ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 1093−1104

Journal of Clinical Epidemiology

A systematic review classifies sources of bias and variation in diagnostic test accuracy studies

Penny F. Whiting[a,b,*], Anne W.S. Rutjes[c,d], Marie E. Westwood[a], Susan Mallett[e], and the QUADAS-2 Steering Group[1]

[a]*Kleijnen Systematic Reviews Ltd, Unit 6, Escrick Business Park, Riccall Road, Escrick, York YO19 6FD, United Kingdom*
[b]*School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, United Kingdom*
[c]*Division of Clinical Epidemiology & Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland*
[d]*Centre for Aging Sciences (Ce.S.I.), G. d'Annunzio University Foundation, Chieti, Italy*
[e]*Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, United Kingdom*

Accepted 15 May 2013; Published online 17 August 2013

Abstract

Objective: To classify the sources of bias and variation and to provide an updated summary of the evidence of the effects of each source of bias and variation.

Study Design and Setting: We conducted a systematic review of studies of any design with the main objective of addressing bias or variation in the results of diagnostic accuracy studies. We searched MEDLINE, EMBASE, BIOSIS, the Cochrane Methodology Register, and Database of Abstracts of Reviews of Effects (DARE) from 2001 to October 2011. Citation searches based on three key papers were conducted, and studies from our previous review (search to 2001) were eligible. One reviewer extracted data on the study design, objective, sources of bias and/or variation, and results. A second reviewer checked the extraction.

Results: We summarized the number of studies providing evidence of an effect arising from each source of bias and variation on the estimates of sensitivity, specificity, and overall accuracy.

Conclusions: We found consistent evidence for the effects of case−control design, observer variability, availability of clinical information, reference standard, partial and differential verification bias, demographic features, and disease prevalence and severity. Effects were generally stronger for sensitivity than for specificity. Evidence for other sources of bias and variation was limited.

Keywords: Bias; Test accuracy; Sensitivity; Specificity; Systematic review; Variation

1. Introduction

Evidence on diagnostic accuracy contributes to the appropriate use of diagnostic tests in clinical practice. The use of inaccurate tests can result in serious errors in diagnosis, which may affect treatment decisions and patient outcome. Primary diagnostic test accuracy (DTA) studies compare the results of the test of interest (index test) with those of the best available method of determining disease status (clinical reference standard). The results are cross-tabulated to produce a 2 × 2 table of results based on which measures of the accuracy of the index test can be calculated, for example, sensitivity, specificity, likelihood ratios, and predictive values.

If a study has limitations in its design or conduct, estimates of diagnostic accuracy can differ systematically from the true accuracy, leading to *bias*. In contrast, a source of *variation* is a feature that can result in differences in the (true) diagnostic accuracy across studies. Sources of such variation may be differences in test protocol, differences in study populations, or differences in how the target condition is defined [1,2]. Because of this variation, reported estimates of accuracy, although possibly unbiased, may have limited *applicability* to a specific clinical question. When evaluating a diagnostic accuracy study, it is, therefore, essential to consider both the potential for bias and sources of variation, which determine applicability.

[1] The members of the QUADAS-2 Steering Group: *University of Oxford*: Susan Mallett; *University of Bristol*: Jonathan A.C. Sterne, Penny F. Whiting; *University of Amsterdam*: Patrick M.M. Bossuyt, Mariska M.G. Leeflang; *University of Birmingham*: Jonathan J. Deeks; *University Medical Center Utrecht*: Johannes B. Reitsma; *University of Bern and G. d'Annunzio University Foundation*: Anne W.S. Rutjes; *Kleijnen Systematic Reviews*: Marie E. Westwood.

* Corresponding author. Tel.: +44 (0)1904 727980; fax: +44 (0)1904 720429.

E-mail address: penny@systematic-reviews.com (P.F. Whiting).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.05.014

What is new?

- This review found consistent evidence for the effects of case−control design, observer variability, availability of clinical information, reference standard, partial and differential verification bias, demographic features, and disease prevalence and severity.
- Effects were generally stronger for sensitivity than for specificity.
- Evidence for other sources of bias and variation was limited.
- This review provides an updated classification and overview on the sources of bias and variation in test accuracy studies.
- Primary studies should attempt to minimize the potential for bias and variation.
- End users should consider the potential for bias and variation in primary diagnostic accuracy studies.

In 2004, we published a systematic review on the sources of bias and variation in studies of the accuracy of diagnostic tests [3]. Since this review was published, there has been growing interest in the field of diagnostic accuracy studies and DTA reviews. The Cochrane collaboration has started accepting DTA reviews, the UK National Institute for Health and Clinical Excellence now commissions diagnostic assessment reviews, and the German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen and the US Agency for Healthcare Research and Quality also include diagnostic topics. It is, therefore, increasingly important to have an up-to-date overview of the current evidence in this area. This review aims to provide an overview of the effects of bias and variation in DTA studies classified according to the following four domains: patient selection, index test, reference standard, and flow and timing.

2. Methods

2.1. Literature searches

We searched MEDLINE, EMBASE, BIOSIS, the Cochrane Methodology Register, and DARE from 2001 to October 2011. All studies included in the original review [3] (search date to 2001) were eligible for inclusion in this review. Full details of the search strategy are available on request. In addition, we carried out a citation search to identify studies that cited the key papers (Begg, 1987 [4]; Lijmer et al., 1999 [5]; and Whiting et al., 2004 [3]). Search results were screened for relevance independently by two reviewers; disagreements were resolved through consensus.

2.2. Inclusion criteria

Studies were included if their primary objective was the quantification or clarification of bias or variation in the measurement of diagnostic test operating characteristics.

Studies of any design covering any topic area were eligible. Studies had to investigate the effects of bias or variation on the measures of test performance such as sensitivity, specificity, predictive values, likelihood ratios, and diagnostic odds ratios (DORs). Studies that reported the methods for correcting for bias (e.g., verification bias) were excluded. Inclusion was assessed by one reviewer and checked by a second; discrepancies were resolved through discussion or referral to a third reviewer.

2.3. Data extraction

One reviewer extracted data on the following: study details, study design (meta-review, review, primary DTA study, experimental study, or modeling study), study population (included reviews/studies/patients), index test and target condition, category of bias/variation (patient selection, index test, reference standard, or flow and timing), specific source of bias/variation, specific factors investigated, and effects on sensitivity, specificity, and overall accuracy (increased, decreased, associated, associated but direction unclear, or no association). A second reviewer checked the data extraction. Discrepancies were resolved by consensus or consultation with a third reviewer.

2.4. Data synthesis

Because of differences between studies, it was not possible to pool data, and so a narrative synthesis was prepared. Results were stratified according to the source of bias or variation (see Web Table 1 at www.jclinepi.com). Studies were grouped according to the study design (Box 1). We summarized the number of studies providing evidence of an effect arising from each source of bias and variation. We classified the bias/variation effects on sensitivity, specificity, and measures of overall accuracy (e.g., DOR, area under the receiver operating characteristic curve) as increased, decreased, or associated based on the effect estimates reported in the primary studies. Where studies provided a statistical analysis of the association, we used a threshold of $P < 0.05$ to classify studies as showing a significant association. Studies were classified as having no evidence of effect of bias if a nonsignificant effect size was reported; some studies may have been underpowered to show an effect. For studies that only provided a qualitative interpretation of the association, we made a judgment on the effect of the source of bias or variation based on the effect estimates reported in the papers.

2.5. Role of the funding source

This article was funded by the Medical Research Council (G0801405/1). The sponsor had no role in the study