

How do authors of comparative accuracy studies analyse data when reporting a comparative conclusion: methodological review?

Yaxin Chen¹, Yasaman Vali¹, Sue Mallett², Anne Wilhelmina Saskia Rutjes³, Clare Davenport⁴, Bada Yang⁵, Mariska Leeftang¹

1 Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands

2 Centre for Medical Imaging, University College London, London, UK

3 Saint Camillus International University of Health and Medical Sciences (UniCamillus), Rome, Italy

4 Department of Applied Health Science, University of Birmingham, Birmingham, UK

5 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Background

Diagnostic accuracy studies comparing index tests (comparative accuracy studies) often report comparison results where evaluation can be biased by choices in statistical modelling. The quality assessment tool for comparative accuracy studies, QUADAS-C, currently does not include signalling questions to assess biases in the comparison due to statistical choices.

Objectives

Methodological review of statistical analyses used in comparative accuracy studies with the ultimate aim of deriving possible signalling questions for a future QUADAS-C analysis domain.

Methods

We searched all systematic reviews of diagnostic test accuracy published in 2023 using PubMed. Systematic reviews with a comparative aim and containing at least 10 potential comparative primary studies were selected. From all comparative studies included in these reviews, we randomly selected a subset of 200 studies. Single data extraction was conducted by five reviewers.

Results

Of 200 studies, 53 studies compared two tests, and 147 studies compared more than two tests. Eighty-three percent of the studies (166/200) drew a comparative conclusion. The three accuracy measures that were used the most for comparison were sensitivity and specificity (164/200), area under the receiver operating characteristic curve (92/200), and predictive values (52/200). About half of the studies (99/200) formally compared accuracy measures, using a statistical test. The McNemar test (22/99), DeLong's test (18/99), and the Chi-square test (7/99) were the most commonly used. Fifteen studies provided formal sample size calculations, of which nine were based on the comparative questions. Fifteen studies clearly reported missing data: five had none, six excluded patients with missing data, four replaced missing values with zero. Ten studies (of which eight used the fully paired) took confounding into consideration and the most common method was stratification (8/10).

Conclusions

Most studies drew a comparative conclusion, but few studies reported enough transparent details about statistical methods to evaluate the comparison. These results are important to inform risk of bias questions for the QUADAS-C analysis domain. Future steps involve the potential effects of chosen methods on final inference with respect to the comparison.

Keyword Comparative accuracy studies; Diagnostic accuracy; Test comparison; Statistical analyses; Bias