Editorial

Are Reporting Standards of Diagnostic Test Evaluation Unrealistic?

The history of methods of diagnostic test evaluation not surprisingly includes authorities with an interest in rheumatology. One was Prof. Donald Mainland, a "plain language" medical statistician better known (by some) for the early work he undertook with the Cooperating Clinics Committee of the American Rheumatism Association in developing a composite index of disease activity in rheumatoid arthritis (RA)[1]. In *Elementary Medical Statistics*[2] (which I recommend to all young clinician scientists) Prof. Mainland set out questions for use in planning investigations and in evaluating their reports. In subsequent decades many excellent journal publications and texts offered advice on study design, conduct, analysis, and interpretation[3-6]. In 2003 the Statement for Reporting Studies of Diagnostic Accuracy (STARD) (http://www.stard-statement.org/) was published by several journals. STARD is a checklist of items for reporting studies of diagnostic accuracy[7].

Diagnostic test evaluation is particularly difficult when the test involves inconvenience for patient and researcher (time or cost) or when the test itself has intrinsic clinical (face) validity. These characteristics are clearly relevant to contemporary magnetic resonance imaging (MRI) research. In this issue of *The Journal*, Olech, *et al*[8] are to be congratulated for undertaking a large and blinded comparison of MRI abnormalities in healthy controls and in patients with established rheumatoid arthritis (RA). No previous study has both been blinded and contained sufficiently large numbers of healthy controls to comprehensively evaluate the performance of MRI. However, even this commendable study fails to meet some reporting standards set out by STARD and others.

Are standards set by medical statisticians and clinical epidemiologists unrealistic? Prof. Mainland's 1957 paper "Safety in Numbers," itself a wordplay on the numbers needed for statistical significance and the value of a collaborative process that involves clinicians and "biometricians," is instructive in this regard[9]. Even the best-intentioned clinicians occasionally fail as researchers because good clinicians are not necessarily skilled researchers. Although the standards may seem high, I do not believe they should be changed because they facilitate study validity and clinical interpretability.

Why do we as clinician scientists fail[10]? What design or analysis flaws still trip (or trap) the unwary? I will illustrate a few using the Olech report. Olech investigates whether an extremity 0.2-Tesla MRI unit (similar to the MRI unit marketed for use in a rheumatologist's office) can reliably discriminate MRI lesions of erosions, osteitis, and synovitis in wrist and metacarpophalangeal joints in patients with RA compared to healthy subjects. However, (1) MRI lesion definition and image acquisition were not the OMERACT RA MRI scoring system (RAMRIS) standard[7,11], (2) comparator subjects did not include patients with conditions that are easily confused with the disease of interest[8], (3) results did not include 95% confidence intervals[7], and (4) a more powerful method of establishing MRI utility, the likelihood ratio, was not reported[7]. The first 2 are difficult to remedy even with indirect analyses from the published literature. The latter 2, however, can be remedied if raw data are reported.

RAMRIS was not used by Olech for evaluating erosion and synovitis, therefore we cannot determine whether the poorer performance of MRI erosions and synovitis in this report was due to chance, to failure of published standards (lacking data on blindly evaluated healthy controls[12]), or to failure of the MRI scoring method chosen by Olech. RAMRIS requires that erosions be visible in at least 2 MRI planes and a cortical break be seen in at least one plane. This definition was data-driven. Observers scored coronal images alone, and then rescored the MR images based on coronal and axial planes. In 4 of 5 readers, including axial images in addition to coronal images changed the final scores[13]; therefore biplanar imaging was recommended. Lesions seen on only one plane were called bone defects; lesions seen in 2 planes were called bone erosions[14]. MRI is a 3-D imaging tool, and limiting its evaluation to one plane impedes its functionality. RAMRIS also recommends gadolinium contrast to facilitate assessment of synovitis[15-17]. Olech evaluated images in a single plane (coronal) and did not use contrast.

See Bone marrow edema is the most specific finding for RA on noncontrast MRI of hands and wrists, *page 265*

There is a tradeoff between standardization and innovation[18]. The developers of RAMRIS are an informal group of musculoskeletal radiologists, rheumatologists, and biometricians. They published standards for (1) definitions of lesions, (2) MRI acquisitions, and (3) scoring methods, after several years of calibration exercises. Although RAMRIS is unlikely to be the final standard on lesion definition, acquisition, or scoring, RAMRIS as a published standard facilitates the interpretation of innovation. Another difficulty in interpreting clinical applicability of new technologies arises from the need for 2 comparators: healthy controls (optimally, age and gender matched) as well as patients with conditions that can easily be confused with the disease of interest. An ambitious design would incorporate an additional 40 patients with peripheral joint pain but no obvious joint swelling. However, there are many potential populations deserving comparative study. Additionally, longitudinal designs add the important time dimension.

Olech, *et al* provide raw data that can be used to generate additional statistics. Bone marrow edema was the most specific MRI lesion for RA. However, bone edema was also the least sensitive, demonstrating the common tradeoff between sensitivity and specificity. Assuming sensitivity and specificity are of equal clinical utility, their average is a measure of accuracy[19]. The accuracies of bone erosion, bone edema, and synovitis are 62.5% (sensitivity 90%, specificity 35%), 73.75% (sensitivity 65%, specificity 82.5%), and 68.75% (sensitivity 80%, specificity 57.5%), respectively. The lower and upper 95% confidence limits of accuracy are 48% and 78%, 60% and 87%, and 54% and 83%, respectively. These results are not statistically different.

Because sensitivity and specificity are relatively robust to the prevalence of the condition being evaluated, they are the statistics of choice in the research setting for comparing different diagnostic tests in the same setting, or comparing the same diagnostic test in different settings. Sensitivity and specificity, positive and negative predictive values, and positive and negative likelihood ratios can all be calculated in the research setting. However, in clinical practice, the true disease status of patients is not known. That is why we do the test. In clinical practice, the diagnostic utility of the tests is best evaluated with the likelihood ratios rather than predictive values. Since likelihood ratios incorporate previously obtained sensitivity and specificity, they are relatively robust to the prevalence of the condition under evaluation, whereas the predictive values are not. The likelihood ratio is an arcane statistic calculated by sensitivity /1 – specificity: the bigger the value, the better the test.

The disadvantage of the likelihood ratio, and the reason why it is rarely reported, is that for practical use one needs to first apply it to the pre-test odds of disease to yield the post-test odds of disease (likelihood ratio × pre-test odds = post-test odds). Thus there is the requirement for an understanding of odds. Gamblers use odds; clinicians do not. In this study the likelihood ratio for bone edema is 3.7 and for bone erosion, 1.4. The pre-test odds of disease are the ratio of subjects with RA to subjects without RA, which was 1.0 (pre-test probability = 0.5) because there were equal numbers of healthy controls and patients with RA. The post-test odds of RA, given a finding of bone edema, are 3.7 to 1. These are better odds compared to the odds of bone erosion (calculated at 1.4 to 1). For clinicians, odds can then be converted back to probabilities [post-test probability = post-test odds/(1 + post-test odds); bone edema pre-test probability = 0.5 and post-test probability = 0.79]. Are these results clinically useful? Perhaps not. The pre-test odds of RA in "healthy" controls in the community are at most 0.01 (1%) and not 1.0 (50%). Also, healthy controls are not usually referred to rheumatologists and the accuracy of post-test odds is optimally determined by pre-test odds and likelihood ratio data obtained from research studies that reflect the clinical setting in which these results will be used.

Nevertheless, the Olech report performs well on STARD. The title/abstract/key terms identify that the report is a study of diagnostic accuracy (STARD criterion 1). It describes the study population (STARD 3) and participant recruitment criteria (STARD 4). Patient sampling was consecutive [although the healthy control population was not (STARD 5)], data collection was prospective (STARD 6), MRI technical specifications were described (STARD 8), rationale for cutoffs and categories provided (STARD 9), and number and expertise (but not training) of the scorers noted (STARD 10). MRI scoring was blind to subject status (STARD 11). Most methods for calculating measures of diagnostic accuracy (STARD 12) and reproducibility (STARD 13) were reported. The reference standard and rationale (STARD 7) are described (given limitations described above). Of the 11 STARD results reporting standards, 4 were clearly reported (STARD 15, 17, 22, 24), 3 were incompletely reported (STARD 18, 19, 21), and 4 were not reported (STARD 14, 16, 20, and 23). Finally, the clinical applicability of the results (Discussion: STARD 25) was not entirely straightforward, perhaps because the second research question (Introduction: STARD 2) could not be clearly answered.

In honor of the Charles Darwin bicentenary, we "...must begin with a good body of facts and not from principle (in which I always suspect some fallacy) and then as much deduction as you please"[20].

MARISSA LASSERE, MB, BS, Grad Dip Epi, PhD, FRACP, FAFPHM,
Associate Professor in Medicine,
Department of Rheumatology,
St. George Hospital,
University of New South Wales,
Sydney, New South Wales 2217, Australia

Address correspondence to Prof. Lassere;
E-mail: Marissa.lassere@sesiahs.health.nsw.gov.au

REFERENCES

1. Mainland D. The estimation of inflammatory activity in rheumatoid arthritis: role of composite indices. Arthritis Rheum 1967;10:71-6.
2. Mainland D. Elementary medical statistics. 2nd ed. Philadelphia: W.B. Saunders Company; 1963.
3. Guyatt GH, Tugwell PX, Feeny DH, Haynes BH, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587-94.
4. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: A basic science for clinical epidemiology. 2nd ed. Boston: Little, Brown and Company; 1991.
5. Jaeschke R, Guyatt GH, Sackett DL. Users' guide 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. Jaeschke R, Guyatt GH, Sackett DL. Users' guide to the medical literature III. How to use an article about a diagnostic test. What are the results and will they help me in caring for my patients? Are the results of the study valid. JAMA 1994;271:707-7.
7. 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.
8. Olech E, Crues JV, Yocum DE, Merrill JT. Bone marrow edema is the most specific finding for rheumatoid arthritis on noncontrast magnetic resonance imaging of hands and wrists: a comparison of patients with rheumatoid arthritis and healthy controls. J Rheumatol 2010;37:265-74.
9. Mainland D. Safety in numbers. Circulation 1957;16:784-90.
10. Lassere MN, Bird P. Measurements of rheumatoid arthritis disease activity and damage using magnetic resonance imaging. Truth and discrimination: does MRI make the grade? J Rheumatol 2001;28:1151-7.
11. Ostergaard M, Edmonds J, McQueen F, et al. An introduction to the EULAR-OMERACT rheumatoid arthritis MRI reference image atlas. Ann Rheum Dis 2005;64 Suppl 1:i3-7.
12. Ejbjerg B, Narvestad E, Rostrup E, Szkudlarek M, Jacobsen S, Thomsen HS, et al. Magnetic resonance imaging of wrist and finger joints in healthy subjects occasionally shows changes resembling erosions and synovitis as seen in rheumatoid arthritis. Arthritis Rheum 2004;50:1097-106.
13. Ostergaard M, Klarlund M, Lassere M, Conaghan P, Peterfy C, McQueen F, et al. Interreader agreement in the assessment of magnetic resonance images of rheumatoid arthritis wrist and finger joints — an international multicenter study. J Rheumatol 2001;28:1143-50.
14. Ostergaard M, Peterfy C, Conaghan P, McQueen F, Bird P, Ejbjerg B, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system. J Rheumatol 2003;30:1385-6.
15. McQueen F, Lassere M, Edmonds J, Conaghan P, Peterfy C, Bird P, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Summary of OMERACT 6 MR imaging module. J Rheumatol 2003;30:1387-92.
16. Bird P, Ejbjerg B, Lassere M, Ostergaard M, McQueen F, Peterfy C, et al. A multireader reliability study comparing conventional high-field magnetic resonance imaging with extremity low-field MRI in rheumatoid arthritis. J Rheumatol 2007;34:854-6.
17. Ostergaard M, Conaghan PG, O'Connor P, Szkudlarek M, Klarlund M, Emery P, et al. Reducing invasiveness, duration and costs of MRI in rheumatoid arthritis by omitting intravenous contrast injection — does it change the assessment of inflammatory and destructive joint changes? J Rheumatol 2009;36:1806-10.
18. Lassere MN. Imaging: the need for standardization. In: Cimmino MA, Grassi W, Cutolo M, editors. Imaging and musculoskeletal conditions. Best Pract Res Clin Rheumatol 2008;22:1001-18.
19. Last JM. A dictionary of epidemiology. New York: Oxford University Press; 1988.
20. Dainith J, Isaacs A. Medical quotations. London: Collins; 1989.

J Rheumatol 2010;37:220–2; doi:10.3899/jrheum.091254