Principles for clinical evaluation of methods of perinatal monitoring

Adrian Grant

National Perinatal Epidemiology Unit, Radcliffe Infirmary, Oxford OX2 6HE.

Every clinical action has the potential to do harm. This is true not only for formal medical 'treatments' but also for the tests on which decisions to recommend these treatments are based. For this reason, the introduction of new tests into routine practice should be backed up by evidence that, overall, those tested are likely to benefit as a result. These considerations are particularly compelling in contexts, such as antenatal care, where the vast majority of pregnant women will have a good normal outcome whether they are tested or not.

The development of a new test for use in clinical practice can be conveniently (though somewhat simplistically) characterised into three phases, each of which require a particular approach to evaluation (Table I).

Table I - Phase of test development and clinical evaluation		
Phase	Development	Evaluation
1	Technical improvements in measurement	Does the test measure what it is supposed to measure? <u>Accuracy and repeatability</u>
2	Definition of abnormal test results for use in clinical practice	Does the test identify or predict the condition it is supposed to predict? <u>Validity</u>
3(a)	Choice of appropriate clinical management for abnormal and normal test results	Can the clinical use of the test (plus appropriate management) prevent the adverse outcome which it is supposed to prevent? <u>Efficacy</u>
3(b)	Choice of appropriate management policy	Does the defined policy for using the test perform better than alternative policies? <u>Effectiveness</u>

The aim of the first phase of development is to improve technical aspects so that the test measures the parameter it is supposed to measure as satisfactorily as possible. Evaluation of this phase assesses how well this was accomplished. Occasionally (for example, measurement of pH), there is an alternative means of measuring the parameter of interest which is "known to be correct". Comparison of the new test with such a standard measures its 'accuracy'.

For most tests, however, there is no alternative technique which perfectly measures the parameter of interest - placental blood flow and electro-mechanical cardiac time interval may be examples. Evaluation is then limited to estimation of the agreement between replicate measurements: 'repeatability'. There are two components of this: variation of a particular machine or observer (within-observer variation) which is largely random, and variation between machines or observers (between-observer

variation), which is largely systematic. Unlike random variation, systematic variation is not reduced by repeated measurement. It can, however, sometimes be reduced by altering other aspects of the way in which the test is used such as by more rigid standardisation of the methods with which the test is applied.

Good repeatability is encouraging but does not necessarily mean that the test is measuring what it is supposed to be measuring - it may still be inaccurate. Poor repeatability suggests that the test is less likely to be useful clinically and may indicate that there is still room for technical improvement in the test. This statement needs to be qualified, however, because the clinical usefulness of a test also depends on other considerations such as how the results will be used in clinical practice. This is the next stage of development.

What values for the test should be considered normal and which abnormal for use in clinical practice? There is a variety of ways of deciding this (1). Commonly, the distribution of test results in "normal" subjects is first described and extremes of this distribution, defined in statistical terms such as multiples of standard deviation or centiles, are labelled abnormal. A better approach which is often not possible (for reasons which are discussed later) is to relate the test results of a group of subjects to whether they actually have the condition which the test is to be used to identify. A cut-off is then chosen which best separates those with the condition from those without it.

Additional standardisation may be introduced at this point to eliminate other sources of unwanted variation, such as systematic variation due to gestational age or systematic variation between individuals.

The success of this phase is evaluated by how well the clinical condition of interest is identified or predicted by the test when used in the defined way. The jargon word for this is "validity". This is not the test characteristic which is of greatest interest to clinicians. They want to know how likely it is that a patient with an abnormal result actually has the condition (the Positive Predictive Value PPV, Table II, $a/a+b$).

TABLE II
Fourfold table comparing test results

		'Gold standard' ('the truth')		
		Subjects have the condition	Subjects do not have the condition	
'Test result' (conclusion drawn from the results of the test)	Positive Subjects appear to have the condition	True positives (a)	False positives (b)	$a + b$
	Negative Subjects appear not to have the condition	False negatives (c)	True negatives (d)	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

Stable properties: $a/(a+c)$ =sensitivity; $d/(b+d)$ =specificity.

Frequency-dependent properties: $a/(a+b)$ =positive predictive value; $d/(c+d)$ =negative predictive value; $(a+d)/(a+b+c+d)$ =accuracy; $(a+c)/(a+b+c+d)$ =prevalence.

Unfortunately, the predictive value of a test is not fixed but depends on how many of those tested actually have the condition ('the prevalence'). When performed in a high-risk group, a test has a better PPV than when performed in a low-risk group. The reason for this is that, generally speaking, it is the vertical relationships in Table II which are fixed. A test successfully identifies a fixed proportion of those who have the condition (sensitivity $a/a+c$) and correctly identifies a fixed proportion of those who do not have the condition as being 'normal' (specificity $d/b+d$). These parameters are the test's 'validity'. Knowledge of a test's sensitivity, specificity and the prevalence of the condition in those actually being tested allows a clinician to work out the more clinically useful indices - positive and negative predictive values (Table II).

It is important that the series of subjects used to measure the validity are not the same subjects as those on whom the definition of normal limits is based. If they are, the performance of the test is likely to look better than it actually is. Another pitfall is to measure validity at a time when clinical action is already being taken on the basis of the test results. If such intervention prevents the outcome which the test aims to identify, the test appears less valid than it really is. On the other hand, if the intervention causes the condition which the test aims to predict then, paradoxically, the test appears to be better than it actually is (see Grant & Mohide (1) for further discussion).

Measurement of validity in this way implies a knowledge of which of the subjects tested actually had the condition. Commonly this "truth" is not known. (What is 'fetal compromise', for example?) There is then no alternative to using surrogate measures which have poor or unknown relationships to the "truth". This may seriously distort the evaluation of a test's ability to predict or identify a condition, but what is worse is that it is never clear just how bad this distortion is.

In the final phase of development decisions are made about how the test will be used in clinical practice. Some test results may lead to immediate intervention while others (e.g. screening tests) may prompt a second test before treatment is considered.

Evaluation of this phase examines the extent to which the actions prompted by the results of a test actually improve outcome. In the rare instances where the validity of a test and the effectiveness of the treatment which it prompts are both known (e.g. AFP screening for neural tube defects) no further evaluation is needed because the consequence of using the test in practice can be extrapolated from this knowledge. But if uncertainty exists about either of these parameters (and this is true for all tests of fetal well-being) then additional evidence about the effects of using the test plus the intervention which it prompts is necessary. Ideally, this should come from randomised controlled trials. Evaluation of test and treatment together in this way circumvent the problems of evaluating each component in isolation.

There are two aspects of this phase which need to be considered (Table I). Firstly, can the test (plus appropriate management) have the desired effect on outcome (is it efficacious?) and secondly, does the test have the desired effect when used in everyday practice where alternative management policies may already exist (is it effective?).

The use of randomised controlled trials to minimise bias in the comparison of alternative management policies is now well known. What is less clearly recognised is the problem of imprecision in inadequately sized

trials. The size required to be confident of identifying a clinically important difference between two alternative managements is often more than a single centre can possibly generate. Multicentre collaboration seems unavoidable and there are now examples of successfully mounted multicentre trials in many fields including perinatal medicine. The obstacles to such collaboration which undoubtedly exist can be overcome. But if, for understandable reasons, the current fashion of mounting inadequately-sized, single centre trials continues, investigators must be encouraged to share protocols so that the data from these trials can later be pooled most efficiently and thereby greatly increase their scientific and clinical usefulness.

An example of this sort of analysis is the pooling of the results of four RCT's of antenatal cardiotocography (see also Grant & Mohide (1)). The numbers were small but in each of these trials (2-5) there were more perinatal deaths in the groups for which the traces were available to the clinicians. A review of these deaths based on the Wigglesworth classification (6) is shown in Table III.

Table III - Classification of perinatal deaths in four trials of non-stress antepartum cardiotocography		
Class of death	Revealed	Concealed
normally-formed		
macerated stillbirths	0	0
congenital malformations	5	5
conditions associated with prematurity	1	2
asphyxial conditions developing in labour	9	2
specific conditions other than above	4	0
Total	19	9

The most striking difference is in the category 'asphyxial conditions developing during labour'. A detailed review of the hospital notes of these cases, however, revealed only one case in which false reassurance of a 'normal' CTG trace is likely to have been implicated. Even so, this analysis provides little support for non-stress cardiotocography as used in these studies as a supplementary diagnostic test of fetal compromise.

References

- (1) Grant, A., Mohide, P. Screening and diagnosis in antenatal care. In: Enkin, M., Chalmers, I. (eds) Effectiveness and Satisfaction in Antenatal Care. Spastics International Medical Publications/William Heinemann Medical Books, 1982.
- (2) Brown, V.A., Sawyers, R.S., Parsons, R.J., Duncan, S.L.B., Cooke, I.D. The value of antenatal cardiotocography in the management of high-risk pregnancy: a randomized controlled trial. Br J Obstet Gynaecol 89 (1982) 716-722.

- (3) Flynn, A.M., Kelly, J., Mansfield, H., Needham, P., O'Connor, M., Viegas, O. A randomised controlled trial of non-stress antenatal cardiotocography. *Br J Obstet Gynaecol* 89 (1982) 427-433.
- (4) Lumley, J., Lester, A., Anderson, I., Renou, P., Wood, C. A randomised controlled trial of weekly cardiotocography in high-risk obstetric patients. *Br J Obstet Gynaecol* 90 (1983) 1018-1026.
- (5) Kidd, L.C., Patel, N., Smith, R. Personal communication.
- (6) Wigglesworth, J.S. Monitoring perinatal mortality, a pathophysiological approach. *Lancet* 2 (1980) 684-687.

Acknowledgements

I would like to thank the authors of the four trials of non-stress cardiotocography for generously providing me with unpublished data. The National Perinatal Epidemiology Unit is supported by a grant from the Department of Health and Social Security.