Sample sizes of studies on diagnostic accuracy: literature survey
Lucas M Bachmann, Milo A Puhan, Gerben ter Riet, Patrick M Bossuyt

Abstract

Objectives To determine sample sizes in studies on diagnostic accuracy and the proportion of studies that report calculations of sample size.

Design Literature survey.

Data sources All issues of eight leading journals published in 2002.

Methods Sample sizes, number of subgroup analyses, and how often studies reported calculations of sample size were extracted.

Results 43 of 8999 articles were non-screening studies on diagnostic accuracy. The median sample size was 118 (interquartile range 71-350) and the median prevalence of the target condition was 43% (27-61%). The median number of patients with the target condition—needed to determine a test's sensitivity—was 49 (23-108). The median number of patients without the target condition—needed to determine a test's specificity—was 161 (33-363). Two of the 43 studies (5%) reported a priori calculations of sample size. Twenty articles (47%) reported results for patient subgroups. The number of subgroups ranged from two to 19 (median four). No studies reported that sample size was calculated on the basis of predefined analyses of subgroups.

Conclusion Few studies on diagnostic accuracy report considerations of sample size. The number of participants in most studies on diagnostic accuracy is probably too small to analyse variability of measures of accuracy across patient subgroups.

Introduction

Estimates of sensitivity and specificity in small studies on diagnostic accuracy are usually imprecise, with wide confidence intervals. This makes it difficult to assess just how informative a test may be. Subgroup analysis is often needed because sensitivity and specificity may vary across patient subgroups, yet estimates are even less precise when subgroups are considered.1 Investigators should calculate the sample size needed for sufficiently narrow confidence intervals at the planning stages of a study, as is common practice for randomised trials.2,3 For example, if a diagnostic test requires a sensitivity of at least 90% for adequate decision making, the lower boundary of the 95% confidence interval should be at least 90%.

We hypothesised that studies of diagnostic accuracy rarely report considerations of sample size and tend to be small. We assumed that authors would state calculations of sample size if they had been performed. We investigated study sizes, the number of subgroup analyses, and how often studies on diagnostic accuracy reported calculations of sample sizes.

Methods

Two reviewers independently screened all issues of the BMJ, Lancet, New England Journal of Medicine, and JAMA as well as four specialist journals (Thorax, Gastroenterology, American Journal of Obstetrics and Gynecology, and European Journal of Pediatrics) published in 2002 for studies on the accuracy of tests. From each full report we extracted data on the type of test(s) studied (table), study sizes, the number of subgroup analyses, and how often the studies reported calculations of sample size. We calculated 95% confidence intervals, medians, and interquartile ranges.

Results

Fifty seven of 8999 articles reported test accuracy. Fourteen studies focused on a screening test and were excluded, which left 43 clinical studies for analysis. The median sample size was 118 (interquartile range 71-350) and the median prevalence was 43% (27-61%). The median number of patients with the target condition—needed to calculate a test's sensitivity—was 49 (23-108). The median number of patients without the target condition—needed to determine a test's specificity—was 161 (33-363).

Two of 43 studies (5%; 95% confidence interval 1.3% to 15.5%) reported a priori calculations of sample size, but no study reported that the sample size had been calculated on the basis of predefined analyses of subgroups. Twenty articles (47%) reported results for subgroups of patients. The number of subgroups ranged from two to 19 (median four). Four studies used multivariable regression, but none used interaction terms.

Discussion

In this survey of studies on diagnostic accuracy in eight major journals, only 4.7% of the studies reported that they considered sample size. Analysing small numbers of participants with and without the target condition usually yields imprecise estimates of overall diagnostic accuracy, and even less precise estimates of subgroups. For example, when the number of patients with the target condition is 49 the two sided 95% confidence interval of a sensitivity of 81% (40 true positives) is 68% to 91%.

To ensure reasonably precise estimates of sensitivity and specificity investigators should consider sample sizes during the planning stages of the study. Investigators should calculate how precise the estimates of test accuracy should be for a particular diagnostic situation and report these calculations with confidence intervals. Arguably, sample size calculations are not important once data collection has been completed.6 All that matters is the width of the confidence intervals. However, besides determining the minimum study size needed, calculations of sample size have another useful feature that remains important.
after the study has finished. These calculations require authors to think about the minimum precision needed for a test to be clinically meaningful. It is easier for readers to interpret reported confidence intervals if they have access to these data.

In conclusion, few studies on diagnostic accuracy report calculations of sample size. The number of participants in most studies on diagnostic accuracy is probably too small to analyse...
the variability of measures of accuracy across subgroups of patients.

Contributors: All members of the SUBIRAR (subjectivity rationality and reasoning) research collaboration (Klaus Eichler, Madlaima Scharplatz, and Johann Steurer, Horten Centre, University of Zurich, Switzerland, Ulrich Hoffrage, Max Planck Institute for Human Development and Cognition, Berlin, Germany; Alfons G Kessels, Hans Severens, Maastricht University, Germany; Khalid S Khan, University of Birmingham, UK; Jos Kleijnen, Centre for Reviews and Dissemination, University of York, UK) were involved in the design and critical review of the study. LMB, MAP, and GtR developed the protocol. LMB and MAP acquired the data. All authors interpreted the data and helped prepare the manuscript. LMB was guarantor.

Funding: LMB was supported by the Swiss National Science Foundation (grants 3233B0-103182 and 3200B0-103183).

Competing interests: None declared.


doi 10.1136/bmj.38793.637789.2F

Division of Epidemiology and Biostatistics, Department of Social and Preventive Medicine, University of Bern, Switzerland
Lucas M Bachmann senior research fellow

Horten Centre, University of Zurich, CH-8091 Zurich, Switzerland
Milo A Puhan research fellow

Department of General Practice, Academic Medical Centre, 1105 AZ Amsterdam, Netherlands
Gerben ter Riet clinical epidemiologist

Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, Amsterdam, Netherlands
Patrick M Bossuyt professor

Correspondence to: L M Bachmann lucas.bachmann@evimed.ch