Test Research versus Diagnostic Research

The diagnostic workup starts with a patient presenting with symptoms or signs suggestive of a particular disease. The workup is commonly a consecutive process starting with medical history and physical examination and simple tests followed by more burdensome and costly diagnostic procedures. Generally, after each test all available results are converted (often implicitly) to a probability of disease, which in turn directs decisions for additional testing or initiation of appropriate treatment. Setting a diagnosis is a multitest or multivariable process of estimating and updating the diagnostic probability of disease presence given combinations of test results. Each test may be more or less burdensome to the patient, time-consuming, and/or costly. Different tests often provide to various degrees the same information because they are all associated with the same underlying disorder. Relevant for physicians is to know which tests are redundant and which have true, independent predictive value for the presence or absence of the target disease. Accordingly, studies of diagnostic accuracy should demonstrate which (subsequent) test results truly increase or decrease the probability of disease presence as estimated from the previous results, and to what extent.

Various reviews have demonstrated that the majority of published studies of diagnostic accuracy still have methodologic flaws in design or analysis or provide results with limited practical applicability *(1–3)*. This has been attributed to the absence of a proper methodologic framework for diagnostic test evaluations as, for example, exists for studies of therapies and etiologic factors and has motivated various researchers to establish frameworks for studies of diagnostic accuracy, such as the recent STARD Initiative *(4–12)*. In our view, an issue that has received too little attention in most of these methodologic essays is the difference between test research and diagnostic research.

By “test research” we refer to studies that follow a single-test or univariable approach, i.e., studies focusing on a particular test to quantify its sensitivity, specificity, likelihood ratio (LR), or area under the ROC curve (ROC area). We call this test research because it merely quantifies the “characteristics” of the test rather than the test’s contribution to estimate the diagnostic probability of disease presence or absence. By “diagnostic research” we refer to studies that aim to quantify a test’s added contribution beyond test results readily available to the physician in determining the presence or absence of a particular disease. Although the multivariable and probabilistic character of medical diagnosis is slowly gaining appreciation in medical research, the majority of studies on diagnostic accuracy may still be regarded as test research *(2, 3, 8)*.

We believe that test research has limited applicability to clinical practice. Below we describe why we believe this is the case, provide a brief description of a better approach, and give two clinical examples illustrating the hazards of test research. Finally, we describe the few instances in which test research may be worthwhile.

Why Does Test Research Have Limited Relevance to Practice?

study question and object of study

The first reason that test research has limited relevance to practice is the nature of the questions that are usually addressed. The practical utility of estimation of sensitivity, specificity, and LR for a particular test in the diagnosis of a particular disease is not always obvious *(7, 13)*. Consider, for example, the diagnostic workup for patients suspected of deep vein thrombosis (DVT). The relevant research question for patients suspected of DVT would be: “Given patient history and physical examination, which subsequent tests (e.g., d-dimer measurement) truly provide added information to predict the presence or absence of DVT?” The probability of disease presence and quantifying which tests independently contribute to the estimation of this probability should be the objects of study. However, in this respect many studies have aimed only to estimate the sensitivity and specificity of the d-dimer assay. When this is the object of a study, it is only the probability of obtaining a positive or negative test result that is addressed, rather than the probability of disease presence. Moreover, the focus is on the value of a single test rather than on the value of that test in combination with other, previous tests, including patient history and physical examination. We may say that the object of research is the test rather than the (probability of) disease. Hence the term test research.

test characteristics are not fixed

The second reason that results from test research have limited relevance is that a test’s sensitivity, specificity, LR, and ROC area tend to be taken as properties or characteristics of a test. This, however, is a misconception, as we discussed recently *(13)*. It is widely accepted that the predictive values of a test vary across patient populations. However, several studies have empirically shown that the sensitivity, specificity, and LR of a test may vary markedly, not only across patient populations *(14)* but also within a particular study population *(13, 15–17)*. Within different patient subgroups, defined by patient characteristics or other test results, a particular test may have different sensitivities and specificities. This is because all diagnostic results obtained from patient history, physical examination, and additional tests are to some extent related to the same underlying disorder. For example, immobility, gender, and use of oral contraceptives are associated with the development of, and thus the presence of, DVT. In turn, the presence of DVT determines the presence of symptoms and signs and also (the probability of finding) a positive d-dimer assay result. Accordingly, via the underlying disorder, all diagnostic results are somehow correlated and thus mutually determine each other's sensitivity, specificity, and LR to various extents *(13, 15–17)*. A single value of a test's sensitivity, specificity, LR, ROC area, or predictive value that applies to all patients of a study sample does not exist. Hence, there are no fixed test characteristics.

SELECTION BIAS

The most widely acknowledged limitation of test research is that studies often apply an improper patient recruitment and study design *(1–3, 7)*. Investigators often select study participants among those who underwent the reference test in routine practice, i.e., selection based on a "true" presence or absence of the disease. The results of the test(s) under study are retrieved from the medical records and then compared across those with and without the disease. Such a case–control design commonly leads to selection bias, known as verification, workup, or referral bias *(9, 18, 19)*.

Although such patient recruitment methods and study designs have decreased in the past decade, test research is still frequently based on individuals selected based on their final diagnosis *(1–3)*. The need for proper patient recruitment is extensively addressed in the STARD checklist *(11, 12)*. Study participants should be selected in agreement with the indication for diagnostic testing in practice, i.e., on their suspicion of having a particular disease, rather than on the presence or absence of that disease. Such unbiased selection of study participants may indeed be problematic for diagnostic laboratories or imaging centers that do not have access to consecutive series of patients suspected of having the disease. Moreover, most hospital databases code patients according to their final diagnosis rather than by their presenting symptoms or signs. The use of a system to register patients not only on their final diagnosis but also on their clinical presentation would enhance the validity and clinical relevance of diagnostic accuracy research *(20)*.

Proposed Approach for Diagnostic Accuracy Research

We believe that to serve practice, the point of departure and the multivariable and probabilistic character of the diagnostic workup should be reflected in the objective, design, analysis, and presentation of studies of diagnostic accuracy. The aim is to relate the probability of disease presence to combinations of test results, following their typical chronology in practice. The predictive accuracy of the initial tests (including patient history and physical examination) should be estimated first, and the added value of more burdening and costly tests should be estimated subsequently. Hence, all tests typically applied in the workup need to be documented in each patient, even if a study focuses on a particular test. Consider again the question whether the d-dimer assay is relevant to the diagnosis of DVT. A consecutive series of patients suspected of DVT should be selected. The history, physical examination, and d-dimer result should be obtained from each patient. Subsequently, each patient "undergoes" the best reference test currently available; in this example, it would be repeated leg ultrasound. What to do in the absence of a single reference test or when it is unethical to perform the reference test in each patient has been described elsewhere *(8, 10, 21, 22)*.

Because the d-dimer assay will always be applied after history taking and physical examination, the statistical analysis requires a comparison of the (average) probability of disease presence without and with the d-dimer assay, overall or in subgroups. Such sequential modeling of the diagnostic probability as a function of different combinations of test results can be done using, e.g., multivariable logistic regression. Such multivariable analyses account for the mutual dependencies between different test results and thus indicate which tests truly do and which do not independently contribute to the estimation of the probability of disease presence. In addition, various orders of diagnostic testing can be analyzed. The result of such analysis is the definition of one or more diagnostic prediction models including only the relevant tests. If needed, such prediction models can be simplified to obtain readily applicable diagnostic decision rules for use in practice. Various authors have applied or described the details of such an analytical approach *(20, 23–27)*.

Multivariable diagnostic prediction models or rules are not the solution to everything. They may have several drawbacks, such as overoptimism, although methods have been described to overcome some of these drawbacks *(23)*. The need for multivariable modeling in diagnostic research, however, is not different from other types of medical research, such as etiologic, prognostic, and therapeutic research. It is not the singular association between a particular exposure or predictor and the outcome that is informative, but their association independent of other factors. For example, in etiologic research, investigators never publish the crude estimate between exposure and outcome only, but always the association in view of other risk factors (confounders), using a multivariable analysis as well *(13)*. Similarly, in diagnostic accuracy research, multivariable modeling is necessary to estimate the value of a particular test in view of other test results. As in other types of research, such knowledge cannot be inferred from singular, univariable test parameters *(7, 8, 13)*.

Fortunately, a multivariable approach in design and analysis aiming to quantify the independent value of diagnostic tests has gained approval *(20, 23–27)*. In addition, the above study question on the added value of the d-dimer assay in diagnosing DVT has been evaluated in such a way. The d-dimer assay appeared to have an added predictive value to patient history and physical examination, particularly in patients who have a low clinical probability of DVT *(27)*.

Clinical Examples

We now present two clinical examples illustrating how results from a single or univariable test approach can mislead.

In an Australian study, 399 consecutive dyspeptic patients referred for endoscopy underwent two tests, the rapid urease test and the ^{13}C breath test, for *Helicobacter pylori* (HP) with endoscopy as the reference test *(28)*. The investigators found large differences in the test results between patients with a normal and abnormal endoscopy. The sensitivity and specificity were 96% and 67% for the rapid urease test and 91% and 82% for the ^{13}C breath test. The authors concluded that the HP tests might have potential for the initial evaluation of dyspepsia and needed further evaluation in general practice. A second study was done by Weijnen et al. *(26)*. Using a sequential multivariable approach, they found in a consecutive series of 565 dyspeptic patients referred for endoscopy that the HP test did not add diagnostic information to the predictors from history (i.e., history of ulcer, pain on empty stomach, and smoking). The ROC area of the model with only predictors from patient history was 0.71, which was increased to only 0.75 ($P = 0.46$) after addition of the HP test result. They concluded that HP testing in all dyspeptic patients has no value in addition to history taking.

Cowie et al. *(29)* studied a consecutive series of 122 patients suspected of heart failure. They measured in each patient the plasma concentrations of three natriuretic peptides, A-type natriuretic peptide (ANP), N-terminal ANP, and B-type natriuretic peptide (BNP), as well as the presence or absence of heart failure, using consensus diagnosis based on chest radiography and echocardiography as the reference test. They found that the mean concentration of each natriuretic peptide separately (single-test approach) was significantly greater in the patients with heart failure (all $P < 0.001$). They also evaluated all three together in a multivariable logistic prediction model. Only the BNP measurement remained significantly associated with heart failure presence, whereas the other two did not add any predictive information.

Both examples show that one may qualify a test differently (commonly more promisingly) when only the results of a univariable or single-test approach are considered. Evaluating a particular test in view of other test results and accounting for mutual dependencies may decrease or even diminish its diagnostic contribution, simply because the information provided by that test is already provided by the other tests. Because in real life any test result is always considered in view of other patient characteristics and test results, diagnostic accuracy studies that address only a particular test and its characteristics have, in our view, limited relevance to practice. Indeed, as shown by Reid et al. *(30)*, test characteristics are hardly ever actually used by practitioners.

Is There a Place for Test Research?

There are two situations in which pure test research, i.e., studies aiming to estimate the diagnostic accuracy indices of a single test, is indicated. The first situation is when a diagnosis is indeed set by only one test and other test results are not considered. This is, in our view, reserved to the context of screening for preclinical stages of a particular disease only: e.g., screening for breast cancer, prostate cancer, or cervical cancer. Such screening may be considered as a specific case of diagnosis, concerned with the early detection of a disease in a particular age and sex group. Here, only the screening test is considered in the diagnostic process; other patient characteristics or test results are commonly not available and therefore cannot modify the sensitivity, specificity, LR, and predictive values of the screening test. Accordingly, these indices, as estimated from a particular study sample, may be considered characteristics or constants for the corresponding source population. In the presence of a positive screening result, patients are commonly referred for further diagnostic workup. Other test results then become involved, and mutual dependencies between the screening test and these other tests start to play a role, demanding a multivariable approach in design and analysis.

The second situation, as suggested previously, is in the initial phase of developing a new test or evaluating an existing test in a new context; single-test evaluations in these circumstances may be useful for efficiency reasons *(4, 6, 7, 25)*. Such initial test research should apply a case–control approach, preferably starting with a sample of patients with the disease (cases) and a sample of healthy controls. If the test cannot differentiate between these two extreme or heterogeneous outcome categories, the test development process would likely be terminated. In such instances, it will be unlikely that the test does show discriminative value in patients suspected of having the disease, i.e., the population for which the test is intended, because these patients present with similar disease profiles, leading to an even more homogeneous case mixture. However, once the test does yield “satisfactory” diagnostic indices in such an initial test research study, we believe that its independent predictive contribution to existing diagnostic information in a clinical context can and must still be quantified by the above proposed approach.

We gratefully acknowledge The Netherlands Organization for Scientific Research for their support (No. 904-66-112).

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.
2. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.

3. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85–91.

4. Fryback D, Thornbury J. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88–94.

5. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–91.

6. van der Schouw YT, Verbeek ALM, Ruijs JHJ. Guidelines for the assessment of new diagnostic tests. Investig Radiol 1995;30:334–40.

7. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539–41.

8. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 2002;56:337–8.

9. Knottnerus JA. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:237pp.

10. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002;222:604–14.

11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;49:1–6.

12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.

13. Moons KGM, Harrell FE. Sensitivity and specificity should be deemphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–2.

14. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986;104:66–73.

15. Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64–71.

16. Levy D, Labib SB, Anderson KM, Christiansen JC, Kanell WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990;81:815–20.

17. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.

18. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926–30.

19. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:297–315.

20. Oostenbrink R, Moons KGM, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003;56:501–6.

21. Moons KGM, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic research. J Clin Epidemiol 2002;55:633–6.

22. Bossuyt PPM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–7.

23. Harrell FE. Regression modeling strategies, 1st ed. New York: Springer-Verlag; 2001:600pp.

24. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277:488–94.

25. Moons KG, van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology 1999;10:276–81.

26. Weijnen CF, Numans ME, de Wit NJ, Smout AJ, Moons KG, Verheij TJ, et al. Testing for *Helicobacter pylori* in dyspeptic patients suspected of peptic ulcer disease in primary care: cross sectional study. BMJ 2001;323:71–5.

27. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, et al. Application of a diagnostic clinical model for the management of hospitalized patients with suspected deep-vein thrombosis. Thromb Haemost 1999;81:493–7.

28. Fraser AG, Ali MR, McCullough S, Yeates NJ, Haystead A. Diagnostic tests for *Helicobacter pylori*–can they help select patients for endoscopy? N Z Med J 1996;109:95–8.

29. Cowie MR, Struthers AD, Wood DA, Coats AJ, Thompson SG, Poole-Wilson PA, et al. Value of natriuretic peptides in assessment of patients with possible new heart failure in primary care. Lancet 1997;350:1349–53.

30. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. Am J Med 1998;104:374–80.

Karel G.M. Moons[*]
Cornee J. Biesheuvel
Diederick E. Grobbee

Julius Center for Health Sciences and Primary Care
University Medical Center
Utrecht, The Netherlands

[*] Address correspondence to this author at: Julius Center for Health Sciences and Primary Care, University Medical Center, P.O. Box 85500, 3508 GA Utrecht, The Netherlands. Fax 31-30-2505485; e-mail K.G.M.Moons@jc.azu.nl.

DOI: 10.1373/clinchem.2003.024752