Systematic reviews of test accuracy studies in reproductive health

Honest H, Khan KS

Department of Obstetrics & Gynaecology, Birmingham, United Kingdom

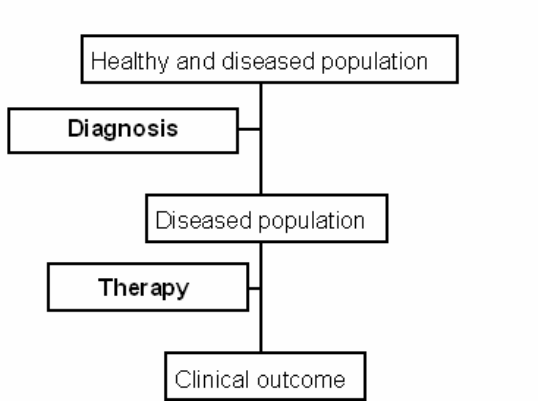
ABSTRACT

Testing, whether used for diagnosis or screening, is a critical part of the clinical process in reproductive health. It is now accepted that absence of clear summaries of individual research studies on clinical tests is a major impediment in evidence-based practice. Just as systematic reviews of effectiveness of therapeutic and preventative interventions have been pursued over the last decades, so attention is now being given to research on systematic reviews of test accuracy studies. This paper delineates the process of reviewing test accuracy literatures in order to allow readers to critically appraise such reviews.

INTRODUCTION

In women's health, over the last decade, there has been a focus on systematic reviews of effectiveness of therapeutic and preventative interventions. This is evident from the large number of reviews found in the Cochrane Library and the WHO Reproductive Health Library. Recently, however, systematic reviews identifying, appraising, and summarising the results of screening and diagnostic test evaluations have gained increasing visibility in the medical literature. Reviewers' attention is becoming focussed on systematic reviews of test accuracy literature (1, 2, 3). Considering the clinical process Figure 1. Temporal relation of the need for diagnostic and therapeutic evidence in the clinical process,

Figure 1
Temporal relation of the need for diagnostic and therapeutic evidence in the clinical process



this development is hardly surprising. After all clinicians cannot use effective therapies without making an accurate diagnosis first. Potential harm might come to patients as results of delayed diagnosis or misdiagnosis (and consequent

administration of wrong treatments). Accurate tests, on the other hand, allow timely diagnosis, correct prognosis, and appropriate treatments.

Considering the clinical process this development is hardly surprising. After all clinicians cannot use effective therapies without making an accurate diagnosis first. Potential harm might come to patients as results of delayed diagnosis or misdiagnosis (and consequent administration of wrong treatments). Accurate tests, on the other hand, allow timely diagnosis, correct prognosis, and appropriate treatments.

In this paper we would like to highlight the process of reviewing test accuracy literature with view to enabling readers to appraise such reviews. First of all the steps involved should be understood (see below). These are similar to those used for typical effectiveness reviews included in the WHO Reproductive Health Library. When undertaking an accuracy review one has to go through:

- Stating the aims and objectives of the review clearly
- Undertaking a thorough search to identify relevant literature
- Assessing study quality for potential biases in accuracy assessment
- Synthesising the extracted data

These steps should be included in a protocol describing how the review is to be conducted. Such a protocol is equivalent to, and as important as, a protocol for primary research. In the absence of a protocol, the review may be unduly influenced by presumption of its findings. Hence, it is the protocol that makes systematic reviews research projects in their own right.

1. STATING QUESTIONS ABOUT ACCURACY OF TESTS

Contrary to popular perception, the term 'test' does not confine itself to signify laboratory tests or the likes of radiological imaging only. Patient's characteristics, history, examination and many simple bedside tests also provide powerful information to reach a diagnosis. These should be considered along with laboratory, radiological and other tests in the diagnostic process when formulating questions for reviews of test accuracy. Focussed and well-structured questions are crucial in making a test accuracy review efficient and valuable to both reviewers and readers alike. The question should state explicitly the target population and their characteristics, the test to be evaluated and the gold standard against which the accuracy of the test is to be compared. An example question is stated in Table 1.

Narrative question
Among pregnant women, what is the accuracy of cervico-vaginal fetal fibronectin test in predicting preterm birth?

Structured question and selection criteria

Narrative question
Among pregnant women, what is the accuracy of cervico-vaginal fetal fibronectin test in predicting preterm birth?

Structured question and selection criteria

Population	Pregnant women at low or high risk of preterm birth (The people at risk of having the condition of interest)
Test	Antenatal cervico-vaginal fetal fibronectin (The test which purports to predict the presence or absence of the condition)
Gold standard	Spontaneous birth with known gestation either at term or preterm (The condition of interest whose existence is confirmed or refuted beyond reasonable doubt independently of the test being evaluated)

Explicit question generation*a-priori* is paramount, as this would dictate the remaining review process. Changing the question *ad-hoc* or*post-hoc* is liable to introduce bias in the review.

2. IDENTIFYING RELEVANT LITERATURE

The review should state how primary accuracy studies were identified. This is done in several steps. These steps should be documented and their conduct should be transparent. Typically, once the question has been formulated, the next step is to construct a strategy for electronic database searching. Search strategy should explicitly state how widely the internet has been cast in an attempt to identify primary studies. These may include, in addition to searching electronic databases, searching the grey literature, searching the reference lists of primary studies and review articles, and contacting the experts (and manufacturers of the test) for unpublished studies. There should be no language restriction. Restriction in the search, either of databases or of languages, has potential to bias accuracy reviews, (4).

General guidelines on methods of electronic searching are available (5, 6, 7). Essentially, it consists of formulation of an appropriate combination of search terms, pilot searches to refine the search term combination, selection of relevant databases (e.g. Medline, Embase, Pascal, Biosis, and BioBase) and citation retrieval from the refined searches for selection of potentially relevant citations. This is done by scrutinising the title and abstract of citations retrieved from the electronic searching using selection criteria derived from the review question Table 1.

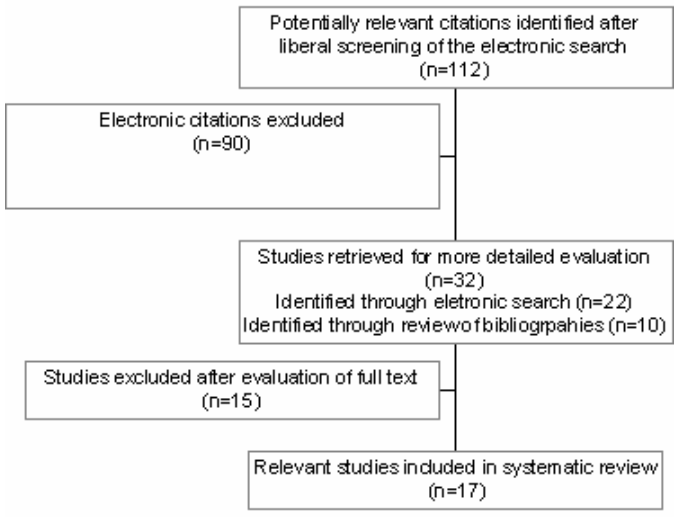
Narrative question
Among pregnant women, what is the accuracy of cervico-vaginal fetal fibronectin test in predicting preterm birth?

Structured question and selection criteria

Narrative question Among pregnant women, what is the accuracy of cervico-vaginal fetal fibronectin test in predicting preterm birth? **Structured question and selection criteria**	
Population	Pregnant women at low or high risk of preterm birth (The people at risk of having the condition of interest)
Test	Antenatal cervico-vaginal fetal fibronectin (The test which purports to predict the presence or absence of the condition)
Gold standard	Spontaneous birth with known gestation either at term or preterm (The condition of interest whose existence is confirmed or refuted beyond reasonable doubt independently of the test being evaluated)

Full papers of all potentially relevant citations are examined to make final inclusion and exclusion decisions based on the explicit selection criteria. The process of literature identification can be a long and drawn out one. An example flow chart representing this process is shown in Figure 2. A flow chart for identification of the literature,

Figure 2
A flow chart for identification of the literature



Based on Chien *et al* [30]

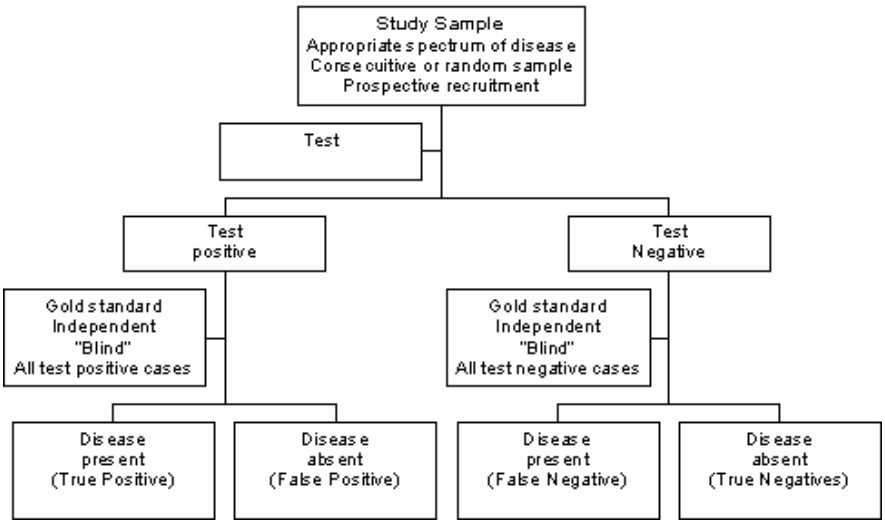
Once potentially relevant papers have been obtained, information is then extracted on methodological quality and accuracy data.

3. ASSESSING QUALITY OF SELECTED STUDIES

Test accuracy studies consist of non-randomised observational studies of defined populations in which the results of the test of interest are compared with the results of a gold standard. These may be prospective or cross-sectional studies. In such studies, methodological quality may be defined as the confidence that the study design, conduct and analysis has minimised biases in estimating the accuracy of the test in question. Variations in study quality may be one source of different results between studies. The extent to which primary research meets methodological standards will influence the strength of any practice recommendations from the review and help make recommendations to improve future studies.

There are several tools available to assess the quality of test accuracy studies (8, 9, 10). The quality features and their relation to an accuracy study design are shown in Figure 3. An accuracy study is designed to generate a comparison between measurements obtained by a test and those obtained by a gold standard. As shown in Figure 3. Design of a test accuracy study and features of its methodological quality,

Figure 3
Design of a test accuracy study and features of its methodological quality



one needs to independently measure the same clinical attribute on two occasions, once by a test and second by a gold standard, and then to discern the relationship between these measurements. In such studies, one possible source of bias is the use a sample which is not representative of the whole spectrum of the clinically relevant population. Accuracy studies may appear to be more optimistic if researchers have deliberately discarded difficult cases from the study. Such omissions are more likely to occur with convenience or arbitrary methods of sampling the study population. Selection bias is less likely to be operative with the use consecutive or random sampling.

The researchers of primary studies on test accuracy should provide sufficient information on the manner in which the test was conducted. For example description of preparation of the patients, measurements of biophysical recordings, details of laboratory assays, computation of results and cut-off levels for defining abnormality should all be provided. Similarly, the gold standard should be an appropriate one, usually a test that is generally acknowledged to be the best available for use as the reference test. In addition, accuracy studies require that observers assessing gold standards verifying the diagnosis be blinded to measurements obtained from the test and vice versa. Blinding avoids bias, as recordings made by one observer are not influenced by the knowledge of the measurements obtained by other observers. Moreover, during the verification process bias may arise if the result of the test under evaluation influences whether study subjects undergo confirmation by the gold standard. This may be the case in some studies where most of the test positive cases but only a minority of the test negative cases are subjected to verification by gold standard.

The purpose of quality assessment is to extract essential information on elements of the study design. In particular, the recruitment, the spectrum and the flow of subjects through the study should be assessed along with the execution of test and blinding of its results to the gold standard. Table 2.

A hierarchy of evidence for primary test accuracy studies

Grade	Level of evidence	Study design
A	1	An independent, blind comparison with reference standard among an appropriate population of consecutive patients.
B	2	An independent blinds comparison with reference standard among an appropriate population of non-consecutive patients or confined to a narrow population of study patients.
B	3	An independent, non-blind comparison with reference standard among an appropriate population of consecutive patients.
B	3	An independent, non-blind comparison with reference standard among an appropriate population of non-consecutive patients or confined to a narrow population of study patients
C	4	An independent, blind comparison among an appropriate population of patients, but reference standard not applied to all study patients.
D	5	Reference standard not applied independently or expert opinion without explicit critical appraisal, based on physiology, bench research or first principles.

Modified from Clark *et al*, (31) Divakaran *et al*, (32) and Sackett *et al* (33)
See Figure 2 for relationship to test accuracy study design.

shows a hierarchy of accuracy evidence based on these features. Empirical evidence of bias is emerging for many of the quality elements (11). It is, therefore, crucial that any test accuracy review should include a comprehensive analysis of the methodological quality of primary studies. These factors, together with characteristics and results of the studies, should be displayed in tabular form, from which, it should be possible to infer whether the test appears accurate when drawing conclusion from a review.

4. SYNTHESISING TEST ACCURACY DATA

Selected studies evaluating test accuracy must provide data on comparison of the test with the gold standard in sufficient detail to allow generation of 2x2 tables for computation of possible accuracy indices. For example, 2x2 tables of the cervico-vaginal fibronectin test result (positive or negative) and spontaneous preterm birth (present or absent) could be produced from each study. Reviewers must obtain missing information from primary investigators. Once the numerical data has been obtained from the various primary studies, the next steps will be exploration of variation in results from study to study (heterogeneity) followed by, if appropriate, synthesis of their results (meta-analysis).

Any variation in results between different studies (heterogeneity) should be investigated. There is likely to be some heterogeneity in population, test, gold standard, and study quality. Conclusions have to be made cautiously if there is significant heterogeneity. Many statistical (12, 13), methods exists to detect whether the apparent differences in test accuracy among studies are due to chance alone. However it is recognised that statistical methods tend to have limited power to detect heterogeneity (14). Therefore it has been recommended that graphical methods (15, 16, 17), should also be used to explore heterogeneity (18). This may involve an exploration of the relationship between sensitivities and specificities for the various studies included in the meta-analysis. Examination of the causes of heterogeneity should be planned *a priori;* otherwise it may be open to bias. Essentially, there are two practical approaches. First, subgroup analyses can be conducted to see whether variations in population, test, outcomes and study quality between different studies affect the estimate of diagnostic accuracy. (19, 20). Second, meta-regression analysis may be performed to determine which one of the several variables considered to be important *a priori;*account for the differences between the studies (21). Where heterogeneity remains unexplained, one should perform data synthesis and interpretation with caution.

In meta-analysis, results from individual studies are pooled together mathematically to generate a summary or pooled result. The various summary measures used to report the pooled results are shown in Table 3.

Summary measures and their use in meta-analysis of test accuracy studies using dichotomous results

Summary measures	**Proportion***
Summary sensitivity (true positive rate)	**58%**

Summary measures	Proportion*
A method of combining the results from primary studies of the proportion of people with disease that is correctly identified as such, independent of specificities.	
Summary sensitivity (true negative rate)	**58%**
A method of combining the results from primary studies of the proportion of people with disease that is correctly identified as such, independent of sensitivities.	
Summary receiver operating characteristics curve (sROC)	**73%**
A method of combining sensitivity and specificity results from individual primary studies that takes into account their relationship between these two measures. The result, which is the average accuracy of the test, obtained by this method is usually presented as area under the curve. This method provides a graphical illustration to the overall accuracy of the test and defined a point where the test was at its most accurate.	
Summary predictive values	**18%**
A method of combining the results from primary studies of the proportions of test positive (or negative) people who truly have (or do not have) disease.	
Summary likelihood ratios	**22%**
A method of combining the results from primary studies of the ratio of the probability of a positive (or negative) test result in the patients with disease to the probability of the same test result in the patients without the disease	
Summary diagnostic odds ratio	**8%**
A method of combining the results from primary studies of the ratio of the odds of a positive test result in patients with disease compared to the odds of the same test result in patients without disease.	

*based on Honest et al 29

Whilst conceptually straightforward, in practice, there is debate about how best to statistically summarise results from several primary test accuracy studies. (2, 22, 23, 24, 25, 26, 27, 28, 29). The lack of consensus was clearly evident in a recent survey of test accuracy reviews found in Database of Abstracts of Reviews of Effectiveness (DARE) from 1994-2000, which showed that pooled sensitivity or specificity was used in 58%, summary receiver operating characteristic (sROC) plots in 73%, pooled

predictive values in 18%, pooled likelihood ratios (LRs) in 22%, and pooled diagnostic odds ratio in 8% of the meta-analyses (29).

From meta-analysis, it should be possible to interpret the result in terms of clinical importance (not just statistical significance). In this respect, LR (25, 26, 27), is believed to represent an improvement over sensitivity, specificity, and predictive values. Many authorities considered pooling of sensitivity, specificity and predictive values as inappropriate as they do not behave independently. On the other hand, pooled (or summary) LRs can be used within a clinical context is shown in Table 4.

An example of clinical application of pooled likelihood ratios

Population & Outcome Measure	**Pretest Probability (95% CI)**	**Likelihood Ratio (95% CI)**	**Posttest Probability (95% CI)**
Delivery <34 weeks'gestation			
Positive test result	32.5 (24.2-40.8)	2.6 (1.8-3.7)	55.6 (43.4-67.3)
Negative test result	32.5 (24.2-40.8)	0.2 (0.1-0.5)	8.2 (3.1-20.1)
Delivery within 1 week of testing			
Positive result	6.6 (4.3-8.9)	5.0 (3.8-6.4)	25.8 (18.0-35.5)
Negative result	6.6 (4.3-8.9)	0.2 (0.1-0.4)	1.2 (0.4-3.1)

Based on Chien *et al 30*

However potentially misleading summary LRs might be obtained from pooling LRs obtained from studies with extreme and diverging prevalence. An alternative way of summarising the average performance of a dichotomous test from multiple studies (particularly those with different thresholds) is to produce a sROC plot. This test takes into account the variation in prevalence and is the preferred meta-analytic method of many experts. The area under curve of a sROC is a mathematical representation of the average accuracy of the test. However, unlike summary LRs, sROC does not lend itself readily to clinical application. Due to lack of consensus about the most appropriate summary measures it may be prudent to use both summary LRs and sROC for performing meta-analysis.

CONCLUSION

Many existing reviews of test accuracy offer limited guidance for practice because they do not apply a rigorous scientific methodology to limit bias in their assembly, appraisal, and synthesis of primary studies. In this paper, we have described methods for conducting a high quality test accuracy review. By understanding this

process, readers should be able to appraise test accuracy reviews with an informed mind thus minimising erroneous inferences.

REFERENCES

1. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. *European Journal of obstetrics gynaecology and reproductive biology* 2001;95:6-11.

2. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC et al. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of internal medicine* 1994;120:667-676.

3. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N et al. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. *JAMA* 1999;281:2029-2034.

4. Song F, Khan KS, Dinnes J, Sutton A. Asymmetric Funnel Plots and the Problem of Publication Bias in Meta-analyses of Diagnostic Accuracy. *International journal of epidemiology* 2002;31:88-95.

5. Devillé WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *Journal of clinical epidemiology,* 2000;53:65-69.

6. Clarke, M., Oxman, AD. Locating and Selecting Studies. Cochrane Reviewers' Handbook 4.1. In: *The Cochrane Collaboration*, 2000.

7. Khan, KS, Kavanagh J. *Clinical Governance Advice No 3: Searching for evidence. London Royal College of Obstetricians and Gynaecologists*. , 2001.

8. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1994;271:389-391.

9. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ* 1986;134:587-594.

10. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *Journal of general internal medicine* 1989;4:288-295.

11. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-1066.

12. Laird NM,.Mosteller F. Some statistical methods for combining experimental results. *International journal of technology assessment in health care* 1990;6:5-30.

13. Dickersin K,.Berlin JA. Meta-analysis: state-of-the-science. *Epidemiology reviews* 1992;14:154-176.

14. Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H. Considerations for the meta-analysis of randomized clinical trials. Summary of a panel discussion. *Controlled clinical trials.* 1989;10:254-281.

15. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials* 1986;7:177-188.

16. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in medicine* 1988;7:889-894.

17. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Annals of internal medicine* 1987;107:224-233.

18. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health technology assessment* 1998;2:1-276.

19. Bulpitt CJ. Subgroup analysis. *The lancet* 1988;2:31-34.

20. Oxman AD,.Guyatt GH. A consumer's guide to subgroup analyses. *Annals of internal medicine* 1992;116:78-84.

21. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *The lancet* 1998;351:123-127.

22. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Academic radiology* 1995;2 Suppl 1:S37-S47.

23. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarising diagnostic test performances: receiver-operating-characteristic-summary point estimates. *CMAJ* 1993;13:253-257.

24. Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests. *Screening and diagnostic tests: Recommended methods*. , 2000.

25. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994;271:703-707.

26. Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. *BMJ* 1997;315:540-543.

27. How to read clinical journals: II. To learn about a diagnostic test. *CMAJ* 1981;124:703-710.

28. Walter SD,.Jadad AR. Meta-analysis of screening data: a survey of the literature. *Statistics in medicine* 1999;18:3409-3424.

29. Honest, H, Khan, KS. *Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC health services research*. , 2001>.

30. Chien PF, Khan KS, Ogston S, Owen P. The diagnostic accuracy of cervico-vaginal fetal fibronectin in predicting preterm delivery: an overview. *British journal of obstetrics and gynaecology 1* 1997;104:436-444.

31. Clark TJ, Mann CH, Shah N, Khan KS, Song F, Gupta JK. Accuracy of outpatient endometrial biopsy in the diagnosis of endometrial hyperplasia. *Acta Obstetrica et gynecologia Scandinavia* 2001;80:784-793.

32. Divakaran TG, Waugh J, Clark TJ, Khan KS, Whittle MJ, Kilby MD. Noninvasive techniques to detect fetal anemia due to red blood cell alloimmunization: a systematic review. *Acta Obstetrica et gynecologia Scandinavia. 2* 2001;98:509-517.

33. Sackett DL, Straus S, Richardson WS, Rosenberg W, Haynes RB. 2nd edition. *Evidence-Based Medicine: How to practice and teach EBM*. , Edinburgh: Churchill Livingstone, 2000.

This document should be cited as: Honest H, Khan KS. Systematic reviews of test accuracy studies in reproductive health: The WHO Reproductive Health Library, No 5, Geneva, The World Health Organization, 2002 (WHO/RHR/02.1).