Diagnostic methods I: sensitivity, specificity, and other measures of accuracy

Karlijn J. van Stralen[1], Vianda S. Stel[1], Johannes B. Reitsma[2], Friedo W. Dekker[1,3], Carmine Zoccali[4] and Kitty J. Jager[1]

[1]Department of Medical Informatics, ERA-EDTA Registry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [2]Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [3]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands and [4]Renal and Transplantation Unit, CNR-IBIM Clinical Epidemiology and Pathophysiology of Renal Disease and Hypertension, Ospedali Riuniti, Reggio Cal, Italy

For most physicians, use of diagnostic tests is part of daily routine. This paper focuses on their usefulness by explaining the different measures of accuracy, the interpretation of test results, and the implementation of a diagnostic strategy. Measures of accuracy include sensitivity and specificity. Although these measures are often considered fixed properties of a diagnostic test, in reality they are subject to multiple sources of variation such as the population case mix and the severity of the disease under study. Furthermore, when evaluating a new diagnostic test, it must be compared to a reference standard, although the latter is usually not perfect. In daily practice diagnostic tests are not used in isolation. Several issues will influence the interpretation of their results. First, clinicians have a prior assumption about the patient's chances of having the disease under investigation, based on the patient's characteristics, symptoms, and the disease prevalence in similar populations. Second, diagnostic tests are usually part of a diagnostic strategy. Therefore, it is not sufficient to determine the accuracy of a single test; one also needs to determine its additional value to the patient's diagnosis, treatment, or outcome as part of a diagnostic strategy.

Kidney International (2009) **75,** 1257–1263; doi:10.1038/ki.2009.92; published online 1 April 2009

KEYWORDS: accuracy; diagnosis; diagnostic strategy; epidemiology; nephrology; sensitivity and specificity

Correspondence: *K.J. Jager, Department of Medical Informatics J1b, ERA-EDTA Registry, Academic Medical Center, PO Box 22700, Amsterdam 1100 DE, The Netherlands. E-mail: K.J.Jager@AMC.UVA.nl*

Received 19 December 2008; revised 27 January 2009; accepted 10 February 2009; published online 1 April 2009

In a physician's daily practice, the use of diagnostic tests is common. Tests refer not only to laboratory assessments, but also to medical history, observing signs and symptoms, and imaging techniques. However, tests are prone to errors; sometimes diagnoses are missed when test results are negative, but the opposite (a positive test result in the absence of disease) may also occur. To be able to decide whether it is useful to perform a test and how to interpret its outcome, it is important to have information on the quality of the test.

There are multiple aspects in the evaluation of a test, for example its reproducibility, that is if the same test is done again will it produce the same result; its accuracy, that is the amount of agreement between the results from the diagnostic test under study and those from a reference test; and its additional value to a diagnostic strategy, that is will the implementation of the diagnostic test into the routine package of diagnostic tests improve the patient's diagnosis, treatment, and his or her outcome. This paper will focus on the latter two aspects of a test, namely the different measures of accuracy, their interpretation, and drawbacks, and on its role within the diagnostic strategy.

DIFFERENT MEASURES OF ACCURACY

Diagnostic accuracy refers to the amount of agreement between the results from the diagnostic test under study and those from a reference test.[1] Several issues are important to determine the accuracy, and therefore multiple different measures have been developed.[1] For the illustration of all the concepts, we used an example from a study evaluating a laboratory test for detecting microalbuminuria in morning urine samples.[2] In this study, the general population responded to an invitation for a screening test. As a reference, a person was considered to have microalbuminuria if the urine albumin excretion determined by nephelometry exceeded 2.3 mg/l. All definitions and formulas are shown in Table 1, as are specific results for this test to detecting microalbuminuria, which are also presented in Figure 1.

Table 1 | The definition of the different measures of accuracy of a diagnostic and their application to a test diagnosing microalbuminuria[2]

		Reference standard[a] or 'truth'		
		Target condition[b]	No target condition	Total
Index test[c] result	Positive	TP 130	FP 357	487
	Negative	FN 23	TN 2017	2040
	Total	153	2374	2527

Term	Formula	Results	Definition
Sensitivity	TP/(TP + TN)	130/(130 + 23) = 85%	Probability of a positive test result among those having the target condition
Specificity	TN/(TN + FP)	2017/(2017 + 357) = 85%	Probability of a negative test result among those without the target condition
PPV	TP/(TP + FP)	130/(130 + 357) = 27%	Probability of having the target condition given a positive test result
NPV	TN/(FN + TN)	2017/(2017 + 153) = 99%	Probability of not having the target condition given a negative test result
Percent (positive) agreement	(TP + TN)/Total	(130 + 2017)/2527 = 85%	The percentage of patients correctly qualified
Positive likelihood ratio	Sensitivity/(1 − Specificity)	0.85/(1 − 0.85) = 5.7	Amount of certainty gained after a positive test result
Negative likelihood ratio	(1 − Sensitivity)/Specificity	(1 − 0.85)/0.85 = 0.18	Amount of certainty gained after a negative test result
Pre-test probability of disease	(TP + TN)/Total	(130 + 23)/2527 = 6.1%	Prevalence of disease in a population
Pre-test odds	Prevalence/(1 − Prevalence)	0.061/(1 − 0.061) = 0.064	Odds of disease before performing a test
Post-test odds			Odds of disease after test result
After a positive test result	Pre-test odds × positive likelihood ratio	0.064 × 5.7 = 0.37	Odds of disease after a positive test result
After a negative test result	Pre-test odds × negative likelihood ratio	0.064 × 0.18 = 0.012	Odds of disease after a negative test result
Post-test probability of disease			Probability of disease after test result
After a positive test result	Post-test odds/(post-test odds + 1)	0.37/(0.37 + 1) = 26.9%	Probability of disease after a positive test result
After a negative test result	Post-test odds/(post-test odds + 1)	0.012/(0.012 + 1) = 1.14%	Probability of disease after a negative test result

FN, false negative; NPV, negative predictive value; PPV, positive predictive value; TP, true positive.

[a]The best available method for establishing the presence or absence of the target condition. [b]Disease or other health status. [c](New) Diagnostic test under study.

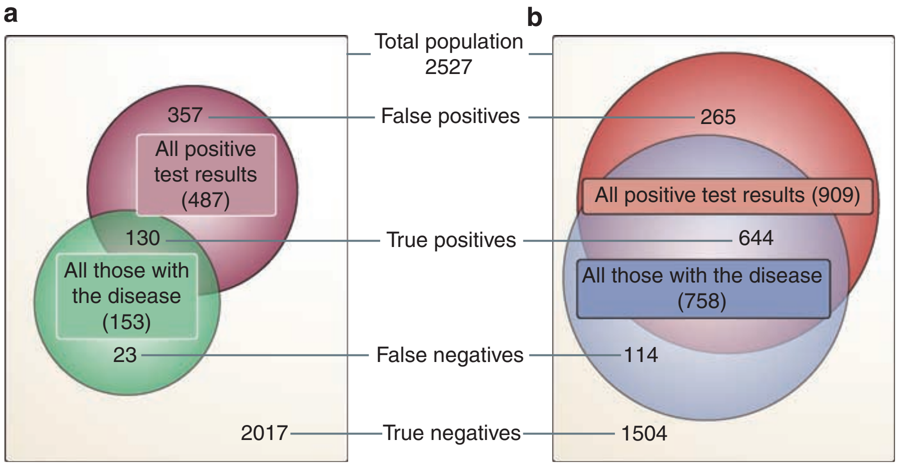
a
b
Total population
2527
357
All positive test results (487)
False positives
265
All positive test results (909)
130
True positives
644
All those with the disease (153)
All those with the disease (758)
23
False negatives
114
2017
True negatives
1504

Figure 1 | Venn diagram of the dependency of true positive, true negative, false negative, false negative, and false positive results in a test for microalbuminuria. Venn diagram of the distribution of the true positive, false negative, and false positive results for a test to detect microalbuminuria (sensitivity 85%, specificity 85%) in a population of 2527 individuals of whom 6% have the disease (**a**), and in a similar size population of whom 30% have the disease (**b**).[2]

A perfect test will show a positive result for all those who have the target condition (for example, the disease or other health outcomes). Its ability to do this is described by its sensitivity.[3] Sensitivity is the proportion of all patients with the disease (true positives + false negatives) who indeed have a positive test result (true positives). However, a high sensitivity alone does not make a test a good test. The test also needs to be negative for all those without the disease. This ability is expressed by the specificity of the test. It is the proportion of all patients without the disease and a negative test result (true negatives) of all those without the disease (true negatives + false positives).

In the study evaluating the diagnostic accuracy of the microalbuminuria test, the sensitivity and specificity were also calculated. Out of the 153 patients with microalbuminuria, 130 had a positive test result with the urine test under study, resulting in a sensitivity of 85% (130/153). Of the 2374 individuals without microalbuminuria, 2017 had a negative urine test, resulting in a specificity of 85% (2017/2374).

In daily practice, the clinician and the patient will be more interested in the positive (PPV) and negative predictive value (NPV).[4] The PPV indicates the probability of having the disease after a positive test result, whereas the NPV is the probability of not having the disease after a negative test result (Table 1). So, when we use the earlier example, the urine test showed a positive result in 487 persons, of whom 130 actually had microalbuminuria. Therefore, the PPV was 27% (130/487); that is, among those with a positive test result, 27% actually had microalbuminuria. The urine test was negative in 2040 patients, of whom 2017 did not have microalbuminuria. Therefore, the NPV was 99%. On the basis of these predictive values of this or other tests, the physician may decide to perform additional tests, start a treatment, or send the patient home.

Determine which accuracy is acceptable and whether one prefers a higher specificity but lower sensitivity or vice versa is not straightforward. Although values close to 100% are ideal, there are situations in which one could prefer a test with a lower sensitivity or specificity over another with a higher sensitivity or specificity. Sometimes a new test is a triage, that is will be used before a second test, and only those patients with a positive result in the triage test will continue in the testing pathway. For such a test, one may accept a lower accuracy than that of the existing tests, as this triage test is not meant to replace the second test. For example, in detecting chronic kidney disease, an inexpensive dipstick test could be preferred as a triage test, allowing many individuals to be tested. In this test, it is important that all patients with chronic kidney disease have a positive test result (high sensitivity), whereas the number of patients with false-positive results (low specificity) is considered somewhat less important, as they would be identified using a second and the subsequent tests.

Another situation in which a diagnostic test with a lower sensitivity and specificity could be preferred is when the test with the better accuracy has a high risk of complications. For example, as performing renal arteriography to test for renal artery stenosis is an invasive diagnostic method, with potential complications, one could prefer to replace arteriography with Doppler testing, which has 89% sensitivity and 73% specificity.[5] In this way, one may attempt to balance the desirable and undesirable consequences of performing different diagnostic tests.[6]

The sensitivity, specificity, PPV, and NPV together result in four different measures, each indicating the accuracy of the test. All these measures have different pros and cons, and they may be difficult to interpret.[7,8] Therefore, one sometimes prefers a combination of them. Frequently used parameters are the percentage of patients correctly qualified, the likelihood ratio, and the pre- and post-test probability. The percentage of patients correctly qualified is the number of concordant individuals, that is those with a positive test with the disease + those with a negative test without the disease (true positives and true negatives in Table 1, respectively), divided by the total population tested. So in the test for microalbuminuria, 2147 (130 + 2017) out of 2527 individuals were correctly classified, resulting in a correctly classified percentage of 85%. This parameter is easy to understand and can be used when there are multiple categories. A disadvantage is that when there are a large number of individuals without the disease, as in our example, the proportion correctly qualified is determined mainly by the specificity and not by the sensitivity.

The likelihood ratio of a positive test result reflects the amount of certainty of having the disease that is gained after positive test reflects result, whereas the likelihood ratio of a negative test result is the amount of certainty gained of not having the disease with a negative test result. The positive likelihood ratio can be calculated by dividing the sensitivity by 1−specificity. It could be interpreted as 'in patients with microalbuminuria, a positive test is found 5.7 times as often as in patients without microalbuminuria.' The negative likelihood ratio on the other hand is the amount of information that is gained after a negative test result. It is calculated by 1−sensitivity divided by the specificity. A likelihood ratio close to 1 indicates that performing the test provides little additional information regarding the presence or absence of the disease. The likelihood ratios have the advantage of putting equal weights to the sensitivity and specificity and therefore being less dependent on the proportion of individuals under study who are diseased versus non-diseased.

One could also compare the pre- and post-test probability of a disease. The pre-test probability is equal to the prevalence of the disease in the population under study. The post-test probability is the chance of the disease after a positive (or negative) test result. For example, the prevalence or pre-test probability for an individual to have microalbuminuria was 6%. Using the calculation in Table 1, after a positive test-result, we could estimate that the post-test probability was equal to 27%. This number is equal to the PPV. Likewise, after a negative test result, the chance of having the disease decreased to a post-test probability of 1% (which is equal to 1−NPV).

SOURCES OF BIAS AND VARIATION

Many diagnostic tests have been introduced with great enthusiasm because of their high sensitivity and specificity, but have nevertheless been rejected at a later stage. One of the reasons for this rejection was that measures of accuracy of a diagnostic test have often been presented as inherent and fixed properties, whereas in reality, they can be affected by different sources of variation and bias.[9] We will now discuss some of the important sources of variation.

Reference standard

To determine the measures of accuracy of a test, we need to compare the test result with a reference standard that reflects 'the truth', that is, that can tell us with great certainty whether or not the patient has the disease. Ideally, this reference standard is an existing test with a sensitivity and specificity of 100%. As reference standards, similar to other tests, can be costly, invasive, or impractical, one may sometimes want to use the newly developed diagnostic test instead of the reference standard.

When performing a diagnostic accuracy study, it may be considered unethical to apply the reference standard to someone who is asymptomatic, for example performing a biopsy in transplant patients without rejection symptoms to identify a new test for determining acute renal failure.[10] In this situation, only cases with a strong indication of rejection, for example those with a positive result using the new diagnostic test, will receive a biopsy. As not all patients are tested with the reference standard, the number of patients with a false-negative result is too low, resulting in an overestimation of the sensitivity of the new test. This has been called workup or verification bias.[11,12]

Although the reference standard is considered the best available test at a certain moment, it will most likely not have a sensitivity and specificity of 100%. Therefore, additional information is needed. A first solution is to combine multiple less accurate tests. Together, these tests could provide a sufficient reference standard.[13] A second solution is to use an expert-based reference standard. A group of experts will then decide whether a patient has the disease. However, this may be subject to bias if the experts have their opinion based on the outcome of one or more diagnostic tests.[11] A third solution is to follow those patients with a negative test result in time to test whether they develop the disease later in life. This has, for example, been done in the diagnosis of pulmonary embolism.[14] In summary, frequently, a reference standard with a perfect sensitivity and specificity is unavailable, and in those cases, the measures of accuracy of a new diagnostic test can only be compared with an imperfect standard, resulting in incorrect estimates of the measures of accuracy of the new test.

Case mix and disease severity

It is known for a long time that both the case mix and the disease severity can affect the measures of diagnostic accuracy of a test. One could study the sensitivity and specificity of C-peptide levels to classify diabetes patients as type I or type II in patients with kidney failure. However, as the kidney is the major site of C-peptide metabolism and excretion,[15] one can easily imagine finding a different sensitivity and specificity in a patient population with an altered kidney function, compared with the results in a diabetic population without chronic kidney disease.[16,17] Consequently, when comparing results between populations with varying rates of kidney failure one would find different values for the sensitivity and specificity. Situations similar to this occur often, and in general for many tests it can be conceivable that one could obtain different sensitivities and specificities in women versus men, for elderly versus younger individuals and in patients with and without (kidney) disease. For this reason, it is needed to present strata-specific estimates of the measures of disease accuracy for relevant subgroups.[18,19]

Also, the stage of disease severity could affect the sensitivity and specificity of a test. Sherwood[20] showed a different sensitivity for renal ultrasound in patients with different renal masses, that is he found a sensitivity of 97% for predicting a renal cyst but that of 60% for detecting a renal carcinoma. In general, the measures of accuracy can be affected by the type of renal disease, stage of cancer, or level of glomerular filtration rate and so on. Therefore, specification of the sensitivity and specificity of the test for each severity of a disease is needed.[11]

This shows that tests may perform differently in different groups of subjects and for different severities of disease. Therefore as a clinician, it is important to compare the patient to be tested with the general characteristics or case-mix of the study population described in a paper to see to what extent the test properties provided in that paper apply to this specific patient.[1]

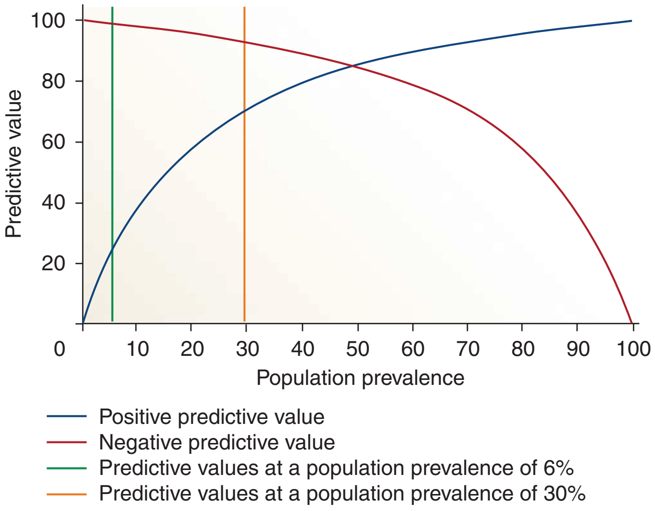

Figure 2 | Effect of population prevalence on the values of positive and negative predictive values, using a test with a sensitivity and specificity of 85%.

Population prevalence

A feature of PPV and NPV is their dependence on the prior probability of the disease (which is equal to the prevalence of disease in the population to be tested).[21] As is shown in Figure 1 and Figure 2, if the disease prevalence increases, a positive test result will have a higher PPV. This is due to a relative decrease in the number of patients with a false-positive result to the number of true positives. As a consequence, the proportion of true positives among the total number of those with a positive test result will rise, resulting in a higher PPV. The opposite will occur for NPV, that is a higher disease prevalence will result in a lower NPV. Therefore, when comparing diagnostic accuracy measures between different populations, different results can be obtained. For example, in the microalbuminuria test with a population prevalence of nearly 6%, the PPV was 27% whereas the NPV was 99%. If the prevalence would have been higher, for example when studying a predialysis population in which 30% could have microalbuminuria, the PPV would have increased to 71%, whereas the NPV would have decreased to 93% (test results as shown in Figure 1b). A similar variation of PPV and NPV, as a result of variation in population prevalence, occurs when one would compare the results of this test across, for example, different ethnicities, different causes of renal disease, and different sexes.

BAYES' THEOREM

When a young woman is presented to a nephrologist with fever, fatigue, and proteinuria, the nephrologist is likely to test for systemic lupus erythematosus (SLE). If a nuclear antibody test (sensitivity 94%, specificity 97%) would confirm the disease, the physician would probably accept the diagnosis. In addition, if the test would be negative, it is likely that the nephrologist will perform additional tests, as this patient with classical symptoms has a high prior risk for SLE. Conversely, when an old man is presenting with the same symptoms, when having a negative test result, the patient will be considered negative for SLE, as he has a much lower prior chance of disease. When this patient would have a positive test result, most physicians would still not be fully convinced of the diagnosis.

As this example shows, the interpretation of the results of a diagnostic test not only depends on the accuracy, the sensitivity, and specificity, but also on the prior chance of having the disease. This has been called the Bayes' theorem. A physician could estimate the prior chance of SLE in this young woman to be around 50%. After a positive test result, by using the formula from Table 1, her post-test probability of the disease has increased from around 50% to nearly 97%. On the other hand, in the man, the pretest probability of the disease can be estimated by the physician to be closer to 5%. Therefore, after a positive test result, his chance of having the disease is still only 62%, which is much lower than that in the young woman.

This also explains why multiple diagnostic tests are needed when screening the general population. When this very good nuclear antibody test would be used to screen the US population for SLE (prevalence 33/100,000 individuals), approximately 9 million individuals would test positive, of whom only 1% have the disease. Although the posterior probability for a single individual in this screening test after a positive test result has increased from 0.03 to 1%, it is still unlikely that this person truly has SLE, making screening using a single diagnostic test not an efficient method.[22]

DIAGNOSTIC STRATEGY

Besides the accuracy of a test and the clinician's interpretation of the result, it is also important to know whether adding this test to the current diagnostic strategy will improve the patient's diagnosis, treatment, and outcome.[6,23] Most test results will have an incremental value on top of each other, and a test may be of value even if there exist other more accurate tests.[12] A new test could be added to the existing diagnostic strategy in three ways.[24] First, as a triage, the new test could be implemented before the existing test, and only patients with a positive result would continue to the existing testing pathway. Triage test may be less accurate, but a very high sensitivity is important. Preferably, when using a triage test, one prevents performing the more expensive or time-consuming second test in everyone. Instead, one selects only those who are at a high risk. An example of a triage test is a dipstick test for microalbuminuria to detect chronic kidney disease. Second, a new test may be needed to replace an existing test. In this situation, the new test may, for example, be more accurate, less invasive, or easier to do. In this regard, one may think of examples like replacement of the angiography by the Doppler ultrasound. Finally, a new test may be added to a diagnostic strategy to improve the overall diagnosis. In this situation, the new test needs to increase the overall sensitivity of the diagnosis, which is probably at the expense of the specificity.

Therefore, when developing a diagnostic test, one should not only determine and compare the accuracy of a new test with that of the old one and the reference standard, but also study the new test as part of the routine package of tests. This gives the possibility to determine the downstream consequences of using a diagnostic test for outcome measures, such as overall mortality, time to discharge, or cost-effectiveness. Besides the type of tests that are performed, the order in which the tests are performed also could be important.[25] Furthermore, one could calculate the post-diagnostic strategy—the chance of disease by multiplying the post-test probabilities of the different individual tests.

The ideal way to determine the effect of the new strategy is to perform a randomized controlled trial, with one group having the 'old strategy' for diagnosing the disease and the other having the 'new strategy' involving the new test.[12] In this way, it is possible to determine the effect of the new therapy on the treatment outcome and the cost-effectiveness of the diagnostic strategy. An example of such a trial is the study on the usefulness of b-type natriuretic peptide in the diagnosis of congestive heart failure. Earlier studies had shown

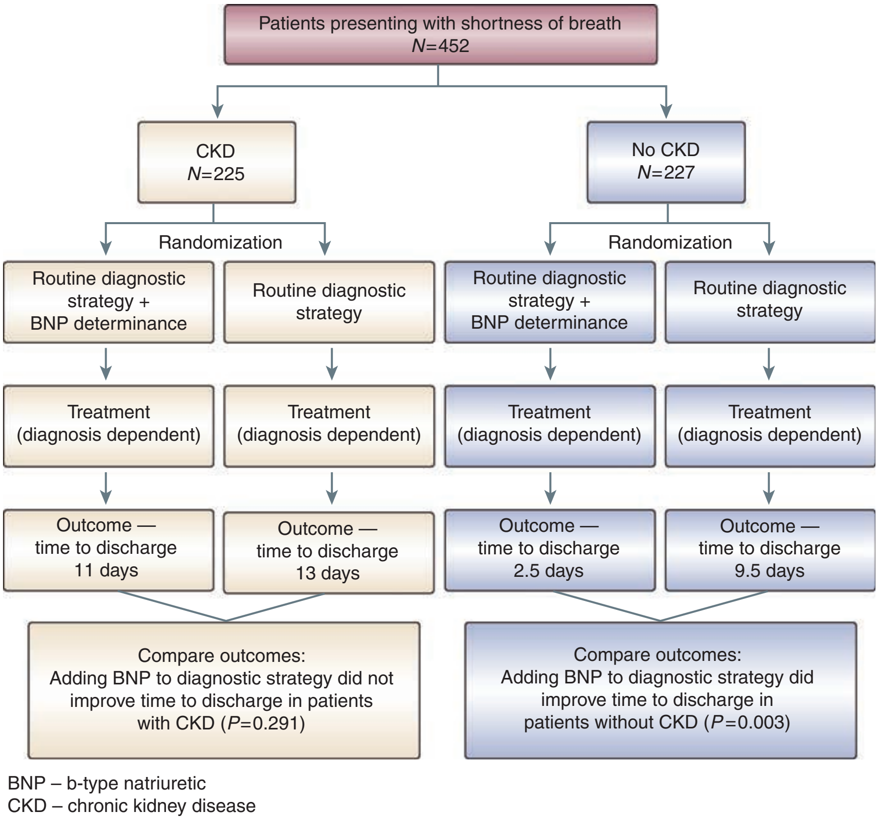
Patients presenting with shortness of breath
N=452
CKD
N=225
No CKD
N=227
Randomization
Randomization
Routine diagnostic strategy + BNP determinance
Routine diagnostic strategy
Routine diagnostic strategy + BNP determinance
Routine diagnostic strategy
Treatment (diagnosis dependent)
Treatment (diagnosis dependent)
Treatment (diagnosis dependent)
Treatment (diagnosis dependent)
Outcome — time to discharge 11 days
Outcome — time to discharge 13 days
Outcome — time to discharge 2.5 days
Outcome — time to discharge 9.5 days
Compare outcomes: Adding BNP to diagnostic strategy did not improve time to discharge in patients with CKD (P=0.291)
Compare outcomes: Adding BNP to diagnostic strategy did improve time to discharge in patients without CKD (P=0.003)
BNP – b-type natriuretic
CKD – chronic kidney disease

Figure 3 | Example of a randomized controlled trial comparing two diagnostic strategies, based on the study by Mueller *et al.*[27]

approximately similar sensitivity and specificity for those with and without chronic kidney disease.[26] However, when the additional value of b-type natriuretic peptide measurement as a 'triage' to the diagnostic strategy was studied as shown in Figure 3, results were different.[27] Using this randomized controlled trial it was shown that adding the b-type natriuretic peptide test was efficient in those without chronic kidney disease, but inefficient in those with chronic kidney disease. This indicates that although a test can have measures of accuracy that are very high, this does not necessarily imply that adding them to the current diagnostic strategy will improve a patient's diagnosis, treatment, and/or outcome.

CONCLUSION

Sensitivity and specificity are measures to assess the accuracy of a diagnostic test. In recent years, several guidelines have been published on how to report on diagnostic tests, which are recommended for further reading.[6,28] In this paper, we explained the different measures of accuracy. For their calculation, one needs a reference standard. However, as very few tests are perfect, often an imperfect reference is used. Furthermore, due to several biases and sources of variation, such as differences in case mix, and disease severity, the measures of accuracy cannot be considered as fixed properties of a diagnostic test.

In addition, the measures of accuracy are not isolated instances. In general, clinicians have a prior assumption about the patient's chances of disease based on the patient's characteristics, symptoms, and disease prevalence. This will influence the posterior chance of the disease. Furthermore, as nearly all tests are performed within a range of tests, one needs to consider the additional value of a new test on top of or as replacement for the current diagnostic strategy, especially with respect to the patient's diagnosis, treatment, and outcome.

DISCLOSURE

All the authors declared no competing interests.

REFERENCES

1. Bossuyt PM, Reitsma JB, Bruns DE *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003; **49**: 7–18.
2. Gansevoort RT, Verhave JC, Hillege HL *et al.* The validity of screening based on spot morning urine samples to detect subjects with microalbuminuria in the general population. *Kidney Int Suppl* 2005; **67**: S28–S35.
3. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 1994; **308**: 1552.
4. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ* 1994; **309**: 102.
5. Avasthi PS, Voyles WF, Greene ER. Noninvasive diagnosis of renal artery stenosis by echo-Doppler velocimetry. *Kidney Int* 1984; **25**: 824–829.
6. Schunemann HJ, Oxman AD, Brozek J *et al.* Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336**: 1106–1110.

7. Puhan MA, Steurer J, Bachmann LM *et al.* A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. *Ann Intern Med* 2005; **143**: 184–189.
8. Steurer J, Fischer JE, Bachmann LM *et al.* Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002; **324**: 824–826.
9. Whiting P, Rutjes AW, Reitsma JB *et al.* Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004; **140**: 189–202.
10. Aquino-Dias EC, Joelsons G, da Silva DM *et al.* Non-invasive diagnosis of acute rejection in kidney transplants with delayed graft function. *Kidney Int* 2008; **73**: 877–884.
11. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978; **299**: 926–930.
12. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6**: 411–423.
13. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med* 1999; **18**: 2987–3003.
14. van Belle A, Buller HR, Huisman MV *et al.* Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *JAMA* 2006; **295**: 172–179.
15. Robaudo C, Zavaroni I, Garibotto G *et al.* Renal metabolism of C-peptide in patients with early insulin-dependent diabetes mellitus. *Nephron* 1996; **72**: 395–401.
16. Covic AM, Schelling JR, Constantiner M *et al.* Serum C-peptide concentrations poorly phenotype type 2 diabetic end-stage renal disease patients. *Kidney Int* 2000; **58**: 1742–1750.
17. Service FJ, Rizza RA, Zimmerman BR *et al.* The classification of diabetes by clinical and C-peptide criteria. A prospective population-based study. *Diabetes Care* 1997; **20**: 198–201.
18. Coughlin SS, Trock B, Criqui MH *et al.* The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *J Clin Epidemiol* 1992; **45**: 1–7.
19. Diamond GA. Clinical epistemology of sensitivity and specificity. *J Clin Epidemiol* 1992; **45**: 9–13.
20. Sherwood T. Renal masses and ultrasound. *BMJ* 1975; **4**: 682–683.
21. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997; **16**: 981–991.
22. Loong TW. Understanding sensitivity and specificity with the right side of the brain. *BMJ* 2003; **327**: 716–719.
23. Guyatt GH, Tugwell PX, Feeny DH *et al.* A framework for clinical evaluation of diagnostic technologies. *CMAJ* 1986; **134**: 587–594.
24. Bossuyt PM, Irwig L, Craig J *et al.* Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006; **332**: 1089–1092.
25. England WL, Grim CE, Weinberger MH *et al.* Cost effectiveness in the detection of renal artery stenosis. *J Gen Intern Med* 19883: 344–350.
26. Burke MA, Cotts WG. Interpretation of B-type natriuretic peptide in cardiac disease and other comorbid conditions. *Heart Fail Rev* 2007; **12**: 23–36.
27. Mueller C, Laule-Kilian K, Scholer A *et al.* B-type natriuretic peptide for acute dyspnea in patients with kidney disease: insights from a randomized comparison. *Kidney Int* 2005; **67**: 278–284.
28. Bossuyt PM, Reitsma JB, Bruns DE *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003; **138**: 40–44.