How To:

Implement a Screening Programme for Distress in Cancer Settings

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 Screening Terminology

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 identify those with distress (or depression or anxiety disorder) as well as identify those without distress. It is a common mistake to focus on those with distress, looking at detection sensitivity but overlook detection specificity. These two attributes should be considered independent of each other. To illustrate this imagine I decide to diagnose everyone in the clinic as distressed. My detection sentivity would be 100% (I would by accident correctly identify all cases) but my detection specificity would be 0%.

A better way of thinking about reporting diagnostic accuracy is in terms of rule-in and rule-out performance. Sensitivity and specificity are somewhat abstract for clinicians who attempt to spot cases and reassure non-cases. The clinicians ability to spot true cases as a proportion of all their attempts is called the *positive predictive value* (PPV). In my view the PPV is one measure of case-finding ability. However some say that case finding is the overall accuracy when applied to those with a clear disorder that was hitherto unknown to the clinician. The clinicians ability to spot true non-cases (usually healthy individuals) as a proportion of all their attempts is called the *negative predictive value* (NPV). In my view the NPV is a measure of screening. However some say that screening is simply the overall accuracy when applied to those with a possible disorder that was hitherto unknown to the clinician.

Screening is usually applied to those at low or modest risk, that is low prevalence settings. The assumption is that a small number of cases or a modest number of those at high risk of being cases can be identifiyed by excluding a larger number of non-cases. A first stage screen may not have perfect PPV but it should have high NPV. *Case-finding* is usually applied in moderate or 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 wihtout 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. Choosing the "Best" Tool

Given that there are a large number of tools how can clinicians select the best one? The best tool really means the most suitable tool for the job. This can be defined as the most informative tool that is accepted by clinicians and patients. Here there is usually a tension between brevity and accuracy. The most accurate method to detect depression or distress would be a fully structured or semi-structured interview applied by experienced clinicians or researchers. Ultra-short methods usually have low specificity and are better at identifying the healthy then

identifying the unhealthy. This might be helpful for screening, but only when there is a follow-up for all those screening positive.

Regarding a comparison of tools, this is an area of controversy because all possible comparisons cannot realistically be conducted head-to-head. Further some comparisons prefer an ICD10 concept of depression or adjustment disorder whilst others prefer DSMIV. In my opinion the best methods will perform well against both standards.

At the moment every test can be considered imperfect but most can be refined by adding or removing items or changing the weighting of scoring or possibly the diagnostic algorithm. There have been recent attempts to improve efficacy of screening instruments using modern psychometrics, most notably using Rasch models. The models are part of a family of measurement models developed for educational psychology, increasingly employed in test development and refinement in medicine. Frequently it is found that conventional instruments may be shortened in length without significantly decreasing screening efficacy. Occasionally that shortening is dramatic, reducing an instrument by a quarter but there is may be a limit to the reducibility. Further, the ability of these adapted instruments to identify levels a key outcome variable such as "distress warranting intervention" still remains less than perfect. Combining items drawn from a number of emotional distress instruments into an item bank may improve screening efficacy, whilst at the same time minimising the number of questions patients are required to answer and consequently reducing patient burden. Item banks, such as these and computer-adaptive tests, which tailor the questions presented to patients' responses have already been successfully developed for assessing emotional distress in a psychiatric population (Fliege et al., 2005; Walter et al., 2007).[1] [2]

4. Analysing Accuracy from a Simple Screening Study

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).[3] 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.[4] 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

Box 1. 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

positives and hence any positive screen will suggest a true case. [5] 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 1).[5]

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),[6] but increasingly clinicians prefer summary statistics which generate clinically meaningful results. These are discussed below.

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**	

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].[7] 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. [8] 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.

5. Analysing Applicability from a Simple Screening Study

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.

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 1). 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.[9] 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.

Box 2: 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

6. Converting Screening Tests into Screening Programmes

Screening tests are usually examined in indivdual 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 3).

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.

Box 3: 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

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.

7. 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.

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.

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 4: Outcomes that can inform Screening programme Implementation

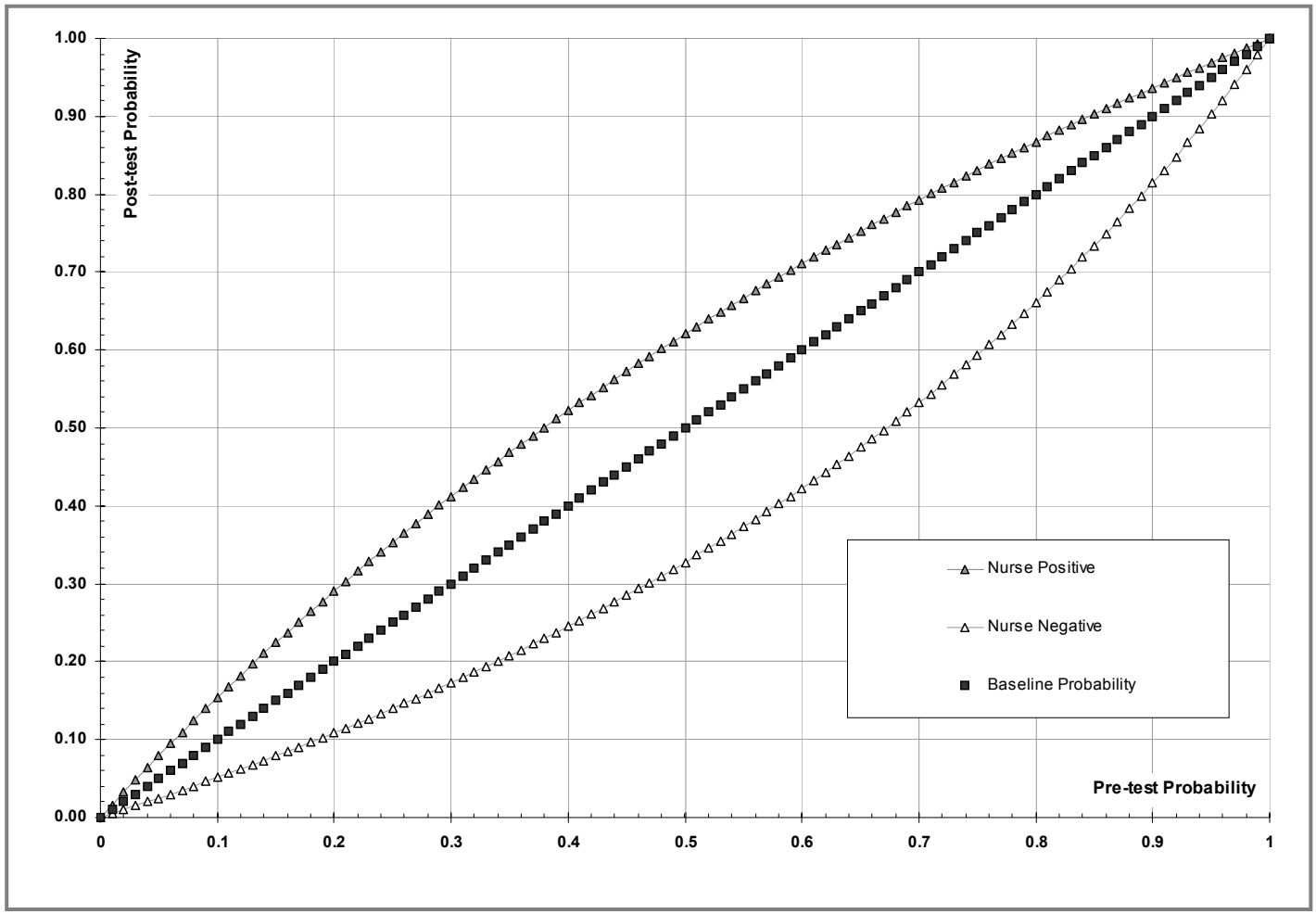
Screening uptake
Diagnostic sensitivity of staff
Diagnostic specificity of staff
Staff satisfaction
Staff burden
% of Staff offering treatment
% of patients offered treatment
Patient satisfaction
Patient burden
Patient wellbeing (HRQoL)
Patient distress / depression
Cost and cost-benefit

8. Conclusions

There are many screening instruments and much evidence concerning diagnostic validation but little research examining whether these method improve quality of care and almost no direct comparison with clinicians unassisted (routine) diagnoses. 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).[10] 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.

9. Illustrations.

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.

10. References

[1] Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res 2005; 14: 2277-2291.

[2] Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). Qual Life Res 2007; 16: S143-S155.

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

4 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

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

[6] 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

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

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

[9] 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.

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