared the validation of two types of introduced errors: (*a*) test results that were changed from normal to abnormal (n = 114) and (*b*) test results changed from abnormal to normal (n = 33). The rest of the errors (n = 16) were changes within the normal or abnormal range. The error recovery rates of LabRespond and VALAB for the first type of error were 79.8% and 82.5%, respectively (P = 0.61). The error recovery rates of LabRespond and VALAB for the second type of error were 75.8% and 39.4%, respectively (P = 0.003). The results of the error recovery of the individual clinical chemists are presented in a ROC curve (Fig. 1, ◇). Test results (n = 281) rejected by three or more clinical chemists out of nine (Fig. 1, ◆) served as a secondary consensus-based reference standard to evaluate both expert systems. The error recovery rate for these test results was 64.8% by LabRespond and 72.2% by VALAB. This difference was not statistically significant (P = 0.06).

Discussion

The manual validation of laboratory test results represents a large effort, and the automation of at least part of this process will decrease the interobserver variability in the validation process, the number of reported laboratory errors, and the time spent on validation, and thus increase the standardization and total quality. These were the main incentives for the Working Group of Clinical Chemometrics of the Dutch Association of Clinical Chemistry to set up the LabRespond project and develop a new expert system for automated test validation *(8)*. Another reason to construct an alternative validation program was that the processing algorithms and decision rules of the commercial solution were considered proprietary and therefore were not available to laboratory users *(3–7)*. In our opinion, the exact reason that a test result is validated or rejected should be clear to and adaptable by the user. VALAB submits the test results and other data contained in each laboratory report to an algorithmic reasoning based on biological, demographic, and clinical considerations. For this purpose, it uses >20 000 rules that have been preset by the manufacturer *(3)*. These rules, however, are not available to the user. Updating a rule-based system is much more difficult than updating the statistical discriminators used by LabRespond.

In this study, we evaluated the new validation method of level 4 of LabRespond. This program uses bivariate frequency distributions. These distributions are derived from historical data obtained from the laboratory itself that applies LabRespond. The LabRespond validation process is completely transparent to the user and can be adapted if wanted.

Bidimensional matrices are also used in the program Plamix *(10)*. In contrast to LabRespond, where the plausibility of a test result is calculated quantitatively, Plamix discriminates only the two classes "plausible" and "not plausible" of combinations of two test results.

We compared validation by visual inspection by nine clinical chemists with automated validation, using the programs LabRespond *(11)* and VALAB. The present study shows that the validation expert system LabRespond is significantly more sensitive (77.9%) in detecting introduced errors when compared with the error recovery rates scored by each clinical chemist, except one (sensitivity ranged from 23.9% to 71.2%). All clinical chemists had fewer incorrect rejected (false positive) test results. The clinical chemists seemed to be willing to sacrifice sensitivity to decrease the number of false alarms.

The error recovery rate of LabRespond (77.9%) was higher than that of VALAB (71.8%), although this difference was not statistically significant. We chose to introduce errors, so the reference standard for the error recovery is truly a "gold standard". As a secondary reference standard to evaluate our statistical validation methods, we compared both programs with a consensus of three or more clinical chemists (n = 9). This was based on a study of Fuentes-Arderiu et al. *(6)*. The secondary reference standard, consisting of the consensus of a number of clinical chemists, is more susceptible to bias because of the subjectivity of the experts. The manual validation process is subject to a large interobserver variation, as shown by the results presented in Table 2 and Fig. 1.

The sensitivity of the validation process can be predefined within LabRespond by changing the threshold plausibility for validation. For the purpose of this study, the validation rate of LabRespond was calibrated to be equal to that of VALAB for optimal comparison. However, the operation point of LabRespond can be easily shifted to a lower sensitivity: a 10% decrease in sensitivity, from 78% to 68%, corresponds to a drop in the false-positive rate of more than one-half.

The validation study carried out here only approaches reality, for several reasons. The number of introduced errors is rather high *(1, 2)*. On 615 reports, 163 erroneous test results were introduced. Of all reports, 26.5% contained an error, and 3% of all test results were erroneous. Some introduced erroneous test results would be complete blunders in reality, others were small errors or even appeared more plausible than the original, strongly deviating test results. In our study, the clinical chemists probably spent more time on validation of the test results than they would in daily practice. Therefore, we expect that the true error recovery rate of the clinical chemists will be lower under routine circumstances.

LabRespond appeared to be significantly more sensitive in detecting erroneous test results that fall within the reference limits than VALAB ($P = 0.003$). LabRespond performs a plausibility check on all test results, irrespective of whether the test result is normal or abnormal. VALAB probably contains more decision rules on abnormal results than on normal results, so normal values are more easily validated.

In the evaluation study of Fuentes-Arderiu *(6)*, the validation results of VALAB for 25 tests were compared with a consensus standard of three or more clinical chemists out of nine, who examined a total of 500 reports. VALAB did not validate any report that was rejected by three or more clinical chemists. This result corresponds to a sensitivity of 100%. In another study, the sensitivity of VALAB to detect reports rejected by consensus of laboratory specialists was evaluated *(3)*. The mean sensitivity was 98.1% for clinical chemistry data and 83.7% for hematological cytology. The results of our study cannot be compared directly because we studied the sensitivity on a test level, not on a report level. The sensitivity for VALAB of 72.2% obtained in our study is definitely lower than the results of previously reported studies *(3, 4, 6, 7)*.

COMPARISON OF THE METHODS OF LabRespond AND VALAB

Both LabRespond and VALAB basically use the same validation items: extreme limits, anteriority (delta check), and correlation between different test results. In the present study, LabRespond took only the relationship between test results and the gender of the patient into account. VALAB can include other data in the validation process: department, age, urgency of the request, ambulatory or hospitalized patient, and medical specialty of the petitioner. In this study, these last data were not included. This setting corresponded to the routine setting used in the Dutch hospital that tested the validation of VALAB for the present study. The use of additional data will allow VALAB to reject fewer test results, improving specificity but not sensitivity. On the other hand, LabRespond is open to improvement by making separate correlation matrices for patient subgroups, such as in- and outpatients and pediatric patients. Further efforts will be directed to making an updated version of the program integrated within the laboratory information system. The number of tests that can be handled by the program can easily be expanded by the user.

In conclusion, we have developed the prototype validation system LabRespond. The statistical patient validation procedure appeared to have a higher error recovery rate than all but one of nine participating clinical chemists. The performance of LabRespond was comparable to that of a similar, established system named VALAB. The statistical validation procedure of LabRespond offers a promising method for the automated validation of clinical laboratory test results, a future trend that is essential for an effective consultation from the laboratory.

Members of the Dutch Working Group of Clinical Chemometrics of the NVKC (Dutch Association of Clinical Chemistry): Drs. Marcel Volmer, Martin van der Horst, Nada O. Osmanovic; Dr. Ir. Remi W. Wulkan; Drs. Rob. N.M. Weijers, Jan Dols, Kees A.J.M. van Dongen, Winfriet Gengler, and Tjitse Dijkstra. We would like to acknowledge Dr. Mario Plebani (Italy); Dr. Heinz Juergen Roth (Germany); Prof. Dr. Richard W. Lent (US); Dr. Richard Jones (UK), and Dr. Hans J.M.L. Hoffmann, Dr. Jeroen D.E. van Suijlen, Dr. Hein H. Kamp, Ir. John F. van de Calseijde, and Dr. Joost C.J.M Swaanenburg (all from The Netherlands) for manual validation of all test results, and Yvonne Bandt and Anton Huisman (The Netherlands) for data acquisition. This work was supported financially by the St. Elisabeth Hospital Quality Committee and the Dutch Association of Clinical Chemistry.

References

1. Plebani M, Carraro P. Mistakes in a stat laboratory: types and frequency. Clin Chem 1997;43:1348–51.
2. Witte DL, VanNess SA, Angstadt DS, Pennell BJ. Errors, mistakes, blunders, outliers, or unacceptable results: how many? Clin Chem 1997;43:1352–6.
3. Valdiguié PM, Rogari E, Corberand JX, Boneu B. The performance of the knowledge-based system VALAB revisited: an evaluation after five years. Eur J Clin Chem Clin Biochem 1996;34:371–6.
4. Valdiguié PM, Rogari E, Philippe H. VALAB: expert system for validation of biochemical data. Clin Chem 1992;38:83–7.
5. Corberand JX, Rogari E, Laharrague P, Fillola G, Valdiguié PM. Computer-assisted validation system applied to hematology: Valab-haemato. Ann Biol Clin (Paris) 1994;52:447–50.
6. Fuentes-Arderiu X, Castiñeiras-Lacambra MJ, Panadero-García MT. Evaluation of the VALAB expert system. Eur J Clin Chem Clin Biochem 1997;35:711–4.

7. Marchand M, Guibourdenche J, Saada J, Le Men H, Porquet D, Demelier J-F. Real-time expert validation of paediatric clinical biochemistry reports using the Valab-Biochemistry system. Ann Clin Biochem 1997;34:389–95.
8. Ulenkate HJLM, Oosterhuis WP, Goldschmidt HMJ. Development of 'RESPOND': an automated validation system for clinical laboratory test results [Abstract]. Ned Tijdschr Klin Chem 1997;22:153–4.
9. Ma[illegible]ch R, S[illegible] R. Sample [illegible]. Cancer Treat Rep 1978;62:1037–40.
10. Po[illegible]h AJ. Pla[illegible]. [illegible] Verfahren [illegible]. Mitt Dtsch Ges Klin Chem 1995;26:91–5.
11. Ulenkate HJLM, Oosterhuis WP, O[illegible] N, Goldschmidt HMJ. Self reporting by validation software (VALAB and LabRespond) compared with clinical chemists [Abstract]. Clin Chem Lab Med 1999;37:S250.
12. Oberholzer M, Ostreicher M, Christen H, Brühlmann M. Methods in quantitative image analysis. Histochem Cell Biol 1996;105:333–55.

Appendix

In its final version, LabRespond will contain five validation levels. In the current Total Quality Management approach, all five levels are essential, but level 4 is the most innovative. In the present study, only level 4, the validation of the test results using a statistical plausibility check, was evaluated. Examples of the checks in the other validation levels are shown in Table A-1.

CALCULATION OF TEST RESULT PLAUSIBILITY

In level 4, the plausibility of a single test result is evaluated by combining pairs of test results. The combinations of test results in Table A-2 are included in the calculation of the plausibility of a test result. For example, the probability check of glucose includes the following test results (when available): previous glucose test result (test result 1) if not older than 7 days, potassium (test result 2), and sodium (test result 3). Other tests and combinations can be applied easily.

Pretest plausibility. The plausibility of a test result, before taking other data except patient gender into consideration, is considered as the "pretest" plausibility. This was set equal to the percentage of the test results within the class of that test result. These frequencies were by definition close to the following percentages: 5%, 10%, 20%, 30%, 20%, 10%, and 5%, respectively (small deviations were attributable to the rounding of the numbers). Patients who had a certain test performed twice within the time window presented in Table A-2 formed a subgroup. Within the same class limits, the pretest plausibilities were different from the percentages above: whereas 5% of all the test results of creatinine were >219 μmol/L in males, 10.31% were above this limit in male patients who had creatinine retested within 30 days. Obviously, the patients who were tested repeatedly were a group with more abnormal test results.

Posttest plausibility. The plausibility of a test result after the frequency of occurrence of the combination of test results is taken into consideration.

Table A-1. Validation levels in LabRespond.

Level	Examples of checks
1. Administrative	Patient identification number, date of birth, gender, sample identification number, sample date
2. Technical	Calibration, quality-control data
3. Sample	Test in number, sample type, sample volume
4. Patient	Plausibility of test results, extreme values
5. Clinical	Reflex testing, LDL-calculation check, protocols

Table A-2. Test combinations used in the current plausibility check.

Number	Test name	Time window, days	Test combinations
1	Glucose	7	1; 2; 3
2	Potassium	7	2; 1; 4; 5; 17
3	Sodium	7	3; 1; 5; 7
4	Creatinine	30	4; 2; 5; 17
5	Urea	14	5; 2; 3; 4; 17
6	Total protein	30	6; 7; 16
7	Albumin	14	7; 3; 6; 12; 16
8	Alkaline phosphatase	14	8; 10; 11; 12; 14; 15; 22
9	Creatine kinase	14	9; 10; 15
10	Aspartate aminotransferase	14	10; 8; 11; 14; 15
11	Alanine aminotransferase	14	11; 8; 10; 12; 14
12	Total bilirubin	14	12; 7; 8; 11; 14
13	Direct bilirubin	14	13; 12
14	γ-Glutamyltransferase	14	14; 8; 10; 11; 12; 15; 22
15	Lactate dehydrogenase	14	15; 8; 10; 14
16	Calcium	30	16; 6; 7; 17
17	Phosphate	30	17; 2; 4; 5; 16
18	Cholesterol	60	18; 19; 20
19	HDL-cholesterol	60	19; 18; 20
20	Triglyceride	60	20; 18; 19
21	Urate	30	21
22	Amylase	14	22; 8; 14
23	C-Reactive protein	14	23; 24; 28
24	White blood cell count	7	24; 23; 25; 27
25	Hemoglobin	14	25; 24; 26; 27; 33
26	Mean corpuscular volume	90	26; 25; 33
27	Thrombocyte count	14	27; 24; 25
28	Erythrocyte sedimentation rate	7	28; 24; 23
29	HbA1c[a]	90	29; 1
30	Free T_4	90	30; 31
31	Thyrotropin	90	31; 30
32	Vitamin B_{12}	90	32; 25; 26
33	Ferritin	90	33; 25; 26

[a] Hb, hemoglobin; T_4, thyroxine.

FREQUENCY MATRIX

The frequency matrices were obtained by dividing the test results into seven classes. The classes were defined by the following percentiles: <5%, 15%, 35%, 65%, 85%, >95%. These percentiles were calculated (for both the male and the female populations) based on all unselected test results. In some tests, it was not possible to divide the test results according to these percentiles, and the classes had to be adapted (e.g., for erythrocyte sedimentation rates, well above 17.91% of the test results have a result <1 mm/h).

Extreme values. Results outside the percentiles 1% and 99% were considered extreme values and were never validated automatically.

Smoothing. The matrix with the frequencies of test results is first smoothed. Combinations of low probability will have a low frequency (e.g., "creatinine high" with "urea low"). These low frequencies are particularly subject to random variation. To overcome this problem, the following approach is used to smooth the results. The central matrix element is replaced by an average of this central element with the surrounding elements using the following weights ["neighborhood averaging" *(12)*]:

$$\begin{matrix} 1 & 2 & 1 \\ 2 & 8 & 2 \\ 1 & 2 & 1 \end{matrix}$$

Calculation of the expected frequency. The expected frequency is calculated by multiplying the actual frequencies of the two classes. For example, for calculation of the expected frequency of the combination of creatinine and urea (males), the observed frequency of a creatinine >219 μmol/L is 4.98%, of urea >27.7 mmol/L is 4.93%. The expected frequency of the combination of both results is (0.0498 × 0.0493) × 100% = 0.246%. The observed (smoothed) frequency of this combination is much higher. Because the positive correlation between creatinine and urea, 1.537% of the test results lie within this cell of the matrix.

For each matrix element, the ratio of the observed (smoothed) frequency to the expected frequency is calculated. In our example of creatinine and urea, this ratio is 1.537%/0.246% = 6.25. The pretest probability of the test result is multiplied by each ratio of tests as defined in Table A-2.

EXAMPLES OF CALCULATION OF TEST PLAUSIBILITY

Example 1. Patient is a female with a serum sodium concentration of 144 mmol/L, a previous sodium (1 day before last test result) of 140 mmol/L, glucose of 9.3 mmol/L, and urea of 22.6 mmol/L. The pretest probability of a serum sodium of 144 mmol/L in a patient retested within 7 days (the time window for sodium; Table A-2) is 7.1%.

The ratios of observed vs expected frequencies of test results are: sodium and previous sodium, 0.8; sodium and glucose, 0.7; and sodium and urea, 0.8.

Posttest plausibility = 7.1% × 0.8 × 0.7 × 0.8 = 3.0% (threshold set at <5%; therefore, this result will not be validated automatically).

Example 2. Patient is a female with a hemoglobin concentration of 6.0 mmol/L, erythrocyte sedimentation rate of 7.0 mm/h, thrombocyte count of 414×10^{11}/L, and white blood cell count of 10.8×10^{9}/L. The pretest probability of hemoglobin of 6.0 mmol/L in a patient not retested within 14 days (the time window for hemoglobin; Table A-2) is 11.1%.

The ratios of observed vs expected frequencies of test results are: hemoglobin and erythrocyte sedimentation rate, 0.4; hemoglobin and thrombocyte count, 1.2; hemoglobin and white blood cell count, 1.0.

Posttest plausibility = 11.1% × 0.4 × 1.2 × 1.0 = 4.9% (threshold set at <5%; therefore, this result will not be validated automatically).

EXAMPLES OF ERRORS INTRODUCED IN TEST RESULTS

The errors introduced into the hemoglobin test results were [real value (introduced erroneous value), mmol/L]: 7.4 (6.2), 9.4 (6.3), 7.8 (6.9), 6.7 (9.8), 7.9 (9.9), 5.8 (10.0).

The errors introduced into the creatinine test results were [real value (introduced erroneous value), μmol/L]: 89 (110), 105 (136), 67 (145), 87 (120), 97 (48), 62 (140), 68 (20), 79 (300).

SOFTWARE

The software can be obtained through a written request to the first author.