Applicability of diagnostic studies – statistics, bias and estimates of diagnostic accuracy

Jos Kleijnen

Director, Kleijnen Systematic Reviews Ltd

Clinical Professor, Joanna Briggs Institute, University of Adelaide

Jos's road to success for applicability

- Get the question right
- Get the study design right
- Include patients for whom the test will also be used in practice
- Educate users of research

QUADAS

- Evidence-based quality assessment tool for test accuracy studies.
- Developed approximately ten years ago, published 2003

  (Whiting et al, BMC Medical research Methodology 2003;3:25).
- Cited over 300 times and mentioned in 96 reviews indexed on DARE.
- Modified version adopted by the Cochrane collaboration.

  http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/ch09_Oct09.pdf

Definition of quality for QUADAS 1

"Both the internal and external validity of a study; the degree to which estimates of diagnostic accuracy have not been biased, and the degree to which the results of a study can be applied to patients in practice."

The QUADAS tool

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?			
2.	*Were selection criteria clearly described?*			
3.	Is the reference standard likely to correctly classify the target condition?			
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?			
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?			
6.	Did patients receive the same reference standard regardless of the index test result?			
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?			
8.	*Was the execution of the index test described in sufficient detail to permit replication of the test?*			
9.	*Was the execution of the reference standard described in sufficient detail to permit its replication?*			
10.	Were the index test results interpreted without knowledge of the results of the reference standard?			
11.	Were the reference standard results interpreted without knowledge of the results of the index test?			
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?			
13.	Were uninterpretable/ intermediate test results reported?			
14.	Were withdrawals from the study explained?			

Experience and feedback suggest that revision is needed.

Sources of Variation and Bias in Studies of Diagnostic Accuracy

A Systematic Review

Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Afina S. Glas, MD, PhD; Patrick M.M. Bossuyt, PhD; and Jos Kleijnen, MD, PhD

Background: Studies of diagnostic accuracy are subject to different sources of bias and variation than studies that evaluate the effectiveness of an intervention. Little is known about the effects of these sources of bias and variation.

Purpose: To summarize the evidence on factors that can lead to bias or variation in the results of diagnostic accuracy studies.

Data Sources: MEDLINE, EMBASE, and BIOSIS, and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration. Methodologic experts in diagnostic tests were contacted.

Study Selection: Studies that investigated the effects of bias and variation on measures of test performance were eligible for inclusion, which was assessed by one reviewer and checked by a second reviewer. Discrepancies were resolved through discussion.

Data Extraction: Data extraction was conducted by one reviewer and checked by a second reviewer.

Data Synthesis: The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias, the amount of evidence was limited. Evidence was lacking for other features, including incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts.

Conclusions: Many issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation; however, the empirical evidence about the size and effect of these issues is limited.

Ann Intern Med. 2004;140:189-202. www.annals.org

For author affiliations, see end of text.

Sources of variation and bias in test accuracy studies: An updated systematic review

Objective:

To summarise evidence on factors that can lead to bias or variation in the results of test accuracy studies.

Methods:

Systematic review updating: Whiting et al, Annals of Internal Medicine 2004;140:189-202.

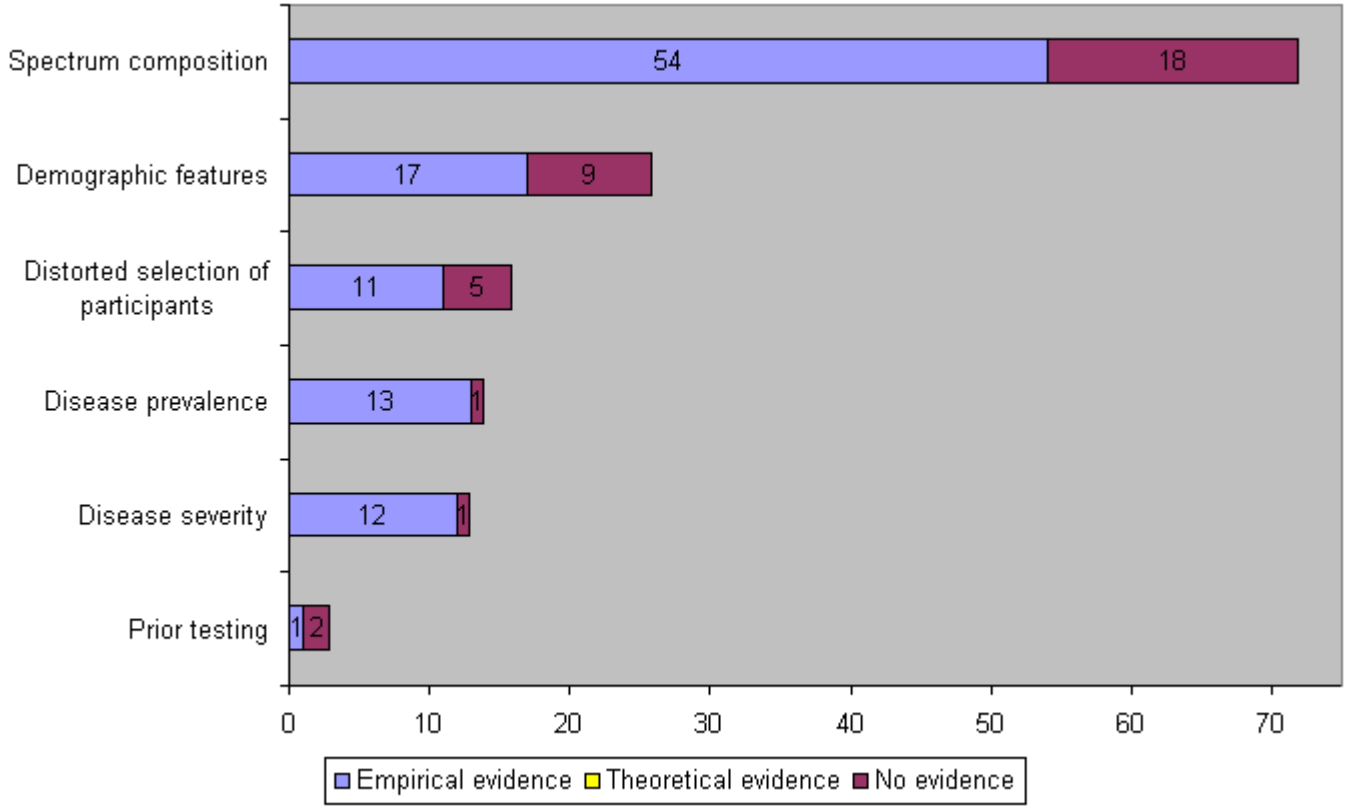
Including empirical studies (experimental studies, diagnostic accuracy studies, systematic reviews) and theoretical studies (modeling studies).

Summary of findings

- 101 reports identified (55 included in previous review, 46 new).
- 91 reports provided empirical evidence and 16 provided theoretical evidence.
- New studies were concentrated in the following areas:
  - Spectrum-related sources of variation.
  - Verification issues.
  - Review biases.
- Direction of effects varied.

Factors evaluated ‘considerably’ (1)

Spectrum composition



The Cochrane Diagnostic Test Accuracy Working Group

Zefram Cochrane – First flight at warp speed in 2063 at 11.15am local time

Viewpoint

Randomised comparisons of medical tests: sometimes invalid, not always efficient

Patrick M M Bossuyt, Jeroen G Lijmer, Ben W J Mol

Lancet 2000; **356:** 1844–47

Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre University of Amsterdam, PO Box 22700, Amsterdam 1100 DE, Netherlands (P M M Bossuyt PhD, J G Lijmer MD, B W J Mol MD)

Correspondence to: Prof Patrick Bossuyt (e-mail: p.m.bossuyt@amc.uva.nl)

Diagnostic tests

- Accurately diagnosing a disease state
- New test finding additional cases compared with old test
- Monitoring
- Screening tests
- New test to replace old test
- Test as triage instrument for further (invasive) testing
- In which order should multiple tests be used?
- Test to decide who needs treatment
- Tests to predict response to treatment and to predict who gets adverse effects

Cite this article as: BMJ, doi:10.1136/bmj.38450.646088.E0 (published 1 June 2005)

Papers

Ultrasonography in screening for developmental dysplasia of the hip in newborns: systematic review

Nerys F Woolacott, Milo A Puhan, Johann Steurer, Jos Kleijnen

Objective

- The objective of this research was to evaluate the effectiveness, clinical impact and cost-effectiveness of ultrasound in screening of newborns for developmental dysplasia of the hip (DDH).

Questions

1. What is the diagnostic accuracy of ultrasound in screening of newborns for DDH?
2. What is the impact of ultrasound in screening of newborns for DDH on the therapeutic decisions and on patient outcomes?
3. What is the cost-effectiveness of ultrasound in screening of newborns for DDH?

Types of studies included per question

- Accuracy studies in unselected newborns allowing the creation of 2 by 2 tables for the first question
- Comparative studies in the relevant population for the other questions

Excluded studies

- Studies with a selected population, e.g., one that only included infants with clinical signs of DDH or with risk factors for DDH.
- Technical reports describing the technique of ultrasound screening, but containing no clinically relevant outcomes.

Results (1)

- We found 683 references. 182 papers were of potential interest and of these 169 were obtained and appraised for inclusion. A total of 58 references describing 57 studies were included in the review.
- No studies of the diagnostic accuracy of ultrasound in screening for DDH in newborns were found.
- A total of seven studies that evaluated the impact of ultrasound in screening newborn infants for DDH on the therapeutic decisions and on patient outcomes were identified.

Results (2)

- The overall quality of the included studies was poor. Of the seven studies included for assessing the impact of screening only one was a randomised controlled trial (RCT) of limited quality, the others were mostly retrospective, with historical controls.
- The populations included in the studies came from five countries (Austria, Jordan, Norway, Poland and the UK) all from various periods between 1980 and 1996.

Results (3)

- Evidence from one RCT of limited quality indicates that ultrasound screening reduces the number of cases of DDH diagnosed after one month of age (late diagnosis). The clinical significance of this is unclear.
- There is no indication of adverse consequences associated with general ultrasound screening of newborns for DDH or any associated treatments

Results (4)

- Evidence from one RCT of limited quality indicates that screening of newborns at birth for DDH using ultrasound appears to increase overall treatment rates and data from one small poor quality study hints that late screening may result in a lower treatment rate
- Evidence from non-randomised, retrospective studies indicates that general screening of newborns at birth for DDH using ultrasound may reduce the severity and invasiveness of the treatments required for DDH and its consequences. Such treatments may be initiated sooner and may be of shorter duration

Conclusions

- The accuracy of hip ultrasound as a screening test is unknown
- The quality of the available evidence is poor
- There is a lack of evidence. Studies that address the questions relating to the true course of DDH, the effect of treatment, and the accuracy of ultrasound screening are required.

ASTHMA

External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?

Justin Travers, Suzanne Marsh, Mathew Williams, Mark Weatherall, Brent Caldwell, Philippa Shirtcliffe, Sarah Aldington, Richard Beasley

Thorax 2007;**62**:219–223. doi: 10.1136/thx.2006.066837

Methods: A postal survey was sent to 3500 randomly selected individuals aged 25–75 years. Respondents were invited to complete a detailed respiratory questionnaire and pulmonary function testing. Participants with current asthma were assessed against the eligibility criteria of the 17 major asthma RCTs cited in the Global Initiative for Asthma (GINA) guidelines.

Findings: A total of 749 participants completed the full survey, of whom 179 had current asthma. A median 4% of participants with current asthma (range 0–36%) met the eligibility criteria for the included RCTs. A median 6% (range 0–43%) of participants with current asthma on treatment met the eligibility criteria.

Interpretation: This study shows that the major asthma RCTs on which the GINA guidelines are based may have limited external validity as they have been performed on highly selected patient populations. Most of the participants with current asthma on treatment in the community would not have been eligible for these RCTs.

Thorax 2007;**62**:219–223.

Stakeholder's comments

- "You have not considered the majority of the evidence, it was all excluded!"
- "In our topic, consequences of wrong care are so extreme (serious morbidity, mortality) that it is unethical to do studies, you have to trust us experts to give optimum treatment"

Statistics

- At the current stage of diagnostic science, statistics are mostly irrelevant, things have gone wrong already before numbers come into play.
- Hope to be invited at 10th symposium to talk about statistics
  - ☐Dealing with heterogeneity
  - ☐Corrections for small study bias

Summary and conclusions – applicability of diagnostic studies

- Get the question right
- Get the study design right
- Include patients for whom the test will also be used in practice
- Educate users of research