Journal of Physiotherapy 67 (2021) 69–71

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Appraisal

Research Note: Diagnostic test accuracy studies

Introduction

In clinical physiotherapy practice, making a diagnosis is an essential part of patient management. The diagnostic process typically consists of history taking, physical examination and any additional investigations deemed necessary by the physiotherapist. A diagnosis enables the physiotherapist to: apply the appropriate evidence to decisions about possible treatments, address the patient's desire for information about their current condition and provide any evidence about the prognosis of people with that condition. A diagnosis therefore facilitates shared decision making by the physiotherapist and the patient about management.[1] It is therefore crucial for physiotherapists to be able to understand research into diagnostic tests and interpret the results of their diagnostic tests on patients.

Research describing the validity or accuracy of diagnostic methods gained prominence in the 1950s, particularly regarding the validity of psychological tests.[2,3] 'Validity' in this context refers to whether a test is able to measure what the clinician aims to measure (ie, the construct of interest). For a number of tests, the validity is obvious: if the physiotherapist wishes to find out how tall someone is, a measuring tape is clearly a valid instrument for this. However, determining the validity of many diagnostic tests relevant to physiotherapy is more complex (eg, tests to confirm whether a patient who presents with knee pain has a meniscal injury or not).

Diagnostic research aims to assess the validity of index tests by comparing them with a reference test. A reference test (previously called a 'gold standard' test) is considered the best available test for the condition of interest (target condition) but it may be difficult to administer, expensive or even invasive. Index tests are generally easier, less expensive and/or safer to administer in clinical practice. Most physiotherapy diagnostic research focuses on physical examination tests as index tests, such as McMurray's test for those with knee pain.[4] However, any element of the history (eg, presence of a cough) or a questionnaire (eg, Cumberland Ankle Instability Tool) can also be regarded as a diagnostic test and evaluated for its accuracy.[5] In addition, while most research focuses on single tests, a diagnostic accuracy study can also investigate combinations of findings, which may be more informative in clinical practice.

This Research Note describes how physiotherapists can interpret classic diagnostic research and explain the difference between classic diagnostic research and diagnostic models. This will help physiotherapists to consider which tests to use in clinical practice and why.

Classic diagnostic test accuracy studies

Scientific research into the value of a diagnostic process typically consists of evaluating the accuracy of individual diagnostic tests. These studies are called diagnostic test accuracy studies. The validity of the index test is assessed by comparing the outcomes of that test with those of a recognised and valid reference test.

In order to quantify the validity of an index test, the index test as well as the reference test commonly divide patients into two categories: test positive and test negative. A 2 × 2 table (also known as a four-field table or cross classification table) can then be created to depict how the study participants have been scored on the two tests (Figure 1). Based on such a table, a range of diagnostic statistics can be calculated, including sensitivity and specificity.

Some index tests or reference tests may be measured as continuous data (eg, age). Therefore, to calculate sensitivity and specificity the researcher needs to dichotomise the test results, which inevitably limits the interpretation of the test performance. The choice of a cut point (eg, age dichotomised at 40 years) is quite often arbitrary. To avoid this, the accuracy of the index test can be evaluated for multiple cut points or included as a continuous variable (with regression analysis).

Critical appraisal

Diagnostic test accuracy studies, like other studies, can be of high or low methodological quality. Tools such as the QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies 2)[6] can be used to rate the quality of classic diagnostic test accuracy studies. This tool

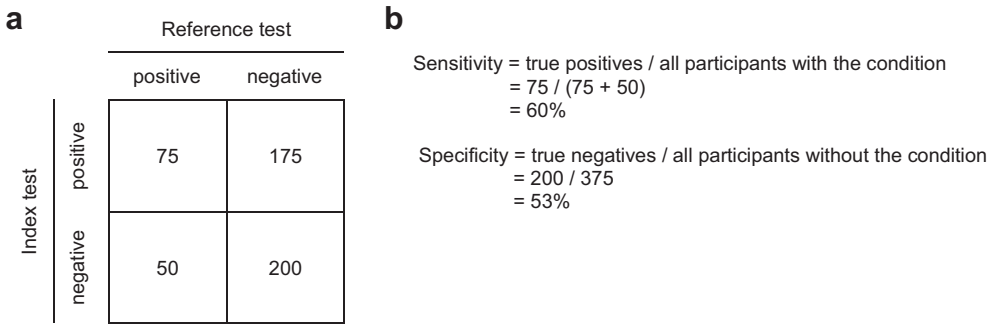

Figure 1. Elements of a classic diagnostic test accuracy study, including (a) a 2 × 2 table and (b) calculation of sensitivity and specificity.

https://doi.org/10.1016/j.jphys.2020.12.004

Box 1. Design elements that increase the believability of a classic diagnostic test accuracy study.

- A consecutive or random sample of patients was enrolled
- The study avoids inappropriate exclusions
- The index test and reference test are assessed independently (ie, without knowledge of the outcome of the other test)
- The threshold for a positive test is clear
- All participants received both tests
- The timing between the performance of the index and reference tests is clearly appropriate (ie, the result on either test is unlikely to change during that time)

comprises four domains: patient selection, index test, reference standard, and flow and timing. For a busy clinician it is helpful to consider that a diagnostic accuracy study will probably be trustworthy if it meets the criteria shown in Box 1.

Interpretation

Many people struggle to interpret diagnostic accuracy measures such as sensitivity and specificity. For example, what does a high degree of sensitivity, such as 92%, mean in practice? Correct interpretation involves both understanding what the terms (eg, specificity) and the numbers (eg, 80%) mean, and just as importantly understanding how the test is intended to be used in a clinical setting. Some tests are used primarily as a screening test, while others are used to make a diagnosis that directs treatment decisions. Screening tests such as the C-Spine rule (to identify possibility of cervical spine fracture) are used to rule out serious pathologies and determine if further testing is needed. Therefore, it is most important that screening tests do not miss people with the diagnosis (ie, screening tests should have a high sensitivity). However, it is less critical that when a screening test is positive the person actually has the diagnosis (ie, it is not crucial that screening tests also have high specificity). In an ideal world, tests would have high sensitivity and specificity. However, this is rarely the case, so the utility of a test for a specific purpose may depend on whether that particular usage requires high sensitivity or high specificity.

The terms SpPIn and SnNOut were introduced in the 1990s to help with the interpretation of sensitivity and specificity.[2,7] SpPIn stands for Specificity high and test Positive rules disease In; and SnNOut for Sensitivity high and test Negative rules disease Out. For instance, in people with low back pain with a suspected disc herniation the overall sensitivity of the straight leg raising test (SLR) is 92% (95% CI 87 to 95).[8] This means that almost all patients with low back pain and a disc herniation will have a positive SLR and very few patients with herniation will have a false negative result (approximately 8%). It can therefore be concluded with a high degree of sensitivity (eg, $>$ 90%) that patients with a negative outcome on the SLR test most likely do not have a herniation (SnNOut). Such an index test is therefore good at ruling out the disease or condition. Because tests rarely have both high sensitivity and high specificity, when a test has high sensitivity it often means that the test also produces many false positive results. The reverse applies to the crossed SLR test that has a high specificity (SpPIn): this test is good at ruling in the disease or condition,[8] but is not good at ruling out the condition (disc herniation).

It is not possible or sensible to provide exact sensitivity and specificity values that are considered 'high' (or at least high enough to comply with the SpPIn and SnNOut rules), since these values depend on the clinical consequences. In a person with low back pain the consequence of missing a diagnosis of acute cauda equina is very different to the consequence of missing a disc herniation, so higher sensitivity would be required in a test for cauda equina. For musculoskeletal disorders, values $\geq$ 90% with reasonably narrow confidence intervals are often considered sufficiently high for clinical utility.[7]

Diagnostic models

In daily clinical practice, the diagnostic strategy rarely, if ever, relies on a single clinical variable or test.[9] In principle, each diagnostic variable or test contributes towards the eventual likelihood that a patient has a particular disease or condition; in other words, the diagnostic process is a multivariable process.[8]

Diagnostic research is increasingly being conducted with the aim of ascertaining which combination of clinical variables or diagnostic tests will enable the clinician to make a diagnosis with a reasonable degree of certainty. Within that combination, how much value a particular test adds to the (final) diagnosis can also be investigated. For example, what is the additional value of a physical examination test (eg, Lachman's test) on top of history taking (eg, mechanism of injury, rapid effusion) for the diagnosis of an anterior cruciate ligament (ACL) rupture? To determine this, an analysis firstly focused on the likelihood of the diagnosis based on the combination of clinical variables from the history, and then added the results of physical examination tests to the analysis, to evaluate whether the likelihood of a diagnosis changed substantially.[10] The combination of three history items (a person has all three positive), each with an independent sensitivity and specificity varying from 43 to 73%, decreases the sensitivity to 18% and increases the specificity to 99%. Adding the anterior drawer test, the only physical examination test with promising results in this study, decreased the sensitivity to 16% and did not change the specificity (SpPIn).[10]

Diagnostic models are multivariable models, meaning that multiple variables (index tests) are evaluated for their collective association with the outcome of a reference standard (eg, ACL injury versus no ACL injury).

Critical appraisal

For diagnostic models it is also worthwhile to think whether the study could have introduced bias concerning: the selection of patients; the conduct or interpretation of the index test and the reference test; the flow of participants; and the order and timing of the index and reference tests. The same criteria of trustworthiness mentioned for single diagnostic tests hold for diagnostic models, although the analysis methods are much more important.

Interpretation

Diagnostic models commonly use a regression analysis, and the contribution of each individual variable or test is often expressed as a regression coefficient (beta) for a continuous measure, or an odds ratio (OR) for a dichotomous measure. For example, if the index test is a dichotomous measure (eg, sex) with an OR of 1.2 for males, this means that the likelihood of a disorder is slightly higher in males (someone with a positive test result for a given sex), compared with females (negative test result). If the index test is continuous (eg, age) with an OR of 1.2, this means that the likelihood of the disorder rises slightly with each year increase in the patient's age.

To evaluate the diagnostic value of a diagnostic model, a receiver operator characteristic (ROC) curve is presented. This is a graph depicting the sensitivity of the model against the specificity (technically, the false positivity rate, ie, 1 – specificity) at different cut-off points. The area under the curve (AUC) indicates how effectively the whole model can discriminate between people with and without the particular condition. A model with an AUC of 1 is a perfect model (ie, it can identify all patients with the condition without producing any false positives or false negatives). In contrast, a model with an AUC of 0.5 has no value whatsoever (ie, it cannot discriminate at all between people with or without the condition). With an AUC between 0.6 and 0.7 the model is considered 'reasonable' and with an AUC $\geq$ 0.7 the model is considered 'good';[11] however, as discussed above, the clinical use and consequences of the diagnostic model need to be considered.

The interpretation of beta coefficients, ORs or an AUC is rather impractical for clinicians. Therefore, a multivariable diagnostic model is ideally converted into a score chart or nomogram, which the clinician can easily use to calculate the likelihood of a certain diagnosis for a particular patient. For example, in patients with shoulder pain, a nomogram was developed that includes one history item (male sex) combined with three physical examination tests (positive lift-off test, Jobe test, and external rotation strength ratio between the affected and unaffected shoulder), which can predict a rotator cuff tear with 83% accuracy.[12]

Diagnostic models versus classic diagnostic test accuracy studies

Classic diagnostic test accuracy studies (ie, the evaluation of the accuracy of an individual test) often have limited direct applicability to daily practice. In daily practice, the diagnostic process almost always consists of a cluster of tests, and the two most important things for a clinician are: the likelihood of a diagnosis based on all tests; and what is the additional contribution of each test (ie, to determine whether it is worth adding that test to the battery of previous tests).[9] A further limitation of classic diagnostic research, compared with diagnostic models, is that it is more often carried out in a select patient group that is not representative of the patients seen in daily practice (selection bias) and is therefore likely to overestimate the accuracy (eg, sensitivity and specificity) of the index test.[13,14] This overestimation resulting from patient selection is most common when the reference test is costly or invasive (eg, a surgical procedure or magnetic resonance imaging).

There are two situations in which classic diagnostic test accuracy studies have a clear role. These are: the use of screening tests among healthy subjects (eg, breast cancer screening using a mammogram) and the evaluation of new tests. A screening test is often performed without the usual history taking and physical examination, with the primary aim of excluding people without the disease or condition from undergoing costly and intensive follow-up investigations (ie, ruling a condition out (SnNOut)).

With a new test, a stand-alone test should be reasonably able to discriminate between patients with and without the condition. If a test is unable to do so, further research into the use of the index test and its value as part of a battery of tests in a more realistic clinical situation is unlikely to be appropriate.

Diagnostic models also have some shortcomings. For instance, inadequate sample size can lead to an overestimation of the likelihood of a particular diagnosis.[15] Furthermore, it is only possible to include in the model those variables that have been measured. There are often considerable differences between studies as to which variables (index tests) were deemed relevant and included, so comparison between studies is difficult. For instance, a systematic review of diagnostic models identified four models for the diagnosis of symptomatic sacroiliac joints in people with low back pain, only evaluating the combination of physical examination tests.[16] Of these four clusters, the 'cluster of Laslett' is most well-known.[17] All these models indicated that three out of five specific tests must be positive, but the tests that were included in the analyses differed somewhat from one another; therefore, the clusters differed. The sensitivity of the clusters varied from 45 to 91%, and the specificity from 57 to 89%. These models used a more classic approach (using sensitivity and specificity) and looked at the number of tests instead of weighting each test for its contribution to the model. Ideally, models developed in one study are then validated in a new population, but examples of this are relatively rare.

Summary

This Research Note summarised and contrasted classic diagnostic test accuracy studies and diagnostic models. When interpreting results from a diagnostic study, it is important to clearly understand the purpose of the test(s) under investigation.[5] The role of a diagnostic test can be: to screen people to exclude the need to undergo further investigation (in this case a high sensitivity (SnNOut) is important); to make a diagnosis to guide prognostication or further treatment decisions; or just to monitor the condition.[18]

Currently, most diagnostic research focuses on the diagnostic accuracy of single tests, but diagnostic accuracy of a model based on multiple variables or tests better reflects clinical practice and may produce better results. A growing number of diagnostic models are being developed. Diagnostic models that are found to be adequately accurate can be translated into nomograms to enhance their clinical usefulness. Ultimately the value of any clinical diagnosis, based on single tests or a model including multiple tests, depends on whether the diagnosis improves prediction of important patient outcomes or helps direct treatment decisions.

Competing interests: Nil.
Sources of support: Nil.
Acknowledgements: Nil.
Provenance: Invited. Peer reviewed.
Correspondence: Arianne Verhagen, Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney, Sydney, Australia. Email: arianne.verhagen@uts.edu.au

Arianne Verhagen[a] and Mark Hancock[b]
[a]*Discipline of Physiotherapy, Graduate School of Health, University of Technology Sydney, Sydney, Australia*
[b]*Faculty of Medicine and Health Science, Macquarie University, Sydney, Australia*

References

1. Davidson M. *Aust J Physiother.* 2002;48:227–233.
2. Cronbach LJ, Gleser GC. *Psychol Bull.* 1953;50:456–473.
3. Cronbach LJ, Meehl PE. *Psychol Bull.* 1955;52:281–302.
4. Kaizik MA, et al. *Physiother Res Int.* 2020;25, e1871.
5. Cohen JF, et al. *BMJ Open.* 2016;6, e012799.
6. Whiting PF, et al. *Ann Intern Med.* 2011;155:529–536.
7. Hegedus EJ, Stern B. *J Manip Physiol Ther.* 2009;17:E1–E5.
8. Van der Windt DA, et al. *Cochrane Database Syst Rev.* 2010;2:CD007431.
9. Moons KGM, et al. *Clin Chem.* 2004;50:473–476.
10. Wagemakers HP, et al. *Arch Phys Med Rehabil.* 2010;91:1452–1459.
11. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* 2nd ed. New York: Wiley; 2000.
12. Jain NB, et al. *Orthop J Sports Med.* 2018;6:2325967118784897.
13. Mower WR. *Ann Emerg Med.* 1999;33:85–91.
14. Sackett DL, Haynes RB. *BMJ.* 2002;324:539–541.
15. Retel Helmrich IRA, et al. *J Physiother.* 2019;4:243–245.
16. Petersen T, et al. *Musculoskelet Disord.* 2017;18:188.
17. Laslett M. *J Man Manip Ther.* 2008;16:142–152.
18. Bossuyt PM, et al. *BMJ.* 2006;332:1089–1092.