How To:

Analyse a Screening or Diagnostic Study

Author:

Alex J Mitchell

Consultant in Liaison Psychiatry, Leicester General Hospital, Leicester LE5 4PW and Honorary SnR Lecturer in Liaison Psychiatry, Department of Cancer & Molecular Medicine, Leicester Royal Infirmary LE1 5WW. Email ajm80@le.ac.uk

Contents:

Useful Abbreviations

Se	– Sensitivity
Sp	– Specificity
PPV	– Positive predictive value
NPV	– Negative predictive value
PSI	– Predictive summary index
NND	– Number needed to diagnose
NNS	– Number needed to screen

1. Overview of Case-Identification Terminology

Screening and case-finding refer to making a diagnosis, either clinically or by some kind of test or tool. Its easy to get confused about screening definitions because authors use terms in non-standardized ways. At its core we are talking about methods to help clinicians identify a disorder or disease and I prefer to call this process case-identication or simply diagnosis. In an ideal world we want to correct spot all "cases" and correctly identify all "non-cases". In effect we want good overall accuracy. However sometimes a clinician (or test) does well in ruling-in cases but poor at ruling out healthy people. Consider an example of 100 men who attend primary care. Hypothetically 20 have an alcohol use disorder and 80 do not. Clinician A might spot 18/20 (sensitivity = 90%) of cases but only 40/80 (specificity = 50%) non-cases. A better way of thinking about diagnostic accuracy is in terms of rule-in and rule-out performance rather than sensitivity and specificity. Sensitivity and specificity are somewhat abstract for clinicians and are really only useful when then approach 100% (when the Sacket addage of $SPIN_{specificity}$ and $SNOUT_{sensitivity}$ apply). Clinicians ability to spot true cases as a proportion of all their attempts is called the *positive predictive value* (PPV). PPV is essentially a measure of case-finding ability. Clinicians ability to spot (true) non-cases as a proportion of all their attempts is called the *negative predictive value* (NPV). NPV is a measure of screening accumen (box 1).

Box 1- Pragmatic Definitions of Case-Identification

Screening

The application of a diagnostic test or clinical assessment in order to optimally rule-out those without the disorder with minimal false negatives (missed cases).

Screening studies are often performed as a broad population strategy as a first step.

Case-Finding

The application of a diagnostic test or clinical assessment in order to optimally identify those with the disorder with minimal false positives.

Case finding studies are often performed in a selected population at high risk the condition

In an epidemiological sense *screening studies* are those where a test is applied to those at low or modest risk, that is low prevalence settings. The aim here is to exclude a larger number of clear non-cases. A first stage screen may not have perfect PPV but it should have high NPV. This is because those ruled-out are unlikely to get a second examination. In an epidemiological sense *case-finding studies are* usually applied in high prevalence settings. The assumption is that the case-finding method is accurate enough to spot cases and non-cases without the need for re-testing. However this is very much an assumption that should be tested and the possibility of further testing not rejected without good reason.

A *screening programme* is the widespread distrubtion of a screening test and screening support system across a health care system. Many staff may be involved in a screening programme. Ideally the impact of the screening programme should be monitored and the programme adjusted accordingly (see below). The effort required to implement an efficient screening programme should not be underestimated even (or especially) in a low resource environment.

2. Designing Studies to Test New Screening Methods

Just as with the introduction of a new drug, a new screening test (and even more so a screening programme) cannot be assumed to be *efficacious* without careful testing. In fact, like a new drug, a screening test may have unforeseen adverse consequences or it may simply be ignored by health professionals. To be *effective* a screening programme must have reasonably high accuracy, very high acceptability and good uptake and a association with subsequent interventions that improve quality of care.

Despite the huge promise of better screening methods for psychological disorders the evidence that any particular method improves patient outcomes is often lacking. The problem lies with a poverty of studies that have examined implementation of screening as opposed to testing just the accuracy (or more correctly diagnostic validity) of a given tool. The evaluation of screening methods should be viewed in a wider context of tool development (table 1). In the pre-clinical phase the tool itself is developed, often by borrowing items from existing scales and usually by consensus rather than by scientific testing. No matter how plausible the new tool, it is essentially untested at this stage. In phases I and II preliminary testing occurs, ideally in a clinically representative sample with several competing comparison groups. An example would be the ability of a tool to detect major or minor depression in cancer compared to those with no symptoms of depression and those with subsyndromal symptoms alone. This "diagnostic validity" testing is an important step which shows the maximum potential of a scale. However it does not show the real-world ability of the test. By analogy, a phase II drug trial may demonstrate potential efficacy of a drug but the effectiveness in clinical practice is unknown to this stage.

The next steps are probably the most important but easily overlooked. In phase III of screening tool development a randomized control trial (RCT) is conducted to directly compare the results of clinicians using the new tool with those using either an older established method or unassisted "diagnosis as-usual" (or ideally both). This is akin to the drug RCT and the outcome of interest is the number of additional cases correctly diagnosed or ruled out compared with assessment as usual. In the final step, phase IV, the success or otherwise of the new method is monitored as it is rolled out in the field. In short the question here is how much does use of the tool by clinicians influence the outcome of patients. This clearly depends on how well the programme is accepted by clinicians (uptake) but also how well clinicians use additional identification to help patients. Ultimately the value of a tool must be proven in the clinical environment by comparison against either an established tool or clinical skills alone.

Table 1. Stages in the Evaluation of the Screening Tool or Diagnostic Test

Stage	Type	Purpose	Description
Pre-clinical	Development	Development of the proposed tool or test	Here the aim is to develop a screening method that is likely to help in the detection of the underlying disorder, either in a specific setting or in all setting. Issues of acceptability of the tool to both patients and staff must be considered in order for implementation to be successful.
Phase I_screen	Diagnostic validity	Early diagnostic validity testing in a selected sample and refinement of tool	The aim is to evaluate the early design of the screening method against a known (ideally accurate) standard known as the criterion reference. In early testing the tool may be refined, selecting most useful aspects and deleting redundant aspects in order to make the tool as efficient (brief) as possible whilst retaining its value.
Phase II_screen	Diagnostic validity	Diagnostic validity in a representative sample	The aim is to assess the refined tool against a criterion (gold standard) in a real world sample where the comparator subjects may comprise several competing condition which may otherwise cause difficulty regarding differential diagnosis.
Phase III_screen	Implementation Study	Screening RCT; clinicians using vs not using a screening tool	This is an important step in which the tool is evaluated clinically in one group with access to the new method compared to a second group (ideally selected in a randomized fashion) who make assessments without the tool.
Phase IV_screen	Implementation Roll-out	Screening implementation studies using real-world outcomes	In this last step the screening tool /method is introduced clinically but monitored to discover the effect on important patient outcomes such as new identifications, new cases treated and new cases entering remission.

3. Analysing Accuracy from a Simple Screening Study

3.1 Simple (One-Sided) Measures of Accuracy

Attempts to separate those with a condition from those without on the basis of a test or clinical method are best represented by the 2x2 table which generates sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV) (figure 1).[1] It is important to understand the difference between looking vertically across cells and looking horizontally. Vertically, the denominator is the number of cases with or without the condition, a number which is unknown to the clinician. Horizontally, the dominator is the number of positive or negative screens, a number that is known and hence the reason why positive predictive value (PPV) and negative predictive value (NPV) are often more important than Se and Sp. Performance of most tests varies with the baseline prevalence of the condition. Put simply it is simple to detect cases when nothing but cases exist (prevalence = 100%) but conversely it is hard for to detect cases when such cases are very rare.[2] Rule in and rule out accuracy should be considered independent variables although a test may perform well in both directions. Rule-in accuracy is best measured by the PPV but a high Sp also implies few false positives and hence any positive screen will suggest a true case. [3] Rule-out accuracy is best measured by the NPV where the denominator is all who test negative but again if the Se is high there will be few false negatives and hence any negative implies a true non-case (box 2).[3]

Figure 1. Generic 2x2 Table

	Gold Standard Disorder	Gold Standard No Disorder	
Test +ve	A	B	A/A + B **PPV**
Test -ve	C	D	D/C + D **NPV**
Total	**A/** A + C **Se**	**D/** B + D **Sp**	

3.2 Summary Measures of Diagnostic Accuracy

Optimal accuracy is often achieved by choosing one test for rule-in (case-finding) and another for rule-out but not uncommonly where resources are limited only a single test can be applied and this single test must perform as well as possible in both directions. Here so called summary statistics are used to test accuracy. These use a combination of either Se and Sp or PPV and NPV. Reciprocal measures are also becoming more common and offer a "number needed" estimate. All such methods work well when the optimum cut-off is known or in binary (yes/no) tests, but where performance varies according to the cut-off threshold then sensitivity versus specificity for each cut-off generates a receiver operating characteristic (ROC) curve and the area under the curve gives a measure of the overall performance. More advanced methods are needed when multiple tests need to be compared (each with different Se and Sp values). For example results can be combined in a summary receiver operator curve (sROC),[4] but increasingly clinicians prefer summary statistics which generate clinically meaningful results. These are discussed below.

Box 2. Basic Measures of Diagnostic Accuracy

Sensitivity (Se) a/(a + c)
A measure of accuracy defined the proportion of patients with disease in whom the test result is positive: a/(a + c)

Specificity (Sp) d/(b + d)
A measure of accuracy defined as the proportion of patients without disease in whom the test result is negative

Positive Predictive Value a/(a+b)
A measure of rule-in accuracy defined as the proportion of true positives in those that screen positive

Negative Predictive Value c/(c+d)
A measure of rule-out accuracy defined as the proportion of true negatives in those that screen negative

Youden's J and Number Needed to Diagnose

Youden's J is based on the characteristics of sensitivity and specificity as follows: $J = sensitivity + specificity - 1$].[5] If a test has no diagnostic value, sensitivity and specificity would be 0 and hence J=-1, a test with modest value where

sensitivity and specificity are both 0.5 would give a *J* of 0. If the test is perfect then *J* =+1. Youden's index is probably most useful where sensitivity and specificity are equally important and where prevalence is close to 0.5.

The reciprocal of Youden's *J* was originally suggested as a method to calculate the number of patients that need to be examined in order to correctly detect one person with the disease. [6] This has been called the *number needed to diagnose* (NND). Thus NND = 1/[Sensitivity - (1 - Specificity)]. However the NND statistic is hampered by the same issues that concern the Youden score, namely that it is insensitive to variations in prevalence and subject to confusion in cases where sensitivity is high but specificity low (or visa versa). Additionally the NND becomes be artificially inflated as the Youden score approaches 0 and this is misleading because the Youden varies between -1 and +1 not +1 and 0. In short the reciprocal of Youden's *J* is not a clinically meaningful number.

The Predictive Summary Index

In most clinical situations when a diagnostic test is applied, the total number of positive results (TP+FP) and negative test (TN+FN) results is known although the absolute number of TP and TN is not. In this situation the accuracy of such a test may then be calculated from the positive predictive value (PPV) and negative predictive value (NPV). Unlike sensitivity and specificity, PPV and NPV are measures of discrimination (or gain). The gain in the certainty that a condition is present is the difference between the post-test probability (the PPV) and the prior probability (the prevalence) when the test is positive. The gain in certainty that there is no disease is the difference between post-test probability of no disease (the NPV) and the prior probability of no disease (1-prevalence). This is best illustrated in a Bayesian plot (figure 2). In the Bayesian plot shown in figure 2 the pre-test probability is plotted (black line) and the post-test probability the dotted line. Thus the overall benefit of a test from positive to negative is a summation of [*PPV* - Prevalence] + [NPV - (1 - Prevalence)] = PPV+NPV-1. This is the predictive summary index (PSI). Where prevalence varies, optimal gain is achieved when the prevalence of the condition is 50%, as shown in the figure 2.

Fraction Correct and Number Needed to Screen

One approach to calculating accuracy is to measure the overall fraction correct. The overall fraction correct is given by A+D/A+B+C+D (figure 1). 1 minus the fraction correct (1-FC) is the fraction incorrect. The fraction correct can be useful because it reveals the real number of correct vs incorrect identifications. The fraction correct minus the fraction incorrect can serve as a useful "identification index" which can be converted into a number needed to screen (below). Fraction correct is also attractive because the performance of two tests may be directly compared using a simple Chi^2 statistic and can support a meta-analysis of diagnostic methods.

The number needed to screen is based on the difference between the real number of correctly diagnosed and incorrectly diagnosed patients. The number needed to screen = 1 / FC – (fraction incorrect) or 1/ Identification index. Unlike the Youden score or NND, the clinical interpretation of the NNS is clinically meaningful. It is the actual number of cases that need to be screened to yield one additional correct identification (case or non-cases) beyond those misidentified.

Take a hypothetical example of a new screening test for depression tested in 100 with the condition and 1000 without which yields a Se of 0.90 and a Sp of 0.50. The Youden score is thus 0.4 and the NND 2.5 suggesting 2.5 individuals are needed to diagnose 1 person with depression. In fact, out of every 100 applications of the test there would be 9 people with depression (prevalence x 100) of whom 90% would be true positives (=8.2), and 81 without

depression (1-prevalence x 100) of whom 50% would negatives (=45.5). In this example there would be 53.6 true cases per 100 screened (fraction correct per 100 cases) but at the expense of 46.4 errors (fraction incorrect) per 100 screened; a net gain of 7.3 identified cases per 100 screened. Thus, the number needed to screen (NNS) would be 13.75 applications of the test to yield one true cases *without error*.

Confusingly there is another definition of number needed to screen which I believe is best called "*screening sensitivity*" as opposed to "*diagnostic sensitivity*". The diagnostic sensitivity is the number of true positive identifications as a proportion of all cases applied a diagnostic test AND a gold standard test. Some calculate the proportion of true positives from all those initially recruited to a screening study. As many may be recruited who are not cases and many may not agree to all tests the "screening sensitivity" is usually very low.

4. Analysing Applicability from a Simple Screening Study

4.1 Clinical Utility Index (Occurrence & Discrimination combined)

It should be clear that Se and Sp are essentially measures of occurrence. If 8 out of 10 people with true anxiety score positive on the distress thermometer then the sensitivity of the distress thermometer for anxiety is 80%. Contrastingly PPV and NPV are essentially measures of discrimination. If nine of those with anxiety to every one without anxiety scores positive on the distress thermometer then the PPV will equal 90%. These two attributes, occurrence and discrimination should both be high for an ideal test. Consider the example of a new "Depression Thermometer" test which if positive has a 90% PPV but is only positive in half of depressed individuals (Se 50%). Clinically relevant rule in accuracy would be product of the PPV and Se. This called the +ve utility index (UI+ = Se x PPV). Similarly clinically relevant rule out accuracy would be product of the NPV and Sp. This called the -ve utility index (UI- = Sp x NPV). The utility index can be considered a measure of the clinical value of a diagnostic test and can be graded using the following scale: < 0.2 poor, > 0.2 <0.4 fair, > 0.4 ≤0.6 moderate, > 0.6 ≤0.8 good and > 0.8 <1 excellent.

4.2 Acceptability and Clinical Feasibility

Even a test with high performance measures cannot be assumed to be beneficial. A number of factors determine whether a screening tool can be usefully translated into a screening programme. Guidelines from the UK National Screening Committee are helpful here (box 3). Feasibility asks whether a tool is practical both in application and scoring to gain acceptance by health professionals and patients. This has been poorly studied in relation to depression severity scales. However, in one example Bermejo et al (2005) looked at attitudes to the Patient Health Questionnaire (PHQ9) in primary care in Germany.[7] In this study 1034 patients from 17 GPs were enrolled and both patients and health professionals asked about acceptability. Patients found the instrument highly acceptably but 62.5% of the GPs felt that the questionnaire as too long and 37.5% too time-consuming, even though it typically took 1-2 minutes. 50% of the GPs rated the PHQ as an impediment to daily practice and 75% thought it was impractical compared with only 25% of patients. One proxy for feasibility is willingness of clinicians to use the test. Any screening roll out will be compromised if front line staff find the tool too difficult to administer or score.

See also section 8. engaging individuals in screening.

Box 3: Simplified UK National Screening Committee Guidelines

The condition should:
- Be an important health issue
- Have a well-understood history, with a detectable risk factor or disease marker
- Have cost-effective primary preventions implemented.

The screening tool should:
- Be a valid tool with known cut-off
- Be acceptable to the public
- Have agreed diagnostic procedures.

The treatment should:
- Be effective, with evidence of benefits of early intervention
- Have adequate resources
- Have appropriate policies as to who should be treated.

The screening program should:
- Show evidence that benefits of screening outweighing risks
- Be acceptable to public and professionals
- Be cost effective (and have ongoing evaluation)
- Have quality-assurance strategies in place.

Adapted from: *UK National Screening Committee Criteria for appraising the viability, effectiveness and appropriateness of a screening programme* http://www.nsc.nhs.uk/pdfs/criteria.pdf

4.3 Anticipated Screening Yields

One important way to calculate the success of a screening programme is in terms of screening yield. If the test is a simple one applied to everyone and with known sensitivity and specificity then the yield is simple to calculate. However real world screening programmes often have two or more steps (an algorithm) and then not all those that screen positive are cases and not all those that screen positive want and accept the help that is offered.

In the following example screening yields from a two-step algorithm are shown. The screening instrument has a 78% sensitivity (Se) and 67% specificity (Sp) and is applied in the first step to all 1000 participants. In those that screen negative no further test is recommended. But in those who screen positive a more lengthy (but more accurate) case-finding instrument with 80% sensitivity and 80% specificity is applied. The true positive (TP), true negative (TN) correct identifications versus false positive (FP), false negative (FN) incorrect identifications are calculated cumulatively after both steps, considering all possible pathways. It can be shown that this algorithm yields an overall sensitivity of 62% and specificity of 93%, a PPV of 80% and an NPV of 85%. Thus the two-step process improves on the single step by taking the PPV from 50% to 80% and the NPV reduces from 88% to 85% with an accompanying gain in overall accuracy but at a cost of a step 2 applied to an addition 465 people.

The screening yield can be calculated as a proportion of all those who enter the study providing all such individuals also receive the criterion test (gold standard). If the yield is predicted on the basis of previous diagnostic accuracy findings (previously known sensitivity and specificity) then the result are “anticipated” or “hypothetical” yields.

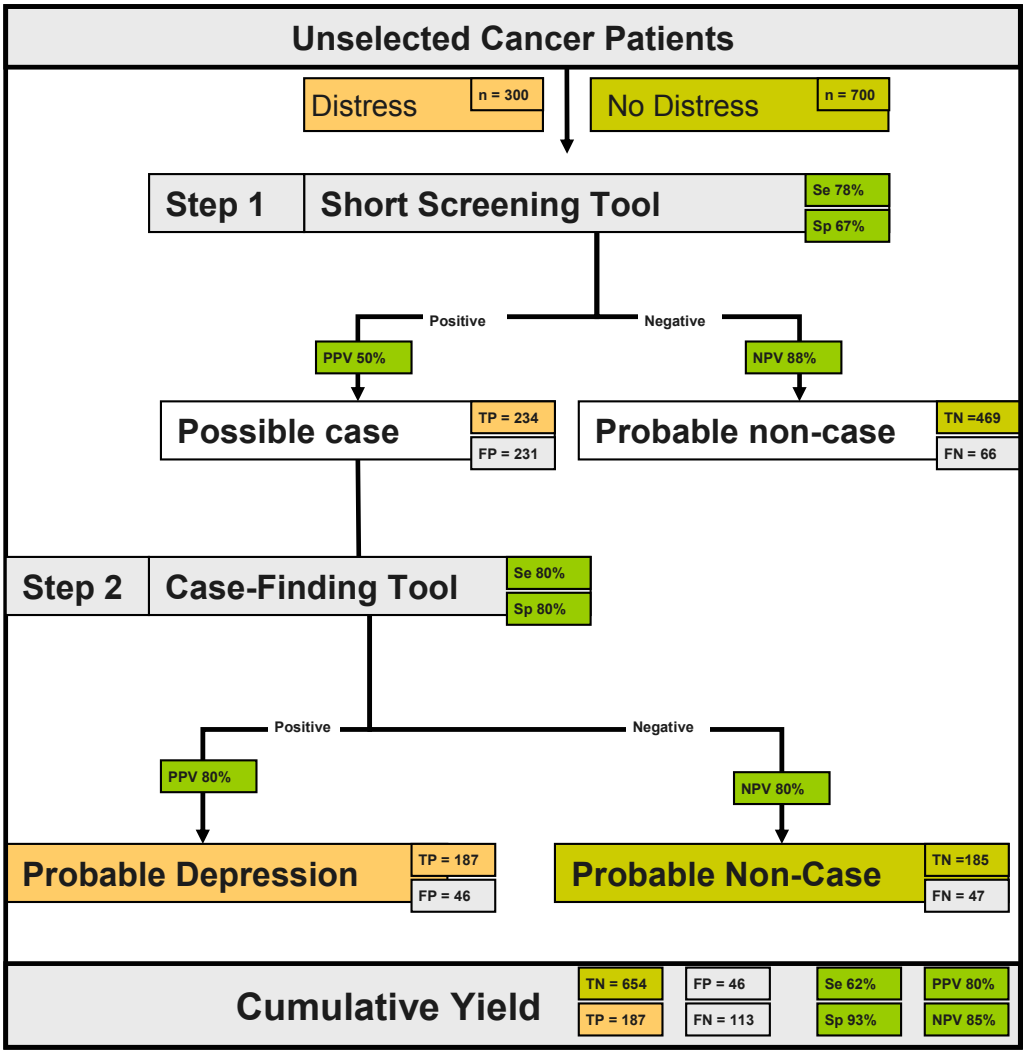


Table 4 shows a simple illustration of yields from multiple application of two different tests in various combinations.

5. Converting Screening Tests into Screening Programmes

Screening tests are usually examined in individual research studies but it is screening programmes that are applied in wide scales clinical studies. Ideally no aspect of screening programme success should be assumed, indeed even the most efficient test may fail roll-out in clinical practice. Additionally in one centre there may be many types of patient who might struggle with a screening programme (box 4).

Box 4: Groups that may Struggle with Screening

Older patients
Younger patients
Those with visual impairment
Those with cognitive impairment
Those with low educational attainment
Those with poor reading ability
Individuals with very high distress
Individuals with high levels of anger
Individuals who fail to attend
Individuals with low trust in health professionals
People who dislike the implementation method

Roll-out usually means that many staff would be expected to use the test to aid in the clinical assessment and diagnosis. Such staff may need to gain basic familiarity with the method or may require more advanced skills through training. Thought needs to be given to the location of the screen, the method of application (eg pencil & paper or computer or touch-tablet) and the timing and number of applications. Much work may be required to assist frontline staff with the roll-out of a new method of screening for psychosocial distress. Not infrequently cancer staff may have no inherent interest in psychosocial issues. Regarding the issue of timing some prefer routine screening others targeted (selected) screening. Routine screening has the advantage of not missing low risk individuals who might nevertheless be in need of help. Targeted screening may be more efficient and have a greater yield due to higher underlying prevalence. How often should a tool be applied? I think the simplest answer is “as often as possible” whilst not compromising staff involvement or patient acceptability. However screening at fixed time-points also has the advantage of ensuring everyone receives at least one test.

Costs of roll-out could vary from nothing at all where existing staff do all the work on a good will basis to millions of dollars/euros for a national distribution using resource intensive methods. Many national programmes for cancer screening (prostate, bowel, cervical) cost tens of millions.

Table 2 Simplified Criteria for Evaluation of Screening Programmes (see appendix)	
1	There must be high quality evidence from RCTs that the screening programme improves outcomes
2	The information provided about the “test” must be of value and readily understood by participants
3	There should be evidence that the complete screening programme is acceptable (clinically, socially and ethically) to health professionals and the public
4	The benefit of the screening programme should outweigh the physical and psychological harm
5	The opportunity cost of the screening programme should be economically balanced against expenditure of medical care as a whole (value for money)
6	There must be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards
7	Adequate staffing should be available prior to launch of the screening programme
8	All other treatment options should have been explored
9	Evidence based information (explaining the possible consequences of testing, diagnosis and treatment) should be available in order to help participants make an informed choice.
10	Public pressure to widen the eligibility criteria should be anticipated and decisions scientifically justifiable.

6. Monitoring Roll-Out Success of Screening Programmes

Several important outcomes can be measured as markers of success. These can be divided into staff reported measures and patient reported measures.

6.1 Staff Outcomes

It is useful to measure frontline clinicians opinion on the screening programme. First does the tool help the front-line staff in the diagnostic decisions? To test this thoroughly an RCT is needed but a before and after design or centre A vs centre B design can also be informative. Second does the tool help clinicians carry out appropriate treatment? Again the above designs apply but clinician reported practices can also be helpful. Third is the tool perceived as a burden (especially after cumulative applications). Clinicians may initially be willing to pilot the tool, but after some time motivation may subside. The tool may have to be revised, data collection simplified. At the end of the study it may be possible to stop collecting evidence and hence the programme can often be much simplified.

6.2 Patient Outcomes

The patient is at the centre of the screening programme and should be involved in its evaluation. First does the patient feel the clinical experience was better with the tool? An RCT can ascertain whether those in the programme have higher satisfaction than those without. However note satisfactions scores in the non-active (TAU) arm may already be high so elucidating differences may require a large study. Second does the patient receive better services under the active arm? Patients should receive better detection and more offers of treatment and more monitoring and also ideally healthy individuals should receive less false positive type interventions. Third and most importantly are patients actually improving at a faster rate or in greater proportion in centres using the tool? The latter may require prolonged follow-up. In addition, although difficult to measure, is there any evidence for extinction in (any) therapeutic differences between arms with time?

Box 5: Outcomes that can inform Screening programme Implementation

Screening uptake
Diagnostic sensitivity of staff
Diagnostic specificity of staff
Staff satisfaction
Staff burden
% of Staff offering treatment
% of patients offered treatment
Patient satisfaction
Patient burden
Patient wellbeing (HRQoL)
Patient distress / depression
Cost and cost-benefit

7. Bias in Screening Programmes

Various factors can cause the screening test to appear more successful than it really is. A number of different biases, inherent in the study method, will skew results.

7.1 Lead time bias

By screening, the intention is to diagnose a disease earlier than it would be without screening. Without screening, the disease may be discovered later once symptoms appear. Even if in both cases a person will die at the same time, because we diagnosed the disease early with screening, the survival time since diagnosis is longer with screening. Unless lead time is accounted for, comparisons of survival rates in screened and unscreened populations will be misleading. There always is a bias toward better survival rates in the screened group because the length of the lead time moves the point at which survival begins to be measured forward. Thus, it is possible that earlier detection only moves forward the time of a patient's diagnosis, without moving back the time of death. If lead time bias is present, screen-detected cancers appear to have better survival, but in fact death occurs at the same point it would have without screening

7.2 Length time bias

Many screening tests involve the detection of cancers. It is often hypothesized that slower growing tumors have better prognosis than tumors with high growth rates. Screening is more likely to detect slower growing tumors (due to longer pre-clinical sojourn time), which may be less deadly. Thus screening may tend to detect cancers that would not have killed the patient or even been detected prior to death from other causes.

7.3 Selection bias

Not everyone will take up a screening program. There are factors that differ between those willing to get tested and those who are not. If people with a higher risk of a disease are more eager to be screened then a screening test will look worse than it really is. Selection bias may also make a test look better than it really is. If a test is more available to young and healthy people (for instance if people have to travel a long distance to get checked) then fewer people in the screening population will get ill, and the test will seem to make a positive difference.

7.4 Over-diagnosis bias

Because screening is more likely than symptom recognition to yield lesions that might never become clinically significant cancers, survival statistics for screening detected cancers may be inflated. Over-diagnosis may be suspected if an imbalance in a cohort persists after an extended period of follow-up between the incidence rate in a screening program and the expected incidence rate in the absence of screening

8. Engaging People in Screening Programmes

8.1 Screened Participants

Many people may be reluctant to participate in even simple screening programmes. From the healthy individuals perspective there may be no benefits and all risks (zero sum game). Even if a diagnosis is revealed this is not necessarily welcome news. There have been many studies of predictors of patient participation in screening studies. In general it is desirable for the individual to be involved in the self-management of his or her own health. The key is often to tie education and information about the underlying disease and allying the screen with effective and acceptable treatment. Some physicians already apply the Five A's construct (Assess, Advise, Agree, Assist, Arrange) in order to bring about change. Motivational interviewing describes these principles: 1. Express empathy. 2. Examine perceived discrepancies between current behaviour and future goals. 3. Roll with resistance, not against it. 4. Support self-efficacy. Some additional tips are shown in figure 3.

8.2 Health Professionals

Just as patients are reluctant staff may be equally ambivalent about engaging in an unfamiliar screening programme. To them they will often say it takes too much time and we are doing this anyway (eg looking for depression). Clearly if they can present evidence of high clinical detection rates they may be correct, but this is rarely the case. The screening programme exists to catch all cases of erroneous diagnosis, particularly in clinicians with below average confidence or below average skills. Occasionally skilled staff must agree to a protocol driven policy to all not-so-skilled colleagues to also engage. Engaging staff is helped by prior education, direct hands-on participation, feedback of results, observation of patient benefit and at all time maintaining low staff burden of the programme.

9. Conclusions

Screening and case-finding of both types of case-identification or diagnosis. Diagnostic tools may be clinical acumen or biomarkers. The development and evaluation of diagnostic (screening) programmes should be approached using the same high standard that is afforded to the evaluation of new drugs. For example a screening RCT would involve evaluation of diagnoses in one group of patients accessed with the new tool compared to a second group randomized to assessment using conventional methods. The aim is to preserve accuracy but deliver it in the briefest, most efficient package. Often the rate limiting step in the effectiveness of any test or tool is its acceptability (for discussion see Mitchell and Coyne 2008).[8] Acceptability to health professional influences clinicians' willingness to apply a screening test and acceptability to patients influences a persons willingness to attend for screening. A small local implementation programme may be performed on a good-will basis with simple before and after monitoring of patient and staff satisfaction. Larger scale roll-outs should be tested in a randomized study where the overall benefit to patients is compared. Often the large potential of screening tests to improve detection are not translated into a successful programme because despite increased recognition of cases staff do not offer treatment or follow-up sufficiently. Building in these elements into a screening programme increases the likelihood of improving the overall quality of care offered and ultimately influencing the wellbeing of patents.

Table 3 – Monitoring Success of a Screening Programme (http://www.screening.nhs.uk/criteria)
Adapted from Wilson, JMG and Junger, G. The Principles and Practice of Screening for Disease. Geneva: World Health Organization, 1966

Stage	Type
	The Condition
	# The condition should be an important health problem # The epidemiology and natural history of the condition, including development from latent to declared disease, should be adequately understood and there should be a detectable risk factor, disease marker, latent period or early symptomatic stage. # All the cost-effective primary prevention interventions should have been implemented as far as practicable. # If the carriers of a mutation are identified as a result of screening the natural history of people with this status should be understood, including the psychological implications.
	The Test
	# There should be a simple, safe, precise and validated screening test. # The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed. # The test should be acceptable to the population. # There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals. # If the test is for mutations the criteria used to select the subset of mutations to be covered by screening, if all possible mutations are not being tested, should be clearly set out. The Treatment # There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment. # There should be agreed evidence based policies covering which individuals should be offered treatment and the appropriate treatment to be offered. # Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.
	The Screening Programme
	# There should be evidence from high quality Randomised Controlled Trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (eg. Down's syndrome, cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened. # There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public. # The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatment). # The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (ie. value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource. # All other options for managing the condition should have been considered (eg. improving treatment, providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available. # There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards. # Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme. # All other options for managing the condition should have been considered (e.g. improving treatment, providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available. # Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice. # Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public. # If screening is for a mutation the programme should be acceptable to people identified as carriers and to other family members.

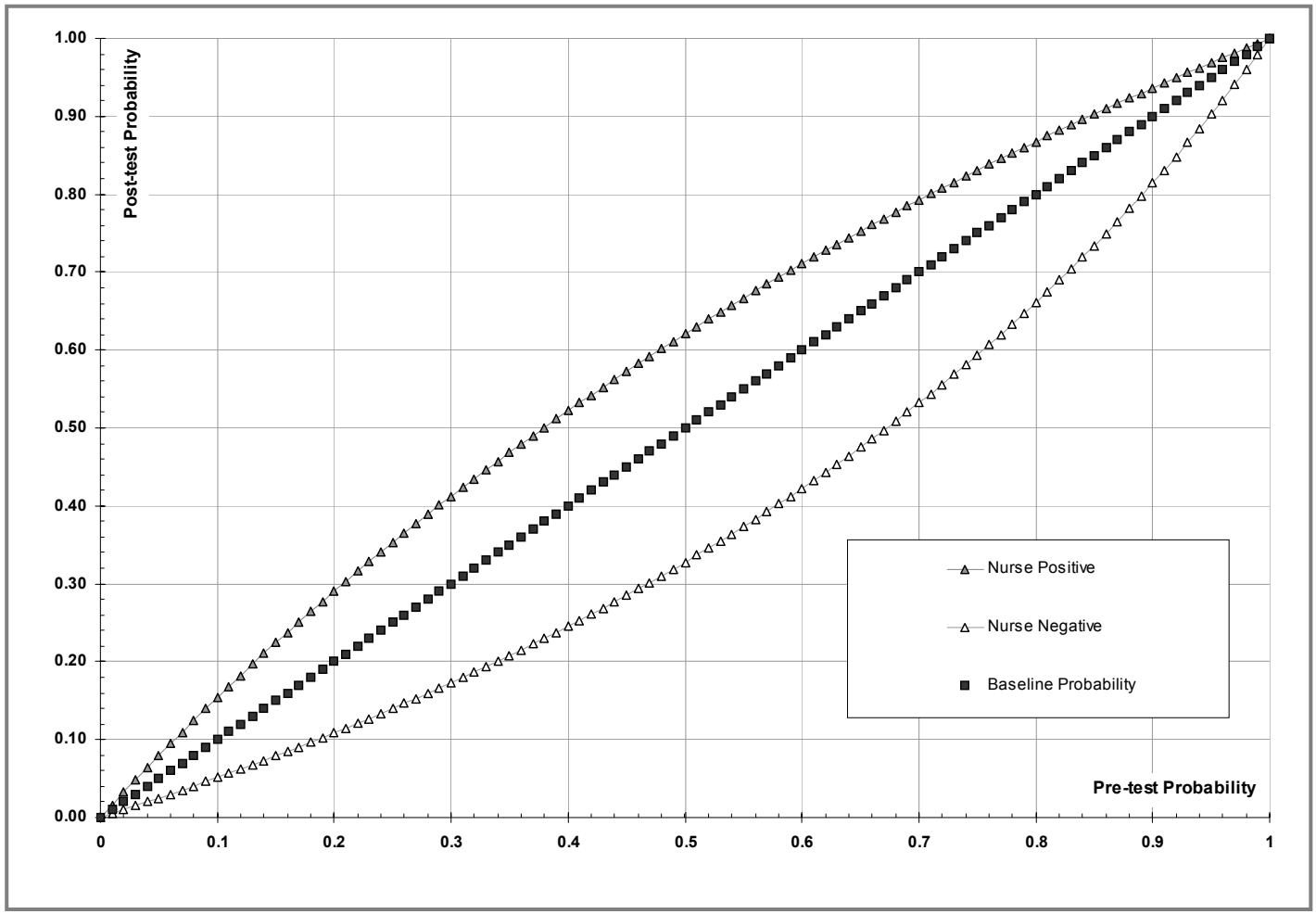
Table 4. Yields Calculated from Multiple Applications of Screening Tests

Step	Distressed	True Positives	Sensitivity	True Negative	Non-Distressed	True Negatives	Specificity	False Positives	PPV	NPV	FC
Test A											
Test A Applied Once	300	234	0.8	66	700	469.0	0.7	231.0	0.50	0.88	0.70
Test A Applied Twice	300	182.52	0.6	117.48	700	623.8	0.9	76.2	0.71	0.84	0.81
Test A Applied Three Times	300	100	0.3	200	700	674.8	1.0	25.2	0.80	0.77	0.77
Test A then Test B											
Test B Applied Once	300	210	0.70	90	700	567	0.81	133	0.61	0.86	0.78
Test A Applied then Test B	300	163.8	0.5	136.2	700	656.1	0.9	43.9	0.79	0.83	0.82
Test A Applied then Test B Twice	300	100	0.3	200	700	691.7	1.0	8.3	0.92	0.78	0.79

Test A Sensitivity = 80% and Specificity = 70%

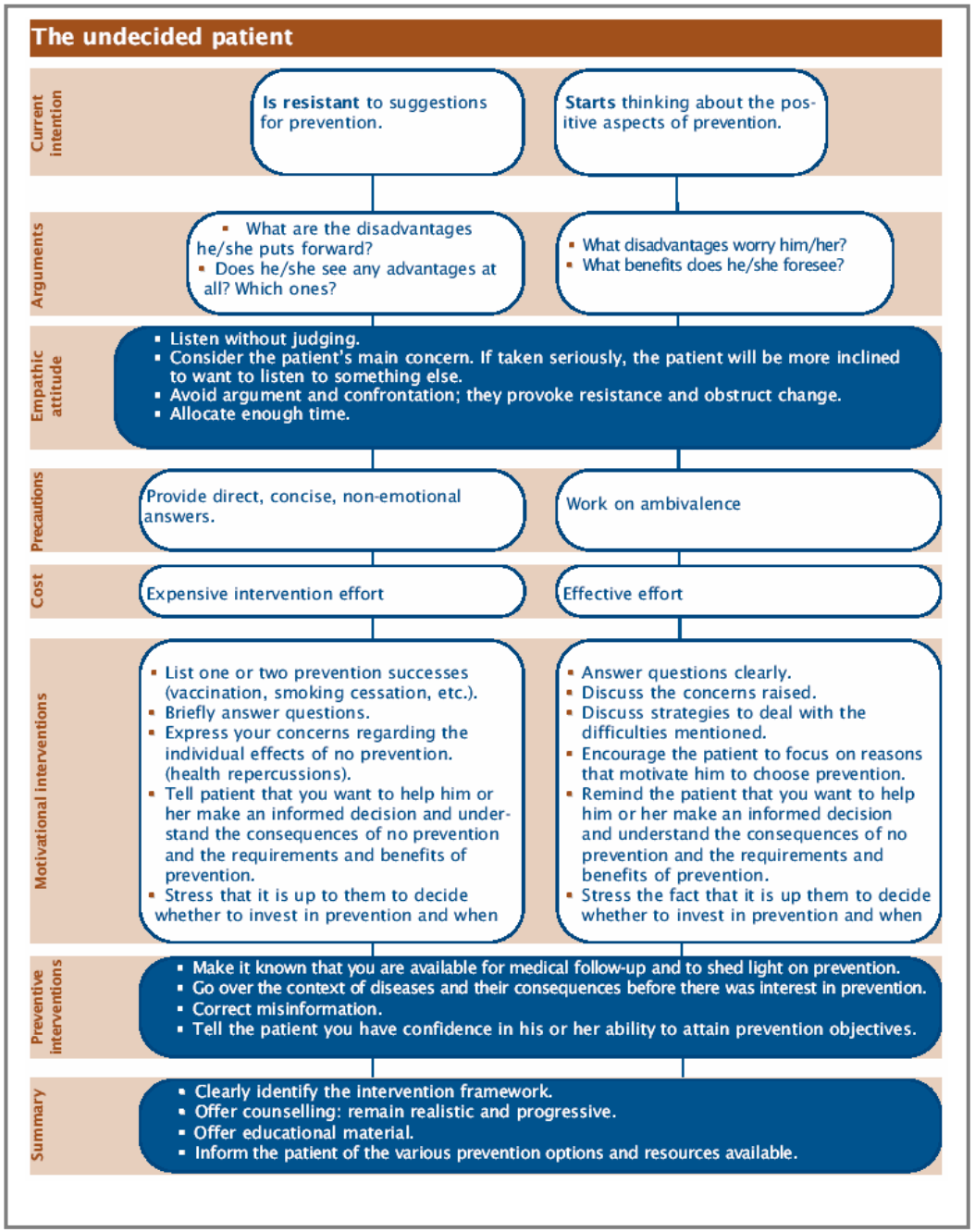
Test B Sensitivity = 70% and Specificity = 81%

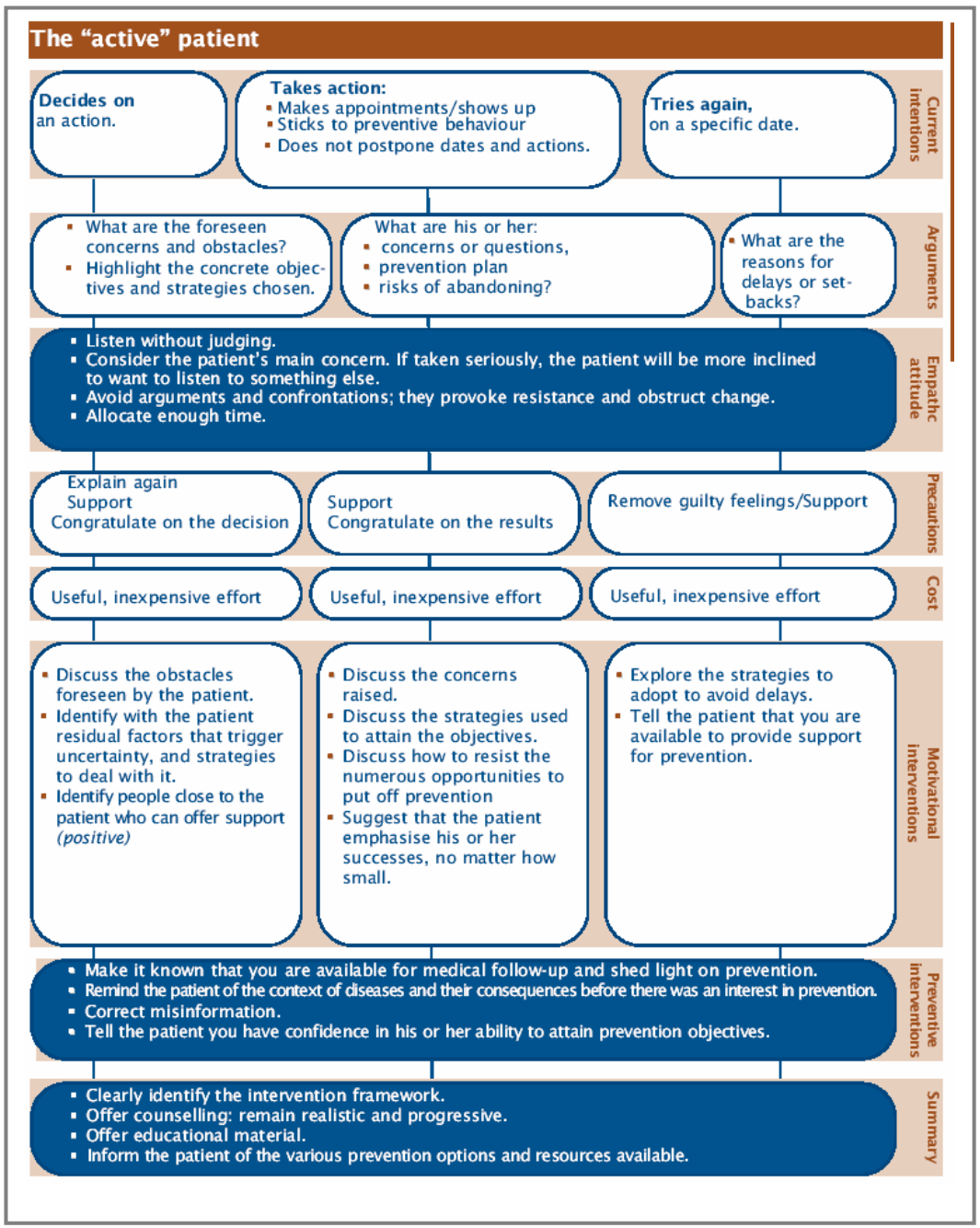
Figure 2. Bayesian Plot of Nurses Judgement Re a Diagnosis of Major Depression in Cancer



Caption: Bayesian graph plots the pre-test post-test gain for each possible prevalence value assuming sensitivity and specificity hold true.

Figure 3. Helping motivate the Undecided and the Active Patient (from Periodic health examination of Adults Preventative Clinical Guidelines)





10. References

[1] Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Pub Health Rep 1947; 62: 1432–49.

2 Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technology Assessment 2004; Vol 8: number 25

[3] Sackett DL, RB Haynes. The architecture of diagnostic research This is the second in a series of five articles BMJ 2002;324:539-541

[4] Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. Journal of Clinical Epidemiology 57 (2004) 925–932

[5] Youden WJ. Index for rating diagnostic tests. Cancer 1950; 3: 32–35.

[6] Bandolier : How good is the test. [http://www.jr2.ox.ac.uk/bandolier/band27/b27-2.html] 1996, 27:2.

[7] Bermejo I, Niebling W, Mathias B, Harter M. Patients' and physicians' evaluation of the PHQ-D for depression screening. Primary Care & Community Psychiatry 10 (4): 125-131 2005.

[8] Mitchell AJ, Coyne J. Screening for postnatal depression: barriers to success. British Journal of Obstetrics and Gynaecology 2008 online first.