Urine Test Strips

How reproducible are readings?

R.A.G. WINKENS, MD
P. LEFFERS, MSC
C.P. DEGENAAR, BSC

SUMMARY

In an experiment, multiple reagent test strips from 90 urine samples were examined twice: observed visually by one of two persons and analyzed by spectrophotometry. Interobserver and intra-observer agreement were calculated and expressed as Cohen's κ. Interobserver and intra-observer agreement were moderate to good, but lower than one might expect. Enhancing discoloration of the test pads could improve reproducibility.

RÉSUMÉ

Au cours d'une expérience comportant 90 échantillons d'urine, on a répété deux fois la lecture des bâtonnets à réactifs multiples: observation visuelle par l'une des deux personnes et analyse spectrophotométrique. Les accords inter et intra-observateurs ont été calculés et exprimés sous forme de coefficient de Cohen. Les accords inter et intra-observateurs ont varié de modérés à bons, mais furent inférieurs aux attentes. Une meilleure qualité de la décoloration sur les bâtonnets réactifs pourrait améliorer la reproductibilité.

Can Fam Physician 1992;38:1095-1099.

THE EXAMINATION OF A URINE sample is a common task in general practice. One of the most commonly used methods is "macroscopic" examination by means of various types of test strips, generally preferred because they are quick and simple to use.

An important prerequisite for a diagnostic test is good reproducibility[1,2] – that is, the test should give the same result when performed by different analysts using the same or different reading techniques (interobserver agreement) and when performed repeatedly by the same analyst (intra-observer agreement). Low reproducibility points to possibilities for improvement of the test procedure and thus for improvement of the diagnostic value of the test.

The reproducibility of urinalysis would be considered good if repeated testing led to the same results – assuming that in the meantime no real change had taken place in the urine sample. Real changes could result from aging of the sample (influenced by storage temperature, pH, and osmolality).[3,4] Test results could also be affected by alterations in the test strip itself (because of storage temperature or strips for which the expiry date is past) and by imperfectly homogenized urine samples.[5]

Reproducibility is influenced by subjectivity in the grading of the test result (for example, different perceptions of color) and by differences in the execution of the test.[6] In the laboratory, changes of the urine sample and of the reagent test strip can be prevented.[7-9] Observational errors are much more difficult to control. It has been suggested that experience, routine, and education are important.[10]

The literature contains limited information about the reproducibility of urine test strips, and particularly little about observational errors.[11] Therefore we chose to study the influence of observational errors on test reproducibility. To measure interobserver and intra-observer agreement, we used visual and spectrophotometric reading of multiple reagent test strips to examine selected urine samples.

Overall agreement as a measure of observer variation has the drawback that, even if the observers randomly assigned test results, some agreement could still occur by chance. The level of chance agreement expected depends on the prevalence of positive test results in the study population. We used Cohen's κ, which is a measure for reproducibility corrected for "agreement by chance"[12,13] (see sidebar).

A negative κ means that the agreement is less than that expected by chance. A κ value of 0 means that the agreement is equal to that expected by chance, and a

Dr Winkens *is a General Practitioner at the Diagnostic Centre Maastricht, The Netherlands.*
Mr Leffers *is an Epidemiologist with the Department of Epidemiology, State University of Limburg, Maastricht.*
Mr Degenaar *is a Chemist at the Department of Clinical Chemistry, University Hospital, Maastricht.*

CORRECTING FOR AGREEMENT BY CHANCE

Suppose two observers perform a test on N cases. The test can give outcomes with κ possibilities. Judging the cases leads to the following comparison with chances for (dis)agreement.

	OBSERVER A (i)					
OBSERVER (B) (j)	**1**	**2**	**.**	**.**	κ	**TOTAL**
1	P_{11}	P_{12}	.	.	$P_{1\kappa}$	$P_{1.}$
2	P_{21}	P_{22}	.	.	$P_{2\kappa}$	$P_{2.}$
.						
.						
κ	$P_{\kappa 1}$	$P_{\kappa 2}$	.	.	$P_{\kappa\kappa}$	$P_{\kappa .}$
TOTAL	$\mathbf{P_{.1}}$	$\mathbf{P_{.2}}$	**.**	**.**	$\mathbf{P_{.\kappa}}$	**1**

Observed agreement = $P_{11} + P_{22} + \ldots + P_{\kappa\kappa}$

Expected chance agreement = $P_{1.}\,P_{.1} + P_{2.}\,P_{.2} + \ldots + P_{\kappa .}\,P_{.\kappa}$

κ corrects for the agreement by chance in the following way:

$$\kappa = \frac{\text{observed agreement (\%)} - \text{expected chance agreement (\%)}}{100\% - \text{expected chance agreement}} \quad or \quad \kappa = \frac{P_o - P_e}{1 - P_e}$$

κ can vary from −1 to +1.

κ larger than 0 means that the agreement is greater than that expected by chance.

METHOD

For the purpose of the experiment, 90 samples were selected from urine samples collected through inpatient and outpatient clinics and delivered daily to the department of clinical chemistry at the university hospital. The samples were selected on the basis of a positive reaction for one or more of the following tests: leukocyte esterase activity, nitrite, blood, and protein. A seven-patch test strip (Nephur-7-RL, manufactured by Boehringer Mannheim, Almere, The Netherlands; Chemstrip-6 in Canada was used. This strip tests for leukocyte esterase activity, nitrite, pH, protein, glucose, ketone bodies, and blood. However, only data from the test pads for leukocyte esterase activity, nitrite, protein, and blood were recorded, because the test pad for acidity is considered to be of limited value for general practice purposes, and only a very small number of urine samples contained glucose or ketone bodies.

The test pads for leukocyte esterase activity, blood, and protein all had four different categories of test results (− up to +++). The test pad for nitrite had two (− or +). The test strips and the spectrophotometric analyzer were used according to the recommendations of the manufacturer.

All selected urine samples were examined in two series by three "observers." Each sample was examined within 1 minute. The three observers were an inexperienced practical nurse trainee (observer 1), an experienced practical nurse (observer 2), and a spectrophotometric analyzer (observer 3), the Urotron RL9 (Boehringer Mannheim, Almere, the Netherlands). During the study, this analyzer was calibrated every day. The analyzer works by scanning the pads of a test strip with light from a light-emitting diode. The test result is determined by the amount of light reflected, which is dependent on the degree of discoloration of the test pad.

Both observers 1 and 2 had had detailed training in urine testing, which is included in the 3-year training course for practical nurses in the Netherlands. The experienced practical nurse finished this training 5 years ago without any reinforcement.

To prevent recognition of urine samples during the second series of measurements, the sequence of the samples was changed after the first series of measurements, using a list of random numbers. To avoid any influence of aging of the urine sample on the test results for the determination of intra-observer agreement, both the first and second measurement of every urine sample were performed within 1 hour. In this period, changes in urine samples are insignificant.[8] Every urine sample was mixed before a test strip was used in order to exclude any influence of insufficient homogenization.

For the determination of the reproducibility, interobserver and intra-observer agreement were calculated and expressed as Cohen's κ.[12]

The results of all measurements were expressed in two different ways: first in the standard graduations of test results (eg, −, +, ++, +++) and second in a more practical classification: normal versus abnormal. A negative test result was classified as normal; any positive test result was classified as abnormal.

RESULTS

All 90 urine samples were positive at the initial screening as follows: leukocyte esterase activity in 36 samples (40%), nitrite in 13 samples (14.4%), protein in 22 samples (24.4%), and blood in 39 samples (43.3%).

Interobserver agreement, expressed as κ, varied from 0.50 to 0.92 (*Table 1*). The highest agreement was achieved for the nitrite test, the lowest for the determination of blood.

Table 1. **INTEROBSERVER AGREEMENT, EXPRESSED AS COHEN'S κ**

	COMPARISON OF OBSERVERS		
TEST PAD	1 VERSUS 2	1 VERSUS 3	2 VERSUS 3
Leukocyte esterase activity	0.57	0.77	0.61
Nitrite	0.75	0.92	0.78
Blood	0.64	0.62	0.50
Protein	0.77	0.70	0.75
Mean	0.68	0.75	0.66

Observer 1 = practical nurse trainee
Observer 2 = practical nurse
Observer 3 = spectrophotometric analyzer

Table 2. **INTRA-OBSERVER AGREEMENT, EXPRESSED AS COHEN'S κ**

	OBSERVER		
TEST PAD	1	2	3
Leukocyte esterase activity	0.73	0.44	0.77
Nitrite	0.84	0.53	0.91
Blood	0.68	0.48	0.86
Protein	0.82	0.67	1.0
Mean	0.77	0.53	0.89

Observer 1 = practical nurse trainee
Observer 2 = practical nurse
Observer 3 = spectrophotometric analyzer

Table 3. **INTEROBSERVER AGREEMENT, EXPRESSED AS COHEN'S κ, AFTER CLASSIFYING TEST RESULTS IN TWO CATEGORIES**

	COMPARISON OF OBSERVERS		
TEST PAD	**1 VERSUS 2**	**1 VERSUS 3**	**2 VERSUS 3**
Leukocyte esterase activity	0.68	0.88	0.66
Nitrite	0.75	0.92	0.78
Blood	0.84	0.83	0.74
Protein	0.84	0.79	0.74
Mean	0.77	0.86	0.73

Observer 1 = practical nurse trainee
Observer 2 = practical nurse
Observer 3 = spectrophotometric analyzer

Table 4. **INTRA-OBSERVER AGREEMENT, EXPRESSED AS COHEN'S κ, AFTER CLASSIFYING TEST RESULTS AS NORMAL OR ABNORMAL**

	OBSERVER		
TEST PAD	**1**	**2**	**3**
Leukocyte esterase activity	0.81	0.55	0.92
Nitrite	0.84	0.53	0.91
Blood	0.83	0.74	0.93
Protein	0.81	0.84	1.0
Mean	0.82	0.67	0.94

Observer 1 = practical nurse trainee
Observer 2 = practical nurse
Observer 3 = spectrophotometric analyzer

Intra-observer agreement is shown in *Table 2.* κ has a wide range, from 0.44 to 1.0. The highest agreement was achieved for protein, the lowest for leukocyte esterase activity.

The observer with the best intra-observer agreement was the spectrophotometric analyzer. The experienced practical nurse had the lowest intra-observer agreement.

Table 3 shows interobserver agreement after test strip results were classified into normal and abnormal groups. Agreement increased in almost all situations.

After the same classification, intra-observer agreement was recalculated as well (*Table 4*). In general, agreement increased.

DISCUSSION

κ is an accepted measure for evaluating reproducibility in clinical medicine. Although there can be no objective interpretation (because κ measures agreement, not correctness), κ values lower than 0.40 are interpreted as low agreement, κ values between 0.40 and 0.75 as moderate to reasonable agreement, and κ higher than 0.75 as good agreement.[13] We believe that agreement should be good if a test is to be applied in clinical practice.

Reproducibility in general was moderate to good. However, it was not as good as is generally assumed. In the experiment the measurements were performed, after detailed instructions, by motivated observers in an optimal laboratory situation, set up to prevent the sort of errors that can occur in the field. Therefore we expect that reproducibility is lower in everyday practice.

Intra-observer agreement permits insight into the consistency of performance of each separate observer. As we expected, the highest intra-observer agreement was achieved by the spectrophotometric analyzer, which is not impeded by factors like lack of experience or weariness. It was, however, not always perfect: κ fell as low as 0.77.

Reproducibility of visual observations by an experienced practical nurse and an inexperienced practical nurse trainee was lower. Our study did not support Fraser's statement[10] that a correct interpretation of the discoloration of test strips depends on experience. On the contrary: the experienced practical nurse performed worse than her prospective colleague. Although we cannot draw generalized conclusions about the effect of experience, this finding is surprising.

Of the four test pads described, the lowest intra-observer agreement was achieved for leukocyte esterase activity. The reason for this might be the sometimes marginal difference in discoloration of the test pad, making differentiation, especially between + and ++, difficult.

Interobserver agreement varied, depending on the type of test pad, from 0.50 to 0.92. Except for nitrite, for which results are already dichotomic, agreement improved markedly after dichotomic classification of the results (normal versus abnormal). This classification is relevant for general practice because it is often enough to know whether the test result is normal or abnormal. Interobserver agreement was now between 0.66 and 0.92, which can be considered to reflect good reproducibility.

Intra-observer agreement after dichotomic classification increased to a range of 0.55 to 1.0. However, the difference between the practical nurse and the trainee remained. Note that, because we used no gold standard in this study, we cannot say that the trainee read the strips more accurately, only that she performed more consistently.

The aim of this study was to determine reproducibility, which was unknown. We expected good reproducibility, and we were surprised by our results.

Our study was, thus, not designed to gather information about the factors that limit reproducibility. Nevertheless, the meticulousness with which the study was performed allows us to exclude any influence of defects in the analyzer or in the test strips, changes in the urine samples, or insufficient knowledge of the observers. We presume that two factors account for the low reproducibility we found: the sometimes subtle differences in discoloration of test pads, and the underestimation of difficulties in reading test strips. This task is apparently not as simple as we might think.

CONCLUSION

Reproducibility of macroscopic urinalysis by means of test strips is in general moderate to good, but is lower than is generally assumed. This is especially true for macroscopic urinalysis using a spectrophotometric analyzer, which is only marginally more consistent than the best visual observer.

Reproducibility of visual observations might be improved if test pads discolored more strongly.

For now, macroscopic urinalysis is not a simple test procedure with guaranteed good reproducibility. ■

Acknowledgment

We thank Boehringer Mannheim, Almere, the Netherlands, for providing us with the test strips and Dr H. Schouten for his statistical advice.

Requests for reprints to: *Dr R.A.G. Winkens, General Practitioner, Diagnostic Centre Maastricht, PO Box 5800, NL-6202 AZ Maastricht, The Netherlands*

References

1. Sturmans F. *Epidemiologie: theorie, methoden en toepassing.* Nijmegen: Dekker & van de Vegt, 1982.
2. Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine.* Boston, Mass: Little, Brown and Company, 1985.
3. Triger DR, Smith JWG. Survival of urinary leucocytes. *J Clin Pathol* 1966;19:443-7.
4. Vaughan ED, Wyker AW. Effect of osmolality on the evaluation of microscopic hematuria. *J Urol* 1971;105:709-11.
5. Nanji AA, Poon R, Hinberg J. Effect of not allowing Reflotron strips to warm to room temperature (techn brief). *Clin Chem* 1988;34:179-80.
6. Spodick DH. On experts and expertise: the effects of variability in observer performance. *Am J Cardiol* 1975;36:592-6.
7. Gadeholt H. Quantitative estimation of urinary sediment, with special regard to sources of error. *Br Med J* 1964;1:1547-9.
8. Kierkegaard H, Feldt-Rasmussen U, Horder M, Andersen HJ, Jørgensen PJ. Falsely negative urinary leucocyte counts due to delayed examination. *Scand J Clin Lab Invest* 1980;40:259-61.
9. Hindman R, Tronic B, Barlett R. Effect of delay on culture of urine. *J Clin Microbiol* 1976;4:102-3.
10. Fraser CG. Urine analysis: current performance and strategies for improvement. *Br Med J* 1985; 291:321-3.
11. Yamane N, Sakamoto F, Matsuura F. Quantification of urinary glucose and protein with test-strips through reflectometric analysis. *Clin Biochem* 1988;21:271-5.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement* 1960;20:37-46.
13. Fleiss JL. *Statistical methods for rates and proportions.* New York, NY: John Wiley and Sons, 1981.