Editorial

Interpreting Studies of Diagnostic Accuracy

In this issue of *The Journal*, Payet, *et al* examine a test for elevated anticyclic citrullinated peptide antibody (anti-CCP) levels and demonstrate its inability to identify rheumatoid arthritis (RA) among anti-CCP-positive patients with rheumatic disorders[1]. Their results seem to be at odds with previous findings, which have shown the high diagnostic accuracy of anti-CCP tests for differential diagnosis of RA[2,3]. It is important, therefore, to consider how to appropriately interpret studies of diagnostic accuracy and assess generalizability. Such considerations are important when planning, reporting, or reading studies of diagnostic accuracy.

Pepe[4] lists the following 6 criteria for identifying settings where diagnostic tests would be useful: (1) the disease should be potentially serious, (2) the disease should be relatively prevalent in the target population, (3) the disease should be treatable, (4) the treatment should be available to those who test positive, (5) the test should not harm the individual, and (6) the test should accurately classify diseased and non-diseased individuals. Given that RA is a chronic disease (with a worldwide prevalence of 1%[5]) that can lead to severe disability, premature mortality[6], and a loss of quality of life[7], and given that appropriate therapeutic intervention can greatly enhance clinical outcomes[6], it is clear that the first 4 criteria have been met in this setting. Anti-CCP antibody tests satisfy the fifth criterion, so it remains to establish that they can be used to accurately classify diseased and non-diseased individuals, which motivates studies of diagnostic accuracy such as that considered by Payet, *et al*[1]. Note that there is evidence that this sixth criterion could be met because anti-CCP tests have been shown to be useful in identifying patients with early-stage RA[8] and predicting which patients will progress from undifferentiated arthritis to RA[3,5,9]. However, as we will discuss, it is important to avoid extrapolating diagnostic study results beyond the particular use of the test under study.

Studies of diagnostic accuracy involve evaluating the ability of a novel index test to detect a target condition whose true status is determined through the use of a reference standard[10]. The agreement between a binary index test and reference standard can be summarized in a 2 × 2 table and can be used to derive a number of different measures of diagnostic accuracy including sensitivity and specificity, positive and negative predictive values, and positive and negative diagnostic likelihood ratios (DLR), as demonstrated in Table 1. These measures summarize different aspects of the test results: sensitivity and specificity summarize the degree to which the test reflects disease status, and the predictive values summarize the likelihood of disease given the test result, while the DLR reflect the degree to which test results affect a potential diagnosis[4].

Studies of diagnostic accuracy have important implications in patient care, but the design and reporting of such studies have often been less than ideal[10]. The Standards for Reporting of Diagnostic Accuracy (STARD) initiative resulted in a broad set of guidelines for the reporting of studies of diagnostic accuracy[10,11]; these STARD guidelines allow readers to assess the generalizability of study results. This editorial will focus on only 2 specific potential sources of error in generalizing results of diagnostic studies: spectrum bias and imperfect reference standard bias.

Spectrum bias occurs when attempting to extrapolate to a population that is different from the sample in terms of patient characteristics[4]. It is well known, for example, that changes in disease prevalence will directly affect the predictive values of a diagnostic test. Consider the test for elevated anti-CCP whose diagnostic accuracy for RA among anti-CCP-positive patients with rheumatic disorders is summarized in Table 1; if this same test were applied as a screening tool in the general population where the prevalence of RA is 1%, then even if the test retained the same sensitivity, specificity, and DLR, the positive predictive value would drop to 0.0107 from 0.83, while the negative

See Anti-CCP in rheumatic disorders, *page 2395*

The Journal of Rheumatology 2014; 41:12; doi:10.3899/jrheum.141109

Table 1. Agreement between a test for high levels of anticyclic citrullinated peptide antibodies (Y) and diagnosis of RA (R) in Payet, *et al*[1].

		R	
		+	–
Y	+	249	50
	–	43	13

Sensitivity = 249/(249 + 43) = 0.85, specificity = 13/(13 + 50) = 0.21, positive predictive value = 249/(249 + 50) = 0.83, negative predictive value = 13/(13 + 43) = 0.23, positive DLR = sensitivity/(1-specificity) = 0.85/(1–0.21) = 1.07, and negative DLR = (1-sensitivity)/specificity = (1–0.85)/0.21 = 0.71. DLR: diagnostic likelihood ratios; RA: rheumatoid arthritis.

Table 2. Agreement between a test for high levels of anticyclic citrullinated peptide antibody (Y) and a hypothetical true diagnosis of RA (D), given that R in Table 1 is an imperfect reference standard.

		D	
		+	–
Y	+	289	10
	–	43	13

Sensitivity = 289/(289 + 43) = 0.87, specificity = 13/(13 + 10) = 0.57, positive predictive value = 289/(289 + 10) = 0.97, negative predictive value = 13/(13 + 43) = 0.23, positive DLR = sensitivity/(1-specificity) = 0.87/(1–0.57) = 2.00, and negative DLR = (1-sensitivity)/specificity = (1–0.87)/0.57 = 0.23. DLR: diagnostic likelihood ratios; RA: rheumatoid arthritis.

predictive value would rise to 0.9928 from 0.23. This calculation of posttest probability of disease can be achieved by taking the product of the appropriate DLR and the pretest odds of disease (i.e., disease prevalence divided by 1 minus disease prevalence) to get the posttest odds of disease, which can then be transformed to probabilities (e.g., $1.07 \times 0.01 / 0.99 = 0.0108$, $0.0108 / (1 + 0.0108) = 0.0107$; $0.71 \times 0.01 / 0.99 = 0.0072$, $1–0.0072 / (1 + 0.0072) = 0.9928$). Fagan[12] presented a nomogram to graphically represent this Bayesian relationship. This calculation, however, relies on the tenuous assumption that DLR is the same in these 2 different situations[4]. While sensitivity, specificity, and DLR are not directly affected by changes in disease prevalence, such changes are often indicative of underlying differences in patient characteristics that will directly affect these measures of accuracy[13]. Therefore, studies should not be interpreted as assessing some absolute diagnostic accuracy of a test, but rather as assessing a particular use of a test in a particular setting[14]. Diagnostic tests that are effective for use in primary care, for example, may be useless in tertiary care settings[15]. This issue can potentially be mitigated through the use of regression modeling, which can help to control for important confounders and identify important subpopulations[4]; however, care should always be taken to avoid extrapolating beyond the population represented by the sample under study.

Imperfect reference standard bias occurs when the reference standard to which the index test is compared is not a perfect indicator of true disease status[4]. In this case, measures of diagnostic accuracy can be over- or underestimated, depending on the error inherent in the reference standard. Suppose, for example, that the reference standard R used in Table 1 is only able to identify late stage RA, and thus is an imperfect reference standard for the true diagnosis of RA, which we will call D. The true diagnostic accuracy of the test might actually be better represented by Table 2, where we have assumed that R, the reference summarized in Table 1, misdiagnosed 40 early-stage RA patients with elevated anti-CCP as having a non-rheumatoid rheumatic disorder. In this case, many of the true measures of diagnostic accuracy would be underestimated using the results from Table 1. Alternatively, rather than viewing R as an imperfect reference for D and the results of Table 1 as biased estimates of the results of Table 2, one might interpret these as representing different uses of the same test: in Table 1, we are summarizing the utility of our test for identifying patients with late-stage RA, while in Table 2 we are summarizing the utility of our test in diagnosing RA more generally. The appropriate interpretation of study results very much depends on the reference standard and patient spectrum included in the study; any attempt to extrapolate to other settings is likely to be problematic.

When reading the results of the study of Payet, *et al*[1], as with any study of diagnostic accuracy, it is very important to consider the study population and the reference standard when considering whether these results can be generalized to your setting. The utility of the test under study would be very different in a generally healthy population (anti-CCP distributions differ between patients with rheumatic disorders and the general population[5]), or even among all patients with non-rheumatoid rheumatic disorders (according to the results of Payet, *et al*, specificity of a test for elevated anti-CCP would be as high as 93% if not restricting to the anti-CCP-positive group where specificity is only 21%). Additionally, it is important to note that Payet, *et al* used diagnoses of RA based on the American College of Rheumatology 1987 revised criteria[16] rather than the 2010 criteria[6] because of the latter's reliance on anti-CCP testing. This was done in an effort to avoid incorporation bias, which could have artificially inflated measures of diagnostic accuracy because of the lack of independence between the index and reference tests[17]. However, as acknowledged in their discussion, this approach potentially led to an under-identification of cases of early-stage RA[6], which is where anti-CCP testing is particularly useful[5]. It could, therefore, result in imperfect reference test bias if one attempted to extrapolate these conclusions as an assessment of the utility of the test for diagnosing early-stage RA or predicting RA development. Such conclusions should only be drawn based

on longitudinal studies that compare baseline test results to disease statuses measured at a later stage using time-dependent measures of diagnostic accuracy[4].

It is necessary to understand the setting in which a test was conducted, to avoid extrapolation biases. Such biases and misunderstandings can be mitigated if those conducting studies of diagnostic accuracy follow these 4 guidelines: (1) explicitly define the particular use of the test of interest, (2) carefully consider whether the population and the reference standard under study are consistent with this use, (3) use regression models to control for important concomitant factors when comparing tests, and (4) follow STARD guidelines in reporting results to ensure that readers can appropriately assess the generalizability of study results and examine potential sources of error.

MICHAEL A. McISAAC, PhD,
Assistant Professor,
Department of Public Health Sciences,
Queen's University,
Kingston, Ontario, Canada.

Address correspondence to Dr. M.A. McIsaac, Public Health Sciences, Queen's University, 99 University Ave., Kingston, Ontario K7L 3N6, Canada. E-mail: mcisaacm@queensu.ca

REFERENCES

1. Payet J, Goulvestre C, Bialé L, Avouac J, Wipff J, Job-Deslandre C, et al. Anticyclic citrullinated peptide antibodies in rheumatoid and nonrheumatoid rheumatic disorders: Experience with 1162 patients. J Rheumatol 2014;41:xxxx.
2. Kudo-Tanaka E, Ohshima S, Ishii M, Mima T, Matsushita M, Azuma N, et al. Autoantibodies to cyclic citrullinated peptide 2 (CCP2) are superior to other potential diagnostic biomarkers for predicting rheumatoid arthritis in early undifferentiated arthritis. Clin Rheumatol 2007;26:1627-33.
3. Taylor P, Gartemann J, Hsieh J, Creeden J. A systematic review of serum biomarkers anti-cyclic citrullinated peptide and rheumatoid factor as tests for rheumatoid arthritis. Autoimmune Dis 2011;2001:815038.
4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
5. Avouac J, Gossec L, Dougados M. Diagnostic and predictive value of anti-cyclic citrullinated protein antibodies in rheumatoid arthritis: a systematic literature review. Ann Rheum Dis 2006;65:845-51.
6. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheumatism 2010;62:2569-81.
7. Aggarwal R, Liao K, Nair R, Ringold S, Costenbander KH. Anti-citrullinated peptide antibody assays and their role in the diagnosis of rheumatoid arthritis. Arthritis Care Res 2009; 61:1472-83.
8. Takasaki Y, Yamanaka K, Takasaki C, Matsushita M, Yamada H, Nawata M, et al. Anticyclic citrullinated peptide antibodies in patients with mixed connective tissue disease. Mod Rheumatol 2004;14:367-75.
9. van der Linden MP, van der Woude D, Ioan-Facsinay A, Levarht EW, Stoeken-Rijsbergen G, Huizinga TW, et al. Value of anti-modified citrullinated vimentin and third-generation anti-cyclic citrullinated peptide compared with second-generation anti-cyclic citrullinated peptide and rheumatoid factor in predicting disease outcome in undifferentiated arthritis and rheumatoid arthritis. Arthritis Rheum 2009;60:2232-41.
10. 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. Ann Intern Med 2003;138:W1-12.
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. Clin Chem Lab Med 2003;41:68-73.
12. Fagan TJ. Nomogram for Bayes theorem [letter]. N Engl J Med 1975;293:257.
13. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185:E537-44.
14. Streiner DL. Diagnosing tests: using and misusing diagnostic and screening tests. J Pers Assess 2003;81:209-19.
15. Knottnerus JA, Buntinx F, editors. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2nd ed. Oxford: Wiley-Blackwell; 2009.
16. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries DF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
17. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": A note of caution. BMJ 2004;329:209-13.

J Rheumatol 2014;41:2340–2; doi:10.3899/jrheum.141109