Open Access Publication

Critical Evaluation of Diagnostic Review Article

by Michael Turlik, DPM[1]

The Foot and Ankle Online Journal 2 (12): 5

This is the third in a series of articles for podiatric physicians discussing the evaluation of a diagnostic article. The author contrasts and compares the critical evaluation of a diagnostic review article to a therapeutic review article utilizing two recent diagnostic publications from the foot and ankle literature.

Key words: Evidence-based medicine.

Accepted: November, 2009 **Published:** December, 2009

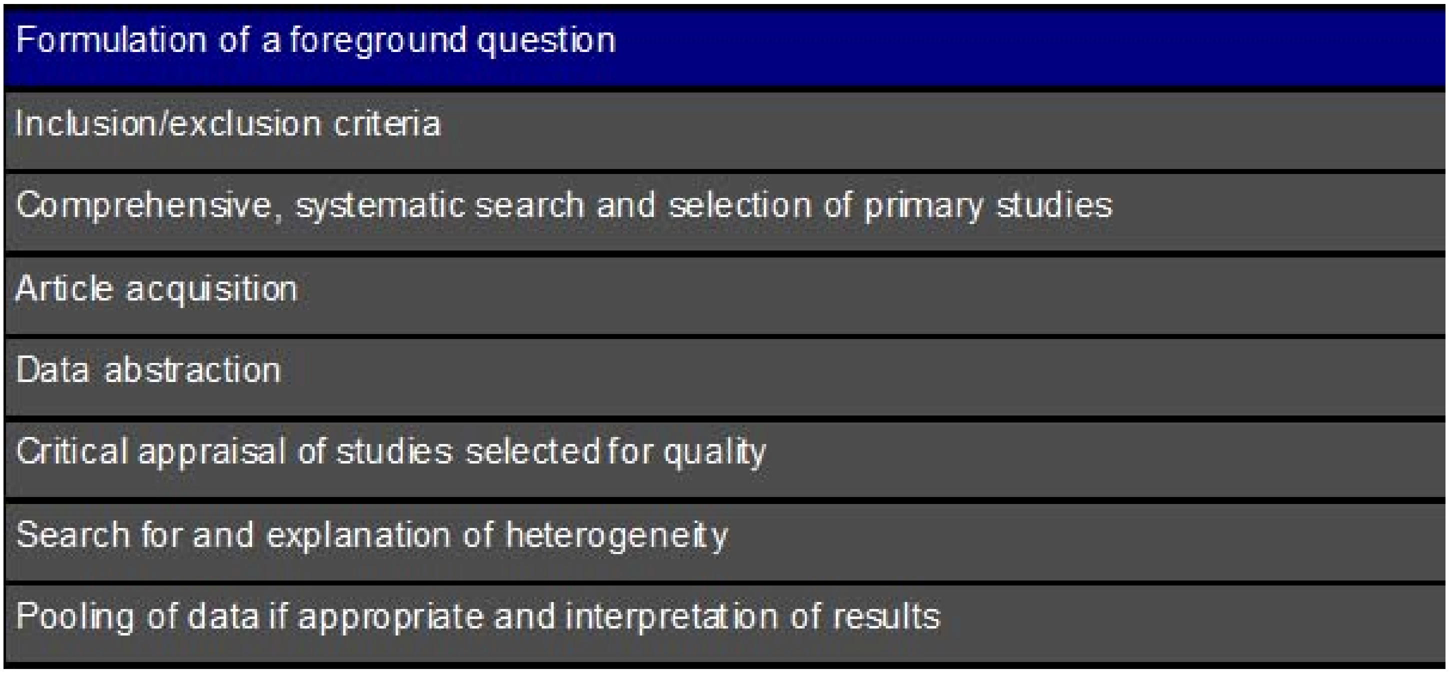
The treatment effect of a therapeutic intervention found in a randomized controlled trial (RCT) is enhanced when multiple studies can be combined quantitatively for a more precise estimate of effect. So too, can the diagnostic accuracy of an index test be enhanced if multiple studies can be pooled into a single diagnostic measurement. If the diagnostic meta-analysis is composed of cross sectional studies of an independent, masked comparison with a reference standard among an appropriate population of consecutive patients the meta-analysis can be considered level I evidence. Like therapeutic interventions this is accomplished through a systematic review/meta-analysis. When evaluating a systematic review/meta-analysis of diagnostic studies many of the same principles apply to the design, conduct and reporting of the study as discussed in an earlier article[1] describing the critical analysis of a systematic review/meta-analysis of therapeutic interventions. Steps in performing a systematic review / meta-analysis for diagnostic tests are the same (Table 1).

Although therapeutic and diagnostic meta-analysis shares many features, they differ in several important ways. The method to determine trial quality differs greatly between the two types of studies. In addition, the method by which heterogeneity and statistical pooling also differ significantly. The ways in which they differ will be the basis for this instructional monograph on how to critically evaluate a systematic review/meta-analysis of diagnostic studies. Two diagnostic systematic review/meta-analysis studies[2,3] will be compared and contrasted to illustrate the principles described in this instructional monograph. Both articles evaluate the use of diagnostic imaging techniques in the evaluation of osteomyelitis of the foot. As discussed in an earlier article[4] in this series the evaluation of infected diabetic foot ulcers often requires additional diagnostic studies to evaluate for the presence or absence of osteomyelitis. It is important for the podiatric physician to accept that the results of a systematic review/meta-analysis is a function of the validity of the methods used in the study and the quality of the primary studies included.

Address correspondence to: Michael Turlik, DPM
Email: mat@evidencebasedpodiatricmedicine.com

[1] Private practice, Macedonia, Ohio.

ISSN 1941-6806 doi: 10.3827/faoj.2009.0212.0005



Formulation of a foreground question
Inclusion/exclusion criteria
Comprehensive, systematic search and selection of primary studies
Article acquisition
Data abstraction
Critical appraisal of studies selected for quality
Search for and explanation of heterogeneity
Pooling of data if appropriate and interpretation of results

Table 1 Steps in systematic review/meta-analysis.

Formulation of the foreground question

The basics of question development have been covered elsewhere.[1] The foreground question for a diagnostic meta-analysis should define the population/disease of interest, the index test, the reference standard, and the outcome of interest.

Both articles[2,3] used as examples in this paper propose a foreground question to initiate the meta-analysis. In addition, Dinh[3] assesses clinical examination as well as, imaging modalities in the evaluation of foot osteomyelitis. Dinh[3] also more narrowly defines the population studied to those with infected diabetic foot ulcers.

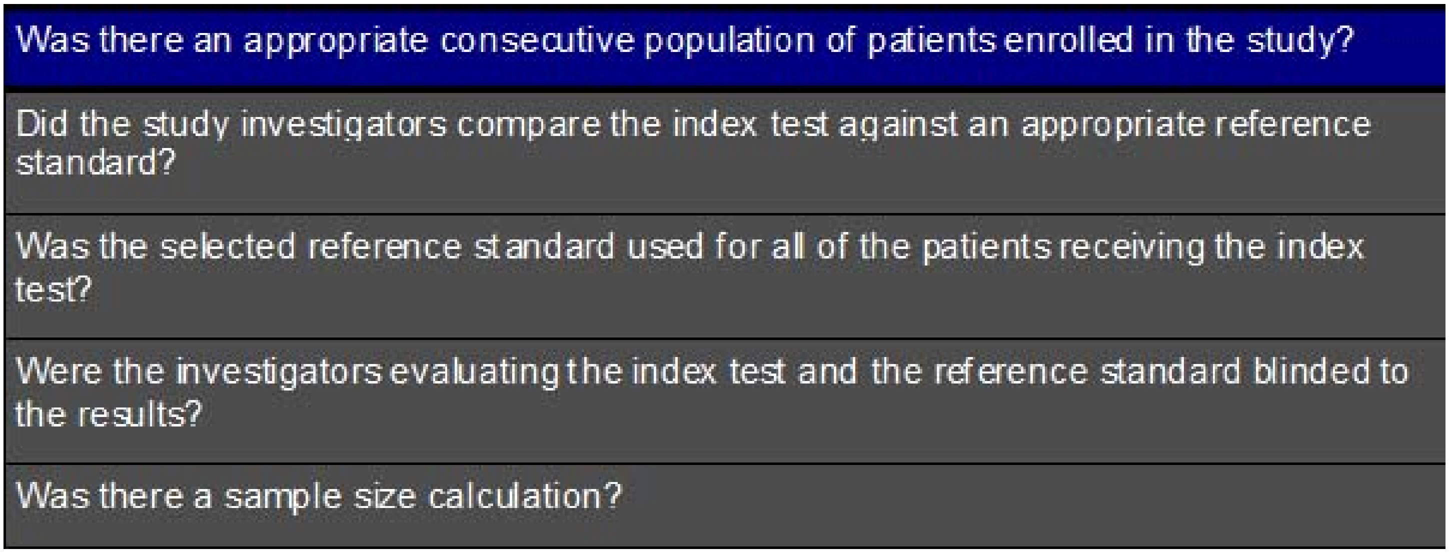
Inclusion/exclusion criteria

An explanation of inclusion and exclusion criteria are provided elsewherere.[1] Inclusion criteria for diagnostic studies should include the index test, reference standard, population studied, prevalence and outcome data to be abstracted.

Kapoor[2] clearly stated that the index test was magnetic resonance imaging (MRI), the comparators were bone scanning and radiographs, the population included all adults suspected of having osteomyelitis and data in the article need to contain information to construct a 2 x 2 diagnostic table. Dinh[3] described the inclusion criteria to consist of all diagnostic and clinical studies to evaluate diabetic foot ulcers for osteomyelitis. Both studies used bone biopsy as the reference standard. Neither described a particular setting for these studies nor was exclusionary criteria specified. However, Dinh[3] would supply exclusionary criteria if requested by the reader.

Comprehensive, systematic search and selection of primary studies

The basics of a systematic searching strategy have been covered earlier.[1] Standardized search terms are not as well-defined for diagnostic studies as they are for therapeutic studies. Special search strategies have been described for articles which evaluate diagnostic accuracy.[5]



Was there an appropriate consecutive population of patients enrolled in the study?
Did the study investigators compare the index test against an appropriate reference standard?
Was the selected reference standard used for all of the patients receiving the index test?
Were the investigators evaluating the index test and the reference standard blinded to the results?
Was there a sample size calculation?

Table 2 Evaluating internal validity of a diagnostic study.

Kapoor[2] described a comprehensive search strategy with additional details available upon request from the author. Dinh's[3] description of the search strategy used seemed less comprehensive than Kapoor's[2] description.

Article acquisition

The process by which the articles are selected from the search for review and data abstraction have been discussed earlier in the series.[1]

Kapoor[2] clearly described a process by which two authors reviewed each study for inclusion with a third author refereeing ties. Dinh's[3] explanation of article acquisition after the search was less clear than Kapoor's.[2]

Data abstraction

This has been covered in an earlier publication regarding the meta-analysis of therapeutic interventions.[1] One of the most important pieces of information to extract from a diagnostic study is the ability to generate a 2 x 2 table for each primary study.

Since diagnostic studies may be incompletely reported the authors of the meta-analysis may need to directly contact the authors of the primary study for additional information.[6]

Kapoor's[2] study described in detail the process by which the data was abstracted by two reviewers using a standardized form from the Cochran Collaboration. Resolution of disagreements between the independent reviewers was not described. Masked conditions were not used nor were inter rater agreements reported. Dinh's[3] description of the data abstraction process was less complete and consisted of a description of a standardized form which was used to abstract the data.

Critical appraisal of studies selected for quality

Diagnostic studies are highly variable with regards to quality and completeness in reporting.[6] The quality of the primary diagnostic accuracy studies published is in general; less mature than studies of therapeutic interventions. Evaluating the quality of a diagnostic study differs greatly from studies of therapeutic interventions. While no commonly accepted standardized method exists the second article of the series[4] describes the critical evaluation of a diagnostic study. The results of which are summarized in Table 2.

Unlike studies of therapeutic interventions often times the primary diagnostic studies may have incomplete information making studies difficult to include in the meta-analysis. If additional information cannot be obtained from the author the primary study results may not be included.

Kapoor[2] referenced the Cochran Methods Group checklist on Systematic Review of Screening and Diagnostic tests as an instrument which was used to assess study quality. Dinh[3] described and referenced an instrument to evaluate study quality.

Heterogeneity

Searching for heterogeneity

The choice of the data analytic method to combine study results is a function of the degree of heterogeneity found. This can be accomplished both graphically and statistically similar to a therapeutic study. A forest plot of study outcomes with 95% confidence intervals can be constructed to visually assess for heterogeneity similar to a therapeutic study.[7] A forest plot can be generated for measures of diagnostic accuracy to include: sensitivity, specificity, likelihood ratios, area under the curve derived from an ROC (receiver operator chacteristic) plot or diagnostic odds ratio.

ROC plots[8] are used in studies of diagnostic accuracy to demonstrate the pattern of sensitivities and specificities observed when the performance of a continuous test is evaluated at several different diagnostic thresholds. The overall diagnostic performance of a diagnostic test can be evaluated by the shape of the receiver operating characteristic curve. The curve is constructed using sensitivity plotted against one minus specificity. (Fig.1) The closer the curve passes to the upper left-hand corner of the plot the more accurate the test is. The area under the curve can be quantified as a measure of test accuracy. In order to be of any use the area under the curve has to be > 0.5. The closer the area under the curve approaches 1 the more accurate the test will be.

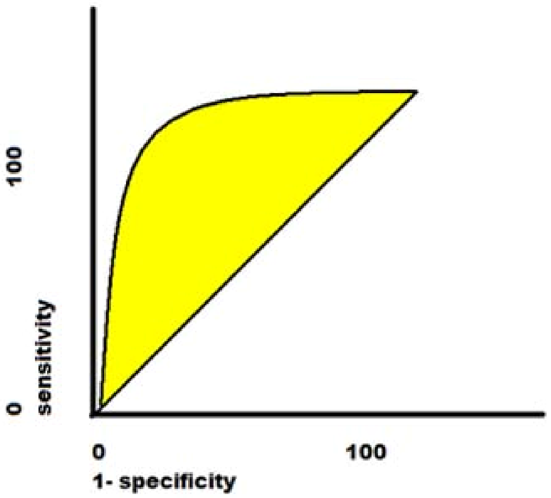


Figure 1 ROC plot.

The diagnostic odds ratio (DOR) is a summary measure of test performance.[9] Unlike other measures of diagnostic accuracy it can be expressed as a single number rather than a pair of numbers. DOR can be expressed from zero to infinity, the higher the number the better the test performance. DOR is derived by dividing the positive likelihood ratio by the negative likelihood ratio. It is a common measure of accuracy used in meta-analysis of diagnostic studies and is thought to be reasonably constant regardless of the diagnostic threshold. However, it is difficult to apply directly to clinical practice

Statistical tests to quantify heterogeneity in diagnostic studies have low statistical power. The most robust test for quantifying heterogeneity in diagnostic studies is a *Q*- statistic.[10] As discussed in an earlier paper, Cochran's *Q* is the traditional test for heterogeneity. It begins with the null hypothesis that the magnitude of the effect is the same across the entire study population. It generates a probability based upon the Chi squared distribution. The test is underpowered therefore; p > 0.1 indicates lack of heterogeneity.

Kapoor[2] in the comment section of the paper discussed the many flaws of the primary studies found in the meta-analysis. Because of small subset numbers the effect of design flaws in the primary studies could not be explored completely. Neither graphical methods nor statistical methods were used to explore for heterogeneity.

Dinh[3] used Cochran's *Q* to evaluate heterogeneity of the primary studies. The results were discussed and presented in table format. Graphical methods were not utilized to search for heterogeneity.

Causes of heterogeneity in diagnostic studies

Differences in study results can be explained by any combination of the following: quality of the study, study location, method of selection of the patients studied, variations in study population, reference standard used in study. As in therapeutic meta-analysis[10] a pre-planned meta-regression can be used to explore causes of heterogeneity found in diagnostic studies. Subgroup analysis is another method to explore heterogeneity. Caution should be used in interpreting the results if they are not planned in advance and should be used to generate rather than confirm hypotheses.

A special consideration when evaluating heterogeneity in diagnostic studies for meta-analysis is to consider the effect of the diagnostic threshold[11] used in the primary studies. Primary studies may use different cutoff points to determine positive and negative results and as a result may introduce heterogeneity in the meta-analysis. By changing the diagnostic threshold to increase sensitivity it results in a decrease in test specificity. This may be explicit as in the case of a continuous measure or implicit in the case of a dichotomous measure.

For example, in the probe to bone test for evaluating osteomyelitis in infected diabetic ulcers the investigators may use the same explicit threshold however, the investigators may differ in what they regard as the boundary between normal and abnormal when performing the test.

An explicit threshold for continuous measures may be derived from a ROC plot. Spearman's correlation coefficient can be calculated between specificity and sensitivity of all diagnostic studies included in the meta-analysis to determine if the heterogeneity of the study is due to differing diagnostic thresholds. If the results of the test are strongly negatively correlated heterogeneity of the studies is unlikely due to a threshold effect.

Kapoor[2] found that studies which did not use bone histology to exclude disease tended to have better test performance. In addition, studies which were published in 1988 or earlier demonstrated lower test performance. Neither study used meta-regression analysis or threshold analysis. It was unclear if Dinh[3] explored for the causes of heterogeneity of the primary studies.

Pooling of data

Combining data in a diagnostic data analysis is not as refined as therapeutic studies. There is no consensus on the best method to pool data in a meta-analysis for primary diagnostic studies.[10] Random and fixed models as described earlier for therapeutic interventions[11] can be used and have been used in meta-analysis of diagnostic studies.[12] Directly combining sensitivities and specificities, predictive values, likelihood ratios and diagnostic odds ratios have been reported. If threshold heterogeneity is detected directly combining measures of diagnostic accuracy should be avoided.

Unlike therapeutic interventions each diagnostic study provides a paired estimate of diagnostic accuracy (sensitivity and specificity). An alternate method of combining this type of data is to generate a summary ROC plot.[12,13] If significant threshold heterogeneity is detected a summary ROC plot is a better meta-analytic model. Unlike a traditional ROC plot the summary ROC plot uses the sensitivity and specificity (diagnostic odds ratio) obtained from each primary study as a data point in constructing the curve. A summary curve is obtained by fitting a regression model to the pairs of sensitivity and specificity (diagnostic odds ratio) from each individual study.

Using the summary ROC plot, a cutoff point can be determined and a global summary measure of ROC the Q* statistic can be determined. The Q* statistic derived can be used to compare the accuracy of different diagnostic studies.

Kapoor[2] used a summary ROC analysis as the method by which to pool the data from the primary studies. In addition, 13 different subsets were evaluated using this method to attempt to explain the heterogeneity of the primary studies.

Dinh[3] pooled diagnostic odds ratios, sensitivities and specificities using a random effects model. In addition, a summary ROC analysis was performed and the area under the curve (Q*) was determined as a measure of test performance.

Results

The results of the study should be presented as point estimates with 95% confidence intervals. Disease prevalence should be reported as a measure of central tendency with ranges. Ideally there should be a comparison in diagnostic accuracy between well done and poorly done primary studies. Finally, the authors should discuss the cost the study evaluated and recommended.

Kapoor[2] and Dinh[3] concluded that MRI was superior in evaluating foot osteomyelitis compared to pedal radiographs, Technetium 99 phosphate bone scans and white blood cell scans. Kapoor[2] reported diagnostic odds ratios as point estimates with 95% confidence intervals. Dinh[3] reported odds ratios but only included 95% confidence intervals with reports of sensitivity and specificity. The overall diagnostic odds ratio for MRI in Kapoor's[2] study was 42.1. Dinh[3] reported overall diagnostic odds ratio for MRI as 24.36. In addition, Dinh[3] reported a summary measure of accuracy (Q*), as well as, pooled sensitivities and specificities which demonstrated the superiority of MRI in the diagnostic assessment of osteomyelitis in diabetic pedal ulcerations. The average prevalence of osteomyelitis in Kapoor's[2] study was 50% (32%-89%).

While the prevalence of osteomyelitis reported in Dinh's[3] study ranged from 12% - 86%. Kapoor's[2] study most likely was composed of hospitalized patients rather than outpatients.[4] Kapoor[2] found that studies prior to 1998 reported lower diagnostic performance likely due to poorer study designs. In addition, Kapoor[2] found that studies which did not use bone histology as a reference standard had higher diagnostic test performance. Both authors discussed the costs of MRI relative to the other imaging studies.

References

1. Turlik M: Evaluation of a review article. The Foot and Ankle Online Journal. 2: 2009.
2. Kapoor A, Page S, LaValley M, Gale D: Magnetic resonance imaging for diagnosing foot osteomyelitis. Archives Internal Medicine 167: 125 – 132, 2007.
3. Dinh M, Abad C, Safdar N: Diagnostic accuracy of the physical examination and imaging tests for osteomyelitis underlining diabetic foot ulcers: Meta-analysis. Clinical Infectious Diseases 47: 519 – 527, 2008.
4. Turlik M: How to evaluate an article on diagnosis for validity. The Foot and Ankle Online Journal 2: 2009.
5. Haynes R, Wilczynski N: Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: analytic survey. BMJ 328: 1040 – 1045, 2004.
6. Halligan S: Systematic reviews and meta-analysis of diagnostic tests. Clinical Radiology 60: 977 – 979, 2005.
7. Leeflang M, Deeks J, Gatsonis C, Bossuyt P: Systematic reviews of diagnostic test accuracy. Ann Intern Med.1 49: 889 – 897, 2008.
8. ROC curve http://www.anaesthetist.com/mnm/stats/roc/Findex.htm Accessed 05/08/2009.
9. Glas A, Lijmer J, Prins M, Bonsel G, Bossuyt P: The diagnostic odds ratio: a single indicator of test performance. J Clin Epid 56: 1129 – 1135, 2003.
10. Honest H, Khan K: Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Services Research 2: 4, 2004.
11. Turlik M. Evaluating the results of a systematic review/meta-analysis. The Foot and Ankle Online Journal. 2: 2009.
12. Devillé W, Buntinx F, Bouter L, Montori V, de Vet H, van der Windt D, Bezemer D: Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Medical Research Methodology 2: 9, 2002.
13. Gatsonis C, Paliwal P: Meta-analysis of diagnostic and screening test accuracy evaluations: Methodologic primer. AJR 187: 271 – 281, 2006.