 doi:10.1093/ije/dyi105

Consequences of different diagnostic 'gold standards' in test accuracy research: Carpal Tunnel Syndrome as an example

Lucas M Bachmann,[1,2] Peter Jüni,[1,3,4*] Stephan Reichenbach,[1,3,4] Hans-Rudolf Ziswiler,[3] Alfons G Kessels[2,5] and Esther Vögelin[6]

Accepted 20 April 2005

Test accuracy studies assume the existence of a well-defined illness definition and clear-cut diagnostic gold standards or reference standards. However, in clinical reality illness definitions may be vague or a mere description of a set of manifestations, mostly clinical signs and symptoms. This can lead to disagreements among experts about the correct classification of an illness and the adequate reference standard. Using data from a diagnostic accuracy study in carpal tunnel syndrome, we explored the impact of different definitions on the estimated test accuracy and found that estimated test performance characteristics varied considerably depending on the chosen reference standard. In situations without a clear-cut illness definition, randomized controlled trials may be preferable to test accuracy studies for the evaluation of a novel test. These studies do not determine the diagnostic accuracy, but the clinical impact of a novel test on patient management and outcome.

Keywords Sensitivity and specificity, ROC curve, reference standards, carpal tunnel syndrome, ultrasonography

The notion of a diagnostic gold standard or reference standard pertains to the best available method for establishing the presence or absence of a condition of interest,[1] i.e. the independent and correct classification of what is meant to be the illness.[2] The traditional concept of a reference standard depends on a high level of biological understanding of the target condition and its causal underlying mechanisms. Typically, a morphological verification such as histopathology or angiography, is used to establish a 'definite diagnosis'. This definite diagnosis is assumed to be a reasonably reliable proxy measure of the true presence or absence of the condition of interest.

In conventional diagnostic accuracy studies, the usefulness of a novel test for the inclusion or exclusion of a specific condition will be determined by comparing the results of the test with the definite diagnosis ascertained by the reference standard. However, in clinical reality the biological understanding of conditions is frequently unclear. Illness definitions are vague or a mere description of a set of manifestations. In fields such as psychiatry and rheumatology, clinicians frequently use 'syndromal diagnoses' consisting of a characteristic pattern of signs and symptoms,[3] while the biological understanding of the condition, of its causes, and its manifestations is incomplete and there is controversy about the manifestations that have to be combined to ensure accurate representation of the condition. In other situations, the biological understanding of the condition may be comprehensive, but the measurement of signs or symptoms is inaccurate.

Two extreme conceptualizations of the reference standard may implicitly or explicitly be used in such circumstances. One extreme ignores potential controversies and assumes a well-defined illness, which is objectively and reproducibly represented by the outcome of one or several laboratory tests. The other extreme ignores potentially useful biological measures and focuses exclusively on patient outcomes or on the need for an intervention. While these two outlooks aim at describing the same issue, they may create a schism when evaluating a diagnostic test. Below, we will explore this in a clinical example of an accuracy study previously published by our group in the field of rheumatology[4] and discuss the potential implications for clinical research into conditions without a clear-cut reference standard by which to establish a diagnosis.

[1] Department of Social and Preventive Medicine, University of Berne, Switzerland.

[2] Horten Centre, University of Zurich, Switzerland.

[3] Department of Rheumatology and Clinical Immunology, Inselspital University of Berne, Switzerland.

[4] MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, UK.

[5] Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Hospital, Maastricht, The Netherlands.

[6] Department of Hand Surgery, Inselspital, University of Berne, Switzerland.

* Corresponding author. E-mail: juni@ispm.unibe.ch

Clinical example

Carpal Tunnel Syndrome (CTS) is an important cause of functional impairment and pain of the hand, which presumably results from a compression of the median nerve at the wrist. Unfortunately, there is no universally accepted reference standard to establish the diagnosis. In our experience, two different approaches towards CTS classification are used. Neurologists traditionally establish the definite diagnosis based more on the outcome of nerve conduction studies than on the patients' signs and symptoms. In contrast, hand surgeons appear to give considerably more importance to the patients' signs and symptoms, the severity of complaints and the likely need for and success of a surgical intervention than to nerve conduction studies when establishing the definite diagnosis. In our accuracy study,[4] we relied on current practice and pre-specified the neurologists' definite diagnosis as the reference standard. Here, we determine the impact of using either of the two 'reference standards' on the estimated test accuracy of sonography in patients with suspected CTS.

Methods and results

Details of methods are reported elsewhere.[4] We assessed 77 patients for eligibility, excluded 3 because of traumatic wrist lesions, and enrolled 74 referred to the outpatient clinic of the Department of Hand Surgery at the University Hospital Berne, Switzerland, between January and December 2002.

Patients included in the study had a mean age of 51 years and 48 were females (65%). The flow of patients through the various stages of the study is described elsewhere.[4] Essentially, 101 wrists from 71 patients were included in the analysis.

Standardized nerve conduction studies were performed by one of several neurologists, who were unaware of the results of the sonographic examination. The sonographic evaluations were performed by a rheumatologist experienced in musculoskeletal sonography, who was unaware of the results of the nerve conduction studies and of the patients' signs and symptoms. He performed transverse imaging of the median nerve for the area ranging from the distal forearm to the outlet of the carpal tunnel and measured the largest cross-sectional area of the median nerve in square millimetres. We used this measure as a single diagnostic indicator, assuming that an increase in cross-sectional areas is associated with an increasing likelihood of disease or disease severity.

Table 1 presents a comparison of definite diagnoses according to neurologists' and hand surgeons' judgements. Overall agreement was 86%. One out of 23 wrists classified as normal by the neurologists was considered as CTS by the hand surgeons (4%). This wrist had normal nerve conduction studies. Conversely, 13 out of 78 wrists classified as CTS by the neurologists were considered normal by the hand surgeons (17%); all 13 wrists had pathological nerve conduction studies. The resulting kappa for the agreement between the two illness definitions was 0.67 [95% confidence interval (CI) 0.48–0.85].

For both reference standards, we fitted a receiver operating characteristic (ROC) curve for diagnosis of CTS by sonography, using a maximum likelihood logistic regression model based on robust standard errors, which allowed for the correlation of characteristics of wrists within patients and compared the area under the ROC curve. Figure 1 shows the fitted ROC curves using either the neurologists' judgements (top) or the hand surgeons' judgements (bottom) as the reference standard. The area under the ROC curve for ultrasound was 0.89 based on neurologists' judgements (95% CI 0.82–0.96) and 0.77 based on hand surgeons' judgements (95% CI 0.68–0.87). The difference between the two areas under the ROC curve was 0.12 (95% CI 0.0–0.23).

Table 1 2 × 2 contingency table comparing reference standard classifications according to neurologists and hand surgeons

	Hand surgeons' judgements		
	CTS present	CTS absent	Total
Neurologists' judgements			
CTS present	65	13	78
CTS absent	1	22	23
Total	66	35	101

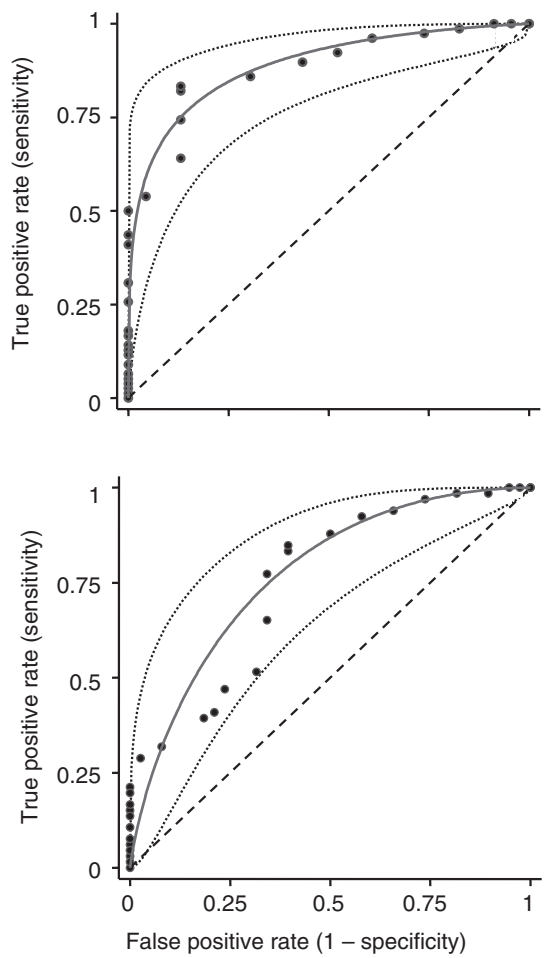

Figure 1 Fitted ROC curves (solid curve) for diagnosis of CTS by sonography with 95% confidence interval (dotted curves), considering the neurologists' definite diagnosis (top) or the hand surgeons' definite diagnosis as the reference standard (bottom). The broken diagonal line represents a hypothetical ROC curve of a test that yields no diagnostic information

Discussion

Even though the agreement between the two employed illness definitions was substantial (a kappa of 0.67), the estimated test performance of ultrasound varied considerably depending on the definition used as the reference standard. The diagnostic accuracy of sonography in patients with suspected CTS was good to excellent according to one reference standard but only moderate according to the other.

The lack of consensus on an illness definition may impede a valid evaluation of diagnostic technology in test accuracy studies. Considering that the final purpose of any novel test is to improve patient management and outcome, the traditional paradigm of test accuracy studies will only be useful if a reference standard is chosen that either has a strong association with patient outcome or a direct relationship with patient management. In our accuracy study[4] we argued, for example, that the neurologists' definite diagnosis directly pertains to clinical decision making and patient management.

Ultimately, the use of a diagnostic test and its potential therapeutic consequences can be considered as two consecutive steps of the same management strategy. Analogous to traditional research into therapeutic interventions, randomized trials may be designed to compare different strategies. In such trials, patients will be randomly allocated to a management strategy that includes the use of a novel test under evaluation, or to a strategy that uses standard tests only. Ascertained outcomes may relate to parameters of patient management (e.g. length of hospital stay), to patient outcome (e.g. pain), or to the total cost of management per patient.[5] If an unanimously accepted reference standard is lacking, as is the case in CTS, such randomized controlled trials may be more appropriate than test accuracy studies to determine the usefulness of a novel diagnostic test.

References

[1] 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.

[2] Wulff HR, Gotzsche PC. Diagnosis. In: *Rational diagnosis and treatment: evidence-based clinical decision making*. Third Edition. Oxford: Blackwell Publishing Ltd, 2000, p. 67.

[3] Wulff HR, Gotzsche PC. The disease classification. In: *Rational diagnosis and treatment: evidence-based clinical decision making*. Third Edition. Oxford: Blackwell Publishing Ltd, 2000, p. 39.

[4] Ziswiler HR, Reichenbach S, Vögelin E, Bachmann LM, Villiger PM, Jüni P. Diagnostic value of sonography in patients with suspected carpal tunnel syndrome: a prospective study. *Arthritis Rheum* 2005;**52:**304–11.

[5] Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;**356:**1844–47.