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