The creation of the diagnostic accuracy quality scale (DAQS)

Chad Cook¹, Joshua Cleland², Eric Hegedus³, Alexis Wright³, Mark Hancock⁴

¹Walsh University, OH, USA, ²Franklin Pierce University, Rindge, NH, USA, ³High Point University, NC, USA,

⁴Macquarie University, NSW, Australia

Background: The utility of a dedicated clinical test is dependent on the diagnostic accuracy values and the quality of the study in which the test was examined. Scales allow a summative scoring of bias within a study. At present, there are no scales advocated to measure the bias of diagnostic accuracy studies.

Objective: The objective of this study was to create a new diagnostic accuracy quality scale (DAQS) that provides a quantitative summary of the methodological quality of studies evaluating clinical tests and measures.

Design: The study used a four-round Delphi survey designed to create, revise, and develop consensus for a quality scale.

Methods: The four-round Delphi involved a work team and a respondent group of experts. An initial round among the work team created a working document, which was then modified and revised, with opportunities to create new items threaded in the second round. Rounds III and IV involved voting on the importance of each of the proposed items and consensus development from the respondent group. Consensus for the selection of an item required a 75% approval for the importance of that item.

Results: Sixteen individuals with a variety of research/professional backgrounds made up the respondent group. Modification and revision of the initial work team instrument created a scale with 21 items that reflected potential areas of methodological bias.

Limitations: The new scale needs validation through weighted assessment. In addition, there was a large proportion of physical therapist/researchers on the work team and the respondent group.

Conclusions: Systematic reviews allow summation of evidence for clinical tests and scales are essential to critique the quality of the articles included in the review. The DAQS may serve this role for diagnostic accuracy studies.

Keywords: Diagnostic accuracy, Validity, Scale, Quality assessment

Background

A clinician's armory for diagnosis of patho-anatomical conditions includes dedicated clinical special tests which function as proxy measures for the targeted condition.¹ These clinical tests are important to orthopedic diagnosticians, and the diagnostic information gained from these tests is used to guide treatment decision-making and estimate prognosis.² Tests that have poor diagnostic accuracy are of little value during the diagnostic process. In turn, tests that have purportedly high diagnostic value, but which were ascertained in low quality studies that exhibit substantial bias, may artificially predispose decision-making and lead to increased decision errors.¹ With the number of clinical diagnostic tests and measures continuing to proliferate and with the recent importance placed on

accurate and appropriate diagnosis by selected professionals as first-point providers, it is necessary to thoroughly evaluate a test's diagnostic utility prior to incorporating it into clinical practice.^{3,4}

Two studies have previously identified a number of methodological factors that increase bias in a diagnostic accuracy study.^{3,5} Using a weighted linear regression analysis with weights proportional to the reciprocal of the variance of the log diagnostic odds ratio, Lijmer and colleagues³ indicated that: (i) use of a case-control design, (ii) use of different reference tests, (iii) lack of blinding of the index test when determining the reference standard, (iv) no description of the index test, and (v) no description of the population, were significantly associated with higher estimates of diagnostic accuracy values and inflated utility. Using a different methodology (meta-epidemiological regression approach) but a similar purpose, Rutjes and associates⁵ reported the effect of design methodological deficiencies on estimates of diagnostic

Correspondence to: Chad Cook, 2020 East Maple Street, Walsh University, USA. E-mail: ccook@walsh.edu

accuracy. Faulty design methodologies reaching statistical significance were: (i) use of a non-consecutive sample and (ii) retrospective data collection. A number of other design methodologies not only exhibited high relative diagnostic odds toward biasing the results (case-control design, incorporation bias, etc.) but also exhibited excessively wide confidence intervals, and were not statistically significant.

Recognition of the influence of bias on diagnostic accuracy results has prompted the creation of guidelines to improve the design and reporting of diagnostic accuracy studies.⁶ The standards for reporting of diagnostic accuracy (STARD)⁶ was designed to provide researchers with a checklist and flow diagram that should be used *prospectively* for optimal study design. The flow diagram outlines a method for patient recruitment, the order of test execution, the number of patients undergoing the index test, and the specific reference test selected. The STARD includes 25 items that are divided into five major topics and six subsequent subtopics.

In contrast to prospective guidelines such as STARD, quality checklists or *scales* exist in order to *retrospectively* rate design bias or overall study quality. Checklists and scales are wholly different tools. Checklists typically do not have a quantitative value or summary score affiliated with the tool.⁷ On the other hand, within a scale, the construct being considered is evaluated on a continuum, with quantitative units that reflect varying levels of a trait of characteristics.⁸ Scales have numeric scores attached to the tool along with overall summary scores that can be used by clinicians to evaluate the quality of the component being measured.⁸ Scales provide numeric ratings for the quality of the published work, thus allowing clinicians to carefully evaluate the source of the data values prior to adoption in clinical practice.²

A quality checklist that was developed for assessment of diagnostic accuracy studies is the original quality assessment of diagnostic accuracy studies (QUADAS)⁹ tool. The QUADAS tool was developed to assess quality by examining pertinent elements of internal and external validity within the study.⁹ The original QUADAS tool was developed through a four-round Delphi panel, which reduced 28 critical criteria for evaluation of a completed diagnostic accuracy study to 14 final components. After evaluation of the potential to use the QUADAS tool as a scale, the authors recommended its use only as a checklist with no summary scoring.¹⁰ The authors¹⁰ were unable to appropriately weigh each item to allow the scale to accurately represent the potential bias in a diagnostic accuracy study. In an attempt to improve on the weaknesses of the original QUADAS, the QUADAS-2¹¹ was developed in 2011.

The QUADAS-2¹¹ was created using a similar four-round Delphi format as the original QUADAS instrument. The QUADAS-2 document was designed to measure both internal bias and applicability, which the authors define as *the degree to which individual study criteria match the review question*. New to the QUADAS-2 is the use of signaling questions that allow an investigator to report the presence of bias or non-applicability within four key domains.¹¹ Key foci of the four domains include patient selection, index tests, reference standards, and flow and timing. There are 11 specific questions within the tool. Each domain has two qualitative summative questions that assess risks of bias, and the matching with the review question. As with the original QUADAS, the QUADAS-2 attempts to function as a qualitative assessment checklist for all forms of diagnostic accuracy; including investigation of measures that are not associated with clinical tests (e.g., imaging, laboratory test results, etc.).

Recently, Schueler and colleagues¹² expressed concerns about QUADAS-2. They questioned the value of the applicability measure, expressed concern about the fact that the QUADAS-2 is more time-consuming than the original instrument, and indicated that calculating inter-rater agreement only on the domain questions may be a limitation. Worth noting is that the creators of QUADAS-2 also found varying reliability when the checklist was piloted.¹¹ Our experience¹³ with the QUADAS-2 has identified some additional concerns. We also found a low level of agreement when scoring among seasoned raters,¹³ and feel the tool is not able to discriminate between a poorly designed and a strongly designed study, and that it has no obvious advantages compared to the original QUADAS. Both the QUADAS-2 and the original QUADAS, are purposively qualitative in nature and do not allow the ability to score a study using a numeric value, a quality that is fundamental to a scale.

It is our impression, based on a review of the literature that there is a need for a quantitative scale to assess the methodological quality of diagnostic accuracy studies. Scales are used for evaluating bias in randomized controlled trials and a number of other study design types. Although not supported by the creators,¹⁰ numerous studies^{14–18} have calculated summary scores when using the original QUADAS to evaluate the quality of diagnostic accuracy studies. Other studies¹⁹ have even weighted individual QUADAS items providing a summative score that they feel more accurately captures the bias elements of a study. The impetus behind summing to a single score is simple: scales provide an easy to use, quick to assess, numeric value that is understood by researchers, clinicians, and the lay population. Consequently, the objective of this study was to create a new diagnostic accuracy scale that provides a quantitative

summary of the methodological quality of studies evaluating clinical tests and measures. It is our hope that the new scale could be validated in the future as a quantitative measure of bias in studies of diagnostic accuracy for clinical tests.

Methods

Design

Our study used a four-round Delphi survey instrument that incorporated both a work team and a respondent group. A Delphi survey is a series of sequential questionnaires designed to distill and obtain the most reliable consensus from a group of experts.²⁰ This survey method is useful in situations where frequent clinical or practical judgments are encountered, yet there is a lack of published research to provide evidence-based decision making.²⁰

Subjects

Respondent group

Delphi participants were targeted if they were first (primary) or last (senior) authors on recently published diagnostic accuracy studies associated with clinical tests (between 2005 and 2012). We were particularly interested in authors who were involved in the research of diagnostic accuracy of clinical tests and measures, which was the targeted focus of the current quality scale. Search terms used when identifying authors included ‘diagnostic accuracy (tw)’, ‘validation (tw)’, and ‘clinical tests (tw)’. We collected 50 consecutive author names of appropriately identified studies that also provided author contact information in the manuscript. The search was performed on Pubmed by a single author (CC). An appropriate article qualified if the title and abstract provided an instance in which a clinical test was validated in a diagnostic accuracy design.

Work team

Because a portion of the creation of a consensus document required qualitative work, a work team was necessary to collate ideas and themes. The work team of this study consisted of the authors of this manuscript. All authors were experienced in publishing research on diagnostic accuracy. A majority was well-experienced in evaluating quality of diagnostic accuracy studies either in textbook or systematic review formats and a minority was experienced with Delphi studies. All the authors were also physical therapists.

Procedure

This Delphi survey consisted of four rounds of questionnaires that respondents consecutively answered.^{21,22} The first round was completed by the work team. The work team developed an initial diagnostic accuracy quality scale (DAQS) following the first three procedures outlined by Streiner and Norman²³ for development of quality assessment tools. Their process outlines

five critical steps, as follows: (i) preliminary conceptual decisions; (ii) item generation; (iii) assessment of face validity; (iv) field trials; and (v) generation of a refined instrument, allowing for the generation, refinement, and potential adaptation of a tool created during a continuous improvement process.

During Rounds II–IV, the respondent group was responsible for modifying the tool that was generated in the first round. Invitations to the respondents for Round II of the study were distributed through e-mail, each providing a direct web link with a web-based consent form and survey. Potential respondents who did not answer the request for participation were sent a reminder to encourage participation using a method suggested by Dillman.²⁴ Two consecutive follow-up reminders were delivered at 10 and 20 days after the initial invitation was sent.^{25,26} Demographics/background information was also captured during Round II. The *Cherries*²⁷ guideline for confidentiality of subjects was followed and the identification of each Delphi participant remained confidential throughout the study.

Assessment of face validity and consensus assessment

As previously stated, Round I involved the identification of potential items from the work team. The team created the document through open-ended suggestions and consensus. Items were created from three primary foci: (i) design bias, (ii) transferability to practice, and (iii) methodological rigor. As consensus-based adjustments occurred, two primary foci, (i) design bias and (ii) transferability to practice, were the principal themes. It was apparent that the items associated with methodological rigor effectively fell within the design bias and transferability to clinical practice categories. Initially, the work team generated a 19-item quality scale. This scale was presented to the respondents for Round II.

Round II involved acceptance, modification, and suggestion of new items from the respondent group. Each participant was provided the same instructions and description of the study’s purpose; which was stated to: *create a new scale that provides clinicians with an easy to use, quick to assess, numeric value that is understood by researchers, clinicians, and the lay population*. Within this round, each respondent group member was allowed to comment on each quality item and make recommendations for removal, addition, modification, or acceptance of the item. Upon completion by the respondent group, literal coding (by the work team) was used for regeneration of items from Round I. After Round II, there were 21 items within the quality scale.

Round III involved scoring the importance of the items created in the literal coding phase of Round II. Delphi participants scored each item as ‘strongly agree’, ‘agree’, ‘disagree’, or ‘strongly disagree’, and

provided comments for each item. Upon completion, scores were tabulated and placed in a bar graph for each item. Round IV involved the respondent group re-scoring items (upon seeing the bar graph scores of the others from Round III respondents). Again, Delphi participants scored each item as 'strongly agree', 'agree', 'disagree', or 'strongly disagree'. At the end of Round IV, items that were scored as 'strongly agree' or 'agree' by 75% or higher of the participants were retained in the DAQS.

Sample size

There is no critical sample size for a Delphi Study;²⁸ however, based on previous experiences we targeted 50 authors in order to capture an expected 15 respondents for the Delphi study. We concluded that 15 respondents would be the ideal number for processing within our electronic medium. Although our process for creating the DAQS was different from the process used for creating the QUADAS (we have involved both a work team and a respondent group of experts, whereas the QUADAS involved only a work team), we had similar numbers to those of the original QUADAS⁹ (created by a Delphi team of nine individuals) and the QUADAS-2¹¹ (generated by 24 individuals).

Results

There were 16 respondents (response rate of 32%) to the Delphi survey and all 16 completed all three designated rounds (100%). There were 11 males and

5 females and the mean age was 43.6 (SD = 8.5; range 29–62 years). The multinational group of respondents was from the Netherlands (2), Ireland (2), the United Kingdom (2), Turkey (1), Canada (1), Australia (1), New Zealand (1), and the United States (6). Primary occupations included university lecturer/researcher (1), physical therapist/researcher (5), epidemiologist (2), medical physician/researcher (2), athletic trainer/researcher (2), chiropractor/researcher (1), and researcher/research fellow (3), with 14/16 (87.5%) indicating an academic appointment of some nature. Eleven (11) of the 16 specialized in orthopedics-related research, whereas others identified dentistry (1), rheumatology (1), physical medicine and rehabilitation (1), and primary care medicine (1). Mean percentage of time devoted exclusively to research was 47.1% (SD = 26.7). Those with clinical experience within their professional fields had practiced on an average of 20.6 years (SD = 9.2; range 8–41) with a mean of 38.1 publications (SD = 45.6; range 2–150).

Initially, the work team had created a 29-item list that was further reduced (after combined concepts) to 19 items. The work team's 19-item scale increased to a 21-item scale in Round II through contributions from the respondent group (Table 1). The expansion from 19 to 21 item in the scale reflected the need to split two items that queried two different concepts in a single question. For example, the work team's item 'all patients in the study were tested with the same,

Table 1 Diagnostic accuracy quality scale (DAQS)

Question	Yes	No	Unclear
Question 1: The study incorporated a consecutive enrollment or random subject sampling.			
Question 2: The spectrum of the full sample was sufficiently described, including a severity measure (e.g., functional measure).			
Question 3: The study included an adequate sample size derived from a power analysis.			
Question 4: The study avoided a case-control design; a design in which subjects with obvious pathology were compared to healthy controls.			
Question 5: The study examined utility of the index test in a situation of diagnostic uncertainty.			
Question 6: Inclusion/exclusion criteria were clearly stated.			
Question 7: All patients in the study were tested with the same index test.			
Question 8: All patients in the study were tested with an adequately-described index test.			
Question 9: A positive and negative index test finding was adequately described.			
Question 10: Within the study, or from past studies, the index test demonstrated sufficient (inter-tester and intra-tester) reliability.			
Question 11: Examiners performing and interpreting the results of the index test were blinded to outcomes of the reference test.			
Question 12: The time interval between the index test and reference test was sufficiently short so that meaningful changes in the patient's clinical profile were not likely.			
Question 13: All patients in the study were tested with the same reference standard.			
Question 14: All patients in the study were tested with an adequately-described reference standard.			
Question 15: The reference standard was appropriate for the pathology examined.			
Question 16: Examiners responsible for performing and interpreting the results of the reference test were blinded to the outcomes of the index test.			
Question 17: The 2 × 2 table information was clearly presented and easily discernible, so that the reader can be confident that all patients received the index test.			
Question 18: Confidence intervals of the diagnostic accuracy values were provided.			
Question 19: Reasons for missing data and the manner in which the missing data were managed in the reduction process and statistical analysis were identified.			
Question 20: The clinical practice setting and or data collection site were adequately described.			
Question 21: The experience and qualifications of those responsible for performing and interpreting the index and reference tests were adequately described.			
Total score			

adequately-described reference standard' was split to capture the concept of the 'same' reference standard and 'adequate description'. During Rounds III and IV, there was very little discord among respondents when voting for the importance of each item. In both rounds, all 21 items reached the 75% cut off for accepting the item.

Discussion

The objective of this study was to create a new diagnostic accuracy scale that provides a quantitative summary of the methodological quality of studies evaluating clinical tests and measures. Our DAQS that was created within the four-round Delphi process involves 21 disparate items that are intended to discriminate studies that are of higher and lower quality. Further, the study focus was on clinical tests and measures only and each item was generated with this conception in mind. Low quality studies may yield diagnostic accuracy values of tests and measures that do not accurately reflect their utility in clinical practice, thus recognition of important biases is essential for the practicing clinician, researcher, and policy maker.

The initiative behind the development of the DAQS was our concern with the QUADAS-2. It appears that the changes made to QUADAS-2 were designed to allow multiple types of studies to be qualitatively represented. In our experience,¹³ the qualitative nature of the QUADAS-2 fails to discriminate between low and high quality papers and may not be effective for clinicians who attempt to differentiate disparate values from competing studies. Further, we feel the tool lacks reliability. As identified by Schueler and colleagues,¹² the type of study targeted during a review would eliminate the need for the applicability questions, a new key feature of the QUADAS-2. We feel some of the new additions of the QUADAS-2 were simply not needed and did not improve the original QUADAS tool.

Our four-round Delphi survey identified 21 items that the work team and the respondent group felt were important for evaluating diagnostic accuracy studies involving clinical tests and measures only. Many of the items are similar to the two QUADAS tools. Nine of the original 14 QUADAS items (64%) were represented within the DAQS, whereas 10 of the 11 domain questions (90.9%) in the QUADAS-2 were identified within the DAQS. The work team and the respondent group elected not to include '*did the study avoid inappropriate exclusions*' a question from the risk of bias domain of the QUADAS-2.¹¹ It is our interpretation that this item would be difficult to fully investigate quantitatively and may be at risk for poor inter-rater agreement. Two of the three items from the original QUADAS (Items 12, 13, and 14) which dealt with same clinical data availability, uninterpretable, indeterminable,

and intermediate test results, and study withdrawals were included within the DAQS; although the language used to assess this element was changed for improved understanding by the reader.

Throughout the previous literature, seven biases were statistically associated with increased bias in diagnostic accuracy studies. The biases include: (i) use of a case-control design,³ (ii) use of different reference tests,³ (iii) lack of blinding of the index test when determining the reference standard,³ (iv) no description of the index test,³ (v) no description of the population,³ (vi) non-consecutive sample,⁵ and (vii) retrospective data collection⁵ were statistically associated with increased bias in diagnostic accuracy studies. The DAQS includes specific items for six of the seven components represented within the literature, failing only to recognize the use of retrospective data as a risk for bias. In contrast, the QUADAS-2 included only *three* of the seven pre-identified areas of bias. Notably missing from the QUADAS-2 are domain queries associated with the use of different reference tests, description of the index test (thresholds are discussed but description is not), description of the population, and as in our instrument, retrospective data collection. We feel that the number of omissions in the QUADAS-2 are concerning and lend further questions toward the utility of the QUADAS-2.

Unique to the DAQS is the inclusion of reporting of confidence intervals, qualifications/experience of the raters, description of clinical or data collection setting, and need of a power analysis to derive sample size. These elements are not represented in the original QUADAS⁹ or QUADAS-2.¹¹ Confidence intervals would provide an estimate of precision for the data values and are generally required for publication in most journals.²⁹ Qualifications/expertise of the raters and description of clinical or data collection setting gives the context on who performed the clinical examinations, in what environment, and lends credence toward transferability of the findings. Although powering a diagnostic accuracy study is controversial,³⁰ a required item may prompt raters toward assessing globally underpowered studies, a historical problem that has been recognized in diagnostic accuracy studies.³¹ Lastly, comparable to selected items in the original QUADAS, a report of reasons for missing data was selected by the work team and the respondent group. This item and the necessity of providing a 2 × 2 table with true and false positives and true and false negatives would further explain if selected participants were removed from accuracy calculations. The QUADAS-2 has a domain question that states 'were all patients included in the analysis'. We feel that the 2 × 2 table and a required report of reasons for missing data would provide transferable information to truly analyze if all patients were included in the analysis.

One notable element of the DAQS is that it was created with the understanding that it would be used exclusively for assessment of bias in diagnostic accuracy studies of clinical tests and measures, only. The original QUADAS and QUADAS-2 were created to assess diagnostic accuracy of all forms of testing, such as laboratory and diagnostic imaging tests. We feel that these forms of diagnostic assessment are quite different than those related to clinical testing. For example, variation in the performance of an index test is much greater in a diverse clinical setting than those represented in laboratory settings. Reliability for interpretation may be an issue as well in clinical testing, since the threshold of the finding may depend greatly on the test interpreter.

We feel the face validation and consensus development of the DAQS has created a logical, easy-to-use scale for assessing bias in diagnostic accuracy of clinical tests and measures. The items within the scale are scored similar to the original QUADAS in that the results are either *present*, *absent*, or *unclear*. When an item is *unclear* it should be scored as *absent* for conservative assessment. We also recognize the need to further investigate the validity of this scale. To exist as a valid scale, a determination of the weightage of the items, similar to the processes used by Lijmer and colleagues³ may further condense the items to those that are discriminative for bias. Using meta-analytic measures such as those used by Lijmer and colleagues³ and Rutjes *et al.*,⁵ is one of a number of possibilities, as are meta-analytic methods that measure test results that provide extreme variations in comparison to majority findings. Nonetheless, we feel the creation of this tool, which has the potential to be used as a summative scale, will meet a substantial need for clinicians and fill a void that was left by the QUADAS-2.

Limitations

This study has a number of limitations. As stated, prior to use as a valid scale, the tool needs validation, most appropriately through weighted assessment. It is our goal to investigate weighted assessment in future research. Certainly, a 21-item scale will take more time to evaluate than a 14-item checklist and even potentially, the QUADAS-2. Although the respondent group was multinational and very experienced, there was a high proportion of physical therapists/researchers. This fact combined with the make-up of the work team (which was wholly comprised of physical therapists) means that the scale was greatly influenced by one profession.

Conclusion

The sheer volume of dedicated clinical special tests^{32,33} and the rapid rate at which new tests are developed suggest the need for systematic review and meta-analysis of these tests. The requisite use of a

scale to critique the quality of the articles included in any review in order to produce clear, clinically useful information. Both the original QUADAS and the QUADAS-2 are not meant to fill this need, necessitating the development of a validated scale that provides a clear ranking of studies based on their quality in design and reporting. We have taken the first step to fill this void with the development of the DAQS.

References

- 1 Bossuyt PMM. The quality of reporting in diagnostic test research: getting better, still not optimal. *Clin Chem*. 2004;50:465–6.
- 2 Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: use of the STARD and QUADAS methodological quality assessment tools. *J Man Manip Ther*. 2007;15(2):93–102.
- 3 Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–3.
- 4 Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;6:9.
- 5 Rutjes A, Reitsma J, DiNisio M, Smidt N, van Rijn J, Bossuyt P. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174:469–76.
- 6 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Ann Intern Med*. 2003;138:40–4.
- 7 Clark L, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess*. 1995;7:309–19.
- 8 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clin Trials*. 1995;16:62–73.
- 9 Whiting P, Rutjes AV, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25.
- 10 Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol*. 2005;5:19.
- 11 Whiting P, Rutjes W, Westwood M, Mallet S, Deeks J, Reitsma J, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529–36.
- 12 Schueler S, Schuetz GM, Dewey M. The revised QUADAS-2 tool. *Ann Intern Med*. 2012;156:323.
- 13 Hegedus EJ, Goode AP, Cook CE, Michener L, Myer CA, Myer DM, *et al*. Which physical examination tests provide clinicians with the most value when examining the shoulder? Update of a systematic review with meta-analysis of individual tests. *Br J Sports Med*. 2012;46(14):964–78.
- 14 Hegedus EJ, Cook C, Hasselblad V, Goode A, McCrory DC. Physical examination tests for assessing a torn meniscus in the knee: a systematic review with meta-analysis. *J Orthop Sports Phys Ther*. 2007;37(9):541–50.
- 15 Hegedus EJ, Goode A, Campbell S, Morin A, Tamaddoni M, Moorman CT 3rd, *et al*. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. *Br J Sports Med*. 2008;42(2):80–92; discussion 92.
- 16 Tjissen M, van Cingel R, Willemssen L, de Visser E. Diagnostics of femoroacetabular impingement and labral pathology of the hip: a systematic review of the accuracy and validity of physical tests. *Arthroscopy*. 2012;28(6):860–71.
- 17 Reneker J, Paz J, Petrosino C, Cook C. Diagnostic accuracy of clinical tests and signs of temporomandibular joint disorders: a systematic review of the literature. *J Orthop Sports Phys Ther*. 2011;41(6):408–16.
- 18 Simpson R, Gemmell H. Accuracy of spinal orthopaedic tests: a systematic review. *Chiropr Osteopat*. 2006;14:26.

- 19 Alqarni A, Schneiders A, Hendrick P. Clinical tests to diagnose lumbar segmental instability: a systematic review. *J Orthop Sports Phys Ther.* 2011;41:130–40.
- 20 Stheeman SE, van't Hof MA, Mileman PA, van der Stelt PF. Use of the Delphi technique to develop standards for quality assessment in diagnostic radiology. *Community Dent Health.* 1995;12:194–9.
- 21 Binkley J, Finch E, Hall J, Black T, Gowland C. Diagnostic classification of patients with low back pain: report on a survey of physical therapy experts. *Phys Ther* 1993;73:138–50.
- 22 Cleary K. Using the Delphi process to reach consensus. *J Cardiopulm Phys Ther* 2001;1:20–3.
- 23 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use.* 4th ed. Oxford: Oxford University Press; 2008.
- 24 Dillman DA. *Mail and internet surveys: the tailored design method.* 2nd ed. New York: John Wiley & Sons; 2000. p.363.
- 25 Lopopolo RB. Hospital restructuring and the changing nature of the physical therapist's role. *Phys Ther* 1999;79:171–85.
- 26 Pesik N, Keim M, Sampson TR. Do US emergency medicine residency programs provide adequate training for bioterrorism? *Ann Emerg Med.* 1999;34:173–6.
- 27 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES). *J Med Internet Res.* 2004;6:e34.
- 28 Delbecq AL, van de Ven AH, Gustafson DH. *Group techniques for program planning: a guide to nominal and Delphi processes.* Glenview, IL: Scott, Foresman and Company; 1975.
- 29 Hegedus EJ, Stern B. Beyond SpPIN and SnNOUT: considerations with dichotomous tests during assessment of diagnostic accuracy. *J Man Manip Ther.* 2009;17:E1–5.
- 30 Chu H, Cole SR. Sample size calculation using exact methods in diagnostic test studies. *J Clin Epidemiol.* 2007;60(11):1201–2; author reply 1202.
- 31 Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol.* 2005;58:859–62.
- 32 Cook C, Hegedus E. *Orthopedic physical examination tests: an evidence-based approach.* 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2013.
- 33 Cleland J, Koppenhaver S. *Netter's orthopaedic clinical examination: an evidence-based approach.* 2nd ed. Philadelphia, PA: Saunders; 2010.