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 calculate a test's sensitivity—was 49 (28-91). The median number of patients without the target condition—needed to determine a test's specificity—was 76 (27-209). 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 preplanned 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 (28-91). The median number of patients without the target condition—needed to determine a test's specificity—was 76 (27-209).

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 preplanned 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%.[4 5]

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.[2] 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.

This article was posted on bmj.com on 20 April 2006: http://bmj.com/cgi/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

BMJ 2006;332:1127–9

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.

Key features of 57 studies on accuracy of diagnostic tests published in eight major medical journals in 2002

First author	Type of test	Prevalence (%)	Sample size	Screening	Subgroup analysis	Multivariable analysis	Stratified reporting	Number of subgroups
Schneider	Imaging	2	8640	Yes	No	No	No	–
Pilcher	Laboratory tests	0.5	8194	Yes	No	No	No	–
Bahado-singh (1)	Laboratory tests	2	5641	Yes	Yes	Yes	No	2
Kulasingam	Laboratory tests	3	4075	Yes	Yes	No	Yes	2
Lu	Physical examination	1	3710	Yes	No	No	No	–
Vintzileos	Imaging	2	3291	Yes	No	No	No	–
Vasan	Laboratory tests	6	3177	Yes	Yes	Yes	Yes	4
Bahado-singh (2)	Imaging	3	3003	No	Yes	Yes	No	5
Selvachandran	History	4	2268	No	Yes	Yes	Yes	2
Maisel	Laboratory tests	47	1586	No	No	No	No	–
Lenders	Laboratory tests	25	858	No	Yes	No	Yes	2
Tibble	Laboratory tests	44	602	No	Yes	No	Yes	2
Bahado-singh (3)	Laboratory tests	3	568	Yes	Yes	Yes	No	–
Azuma	Laboratory tests	6	561	Yes	Yes	No	Yes	3
Ikeda	Physical examination	59	529	No	No	No	No	–
Laing	History	21	458	Yes	No	No	No	–
Schutter	Laboratory tests	55	412	No	No	No	No	–
Wang	Laboratory tests	50	394	No	Yes	No	Yes	2
Muensterer	Imaging	6	386	No	No	No	No	–
Chavarria	Laboratory tests	7	378	No	No	No	No	–
Rettenbacher	Imaging	17	350	No	Yes	No	Yes	3
Rubin	Laboratory tests	39	342	No	No	No	No	–
Luck	History	4	341	Yes	No	No	No	–
Ghezzi	Laboratory tests	3	306	No	No	No	No	–
Riordan	History	73	278	No	Yes	Yes	No	19
Kim	Laboratory tests	36	251	No	Yes	No	Yes	2
Vayssiere	History	5	242	Yes	Yes	Yes	No	2
Virkki	Laboratory tests	85	215	No	Yes	Yes	No	7
Remes	History	16	212	No	No	No	No	–
Hughes	Laboratory tests	4	208	No	Yes	No	Yes	2
Bouin	Other	43	199	No	No	No	No	–
Ribeiro	Laboratory tests	85	177	No	Yes	No	Yes	2
Riskin-Mashiah	Imaging	6	166	Yes	No	No	No	–
Selan	Laboratory tests	27	139	No	No	No	No	–
Oudkerk	Imaging	30	118	No	No	No	No	–
Mihm	Laboratory tests	58	113	No	No	No	No	–
McManus	Other	64	110	Yes	No	No	No	–
McMahon	Physical examination	12	109	No	No	No	No	–
Stiller	Other	6.5	107	No	No	No	No	–
Dueholm	Imaging	69	106	No	No	No	No	–
Andrews	Imaging	53	100	No	No	No	No	–
Joossens	Laboratory tests	32	97	No	Yes	No	Yes	4
DeRoche	Laboratory tests	84	90	No	No	No	No	–
Narang	Laboratory tests	39	80	No	No	No	No	–
Harewood	Imaging	61	80	No	No	No	No	–
Larsen	Other	75	79	No	Yes	No	Yes	2
Warke	Laboratory tests	41	71	No	No	No	No	–
Hara	Imaging	66	60	No	No	No	No	–
Gerber	Laboratory tests	34	53	No	No	No	No	–
Chmait	Imaging	85	53	No	No	No	No	–
Georgakoudi	Laboratory tests	64	44	No	No	No	No	–
Ragette	Other	79	42	No	Yes	No	Yes	3
Parker	Imaging	*	33	No	No	No	No	–
Odunsi	Laboratory tests	39	33	No	No	No	No	–
Cosmi	Imaging	53	32	No	No	No	No	–
Broth	Other	41	29	No	No	No	No	–
Satoh	Laboratory tests	61	23	No	No	No	No	–

This table provides information for both screening (excluded) and non-screening studies.
*Could not be determined.

What is already known on this topic

To assess the minimum size needed for sufficiently narrow confidence intervals of sensitivity and specificity in study groups as a whole and in clinically relevant subgroups in particular, sample sizes should be considered at the planning stage of studies on test accuracy

What this study adds

Few studies on test accuracy report calculations of sample size

Overall size and subgroup size tend to be small in these studies, which leads to imprecise estimates of sensitivity and specificity

Contributors: All members of the SUBIRAR (subjectivity rationality and reasoning) research collaboration (Klaus Eichler, Madlaina 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.

1 Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002;324:669-71.
2 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; 365:1348-53.
3 Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525-37.
4 Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford Statistical Science Series, Oxford University Press, 2003. www.fhcrc.org/science/labs/pepe/book/ (accessed 6 Apr 2006).
5 Pepe MS. Study design and hypothesis testing. In: *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press, 2003:214-51.

(Accepted 7 March 2006)

doi 10.1136/bmj.38793.637789.2F

Commentary: Improving the quality and clinical relevance of diagnostic studies

Frans H Rutten, Karel G M Moons, Arno W Hoes

Bachmann and colleagues show that few studies on diagnostic accuracy include calculations of sample size. Most such studies are too small to provide precise estimates of the overall sensitivity and specificity of a test, let alone for subgroups,[1] and few studies have investigated this issue. We support the authors' recommendation that all diagnostic studies should calculate sample size at the planning phase, especially as straightforward methods are available for assessing simple proportions, such as sensitivity and specificity. However, they used the specificity and sensitivity of single tests to calculate sample size (understandable given the predominance of these tests in research) and did not consider the increasing number of clinically relevant studies that measure the accuracy of several tests in combination.[2]

If you were testing the accuracy of B-type natriuretic peptide (BNP) for excluding heart failure in primary care, for example, precise estimation of the sensitivity and specificity of the test might seem important. Such tests, however, have limited value in clinical practice. Firstly, in daily practice positive and negative values merely help doctors to estimate the probability of disease.[3] Secondly, a diagnosis in practice is seldom based on one test. Doctors would probably use the BNP test only if it provided extra diagnostic information to other measures such as signs and symptoms, which have already been assessed. To improve clinical practice, it would be better to measure the diagnostic accuracy of combinations of readily available tests (applying multivariable regression analysis with receiver operating characteristic curves) and then assess whether the addition of BNP improves accuracy.[4] The BNP test should not be used when the patient's history and physical examination would provide equivalent diagnostic information.

We know even less about determinations of sample size for multivariable diagnostic studies. The number of tests studied is usually limited to allow for adequate data analysis. An often used rule is that at least 10 patients with the disease should be tested for each diagnostic test evaluated.[5] Such ways of determining sample size are not ideal. If the method suggested by Bachmann and colleagues is used to determine sample size in evaluations of multiple tests, many assumptions must be made to achieve acceptable proportions of false negative and false positive diagnoses when a cut-off value is introduced.

Methodological improvements are needed to guide considerations of sample size in diagnostic research. Lack of consensus on some of these issues is no excuse for "complete" lack of prior calculations of sample size in diagnostic studies. Bachmann and colleagues showed that a lack of such calculations is common. We hope that authors of studies on diagnostic tests will soon adopt more rigorous guidelines based on the standards for reporting of diagnostic accuracy (STARD initiative; www.consort-statement.org/Initiatives/newstard.htm).

Contributors: FHR, KGMM, and AWH critically discussed the structure of this article. FHR wrote the first draft and KGMM and AWH critically revised the manuscript.

Competing interests: None declared.

1 Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ* 2006;332:1127-9.
2 Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem* 2004;50:473-6.
3 Moons KG, Harrell FE. Sensitivity and specificity should be deemphasized in diagnostic accuracy studies. *Acad Radiol* 2003;10:670-2.
4 Rutten FH, Moons KGM, Cramer MJM, Grobbee DE, Zuithoff NPA, Lammers JWJ, et al. Recognising heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: a cross-sectional diagnostic study. *BMJ* 2005;331:1379-85.
5 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373-9.

Julius Centre for Health Sciences and Primary Care, University Medical Centre, Utrecht, 3508 AB, Netherlands
Frans H Rutten *general practitioner*
Karel G M Moons *professor of clinical epidemiology*
Arno W Hoes *professor of clinical epidemiology and general practice*
Correspondence to: F H Rutten F.H.Rutten@umcutrecht.nl