# Estimating a test's accuracy using tailored meta-analysis – How setting-specific data may aid study selection

Willis, Brian; Hyde, CJ

*DOI:*
10.1016/j.jclinepi.2013.10.016



*Document Version*
Peer reviewed version

*Citation for published version (Harvard):*
Willis, BH & Hyde, CJ 2014, 'Estimating a test's accuracy using tailored meta-analysis – How setting-specific data may aid study selection', Journal of Clinical Epidemiology, vol. 67, no. 5, pp. 538–546. https://doi.org/10.1016/j.jclinepi.2013.10.016

# Estimating a test's accuracy using tailored meta-analysis - the potential of setting-specific data in aiding study selection

**Brian H Willis[1], Christopher J Hyde[2].**

1. School of Health and Population Sciences, University of Birmingham, UK

2. Public Health and Epidemiology, University of Exeter, UK

**Key words**

Meta-analysis

Diagnosis tests, Routine

Mass Screening

Models, Statistical

Data interpretation, Statistical

Decision making

## Abstract

### Objective

To determine a plausible estimate for a test's performance in a specific target setting using a new method for selecting diagnostic test accuracy studies which are applicable to the setting.

### Study design

It is shown how routine data collected on the test positive rate and the prevalence of disease for the setting of interest may be used to derive a region of performance for the test in receiver operating characteristic (ROC) space. After qualitative appraisal, studies are selected based on the probability that their study accuracy estimates arose from parameters lying in this 'applicable region'. Three methods for calculating these probabilities are developed and used to tailor the selection of studies for meta-analysis. The Pap test applied to the UK NHS cervical screening programme provides a case example.

### Results

The original meta-analysis for the Pap test included 68 studies but at most 17 studies were considered applicable to the NHS. For conventional meta-analysis the sensitivity and specificity (with 95% confidence intervals) were estimated to be 72.8% (65.8-78.8) and 75.4% (68.1-81.5) compared with 50.9% (35.8-66.0) and 98.0% (95.4-99.1) from tailored meta-analysis using a binomial method for selection. The effect of this is that for a prevalence for Cervical Intraepithelial Neoplasia (CIN) 1 of 2.2%, the post-test probability for CIN 1 increases from 6.2% to 36.6% between the two methods of meta-analysis.

### Conclusion

Tailored meta-analysis provides a method for augmenting study selection based on the study's applicability to a setting. As such the summary estimate is more likely to be plausible for a setting and could improve diagnostic prediction in practice.

## 1. Introduction

Ensuring that study results are applicable in practice or in a particular clinical setting poses a challenge to clinicians and policy makers. Yet this may not always be readily appreciated particularly when applying the results of diagnostic or screening test studies in practice; we shall see in the methods developed here, it is possible for studies to provide estimates of a test's performance which are highly improbable in some settings.

To facilitate diagnosis, the performance of a test is most usefully reported in terms of either its sensitivity and specificity or its positive and negative likelihood ratio - the latter two being directly derived from the former two [1,2]. The difficulties arise from these being dependent on external factors that may change individually or multiply between different settings. Variations in the disease prevalence [3,4], the work up of patients with other tests [3], the patient spectrum [5,6], and the execution of the test [7,8] may all contribute to differences in a test's performance between different surroundings [9]. Hence, a study of a diagnostic test represents a snap shot of its performance within a particular setting, with its own set of external factors.

With the potential for these external factors to vary between different settings, it is not surprising that heterogeneity is so widespread in diagnostic test accuracy (DTA) reviews and meta-analyses [10]. Although meta-analysis has allowed the study of test characteristics across a wide range of environments, the summary estimates they provide represent the average performance across all studies [11-13] and may not be representative of an individual clinical setting. This may be the case even when study selection is confined to a particular setting, such as primary care; the resulting summary (average) estimate may still misrepresent some individual primary care settings due to heterogeneity in the true test performance across settings.

This creates a problem for the current paradigm of evidence-based medicine, which is broadly based on ensuring that the best evidence is applied to clinical practice [14]. This usually relies upon the existing methods of critical appraisal being effective at identifying those studies which are representative of the setting in question. But without details of the external factors from both the study and the end user's population, how can we be sure that this is the case?

Rather than wholly relying on the effectiveness of critical appraisal to assess representativeness of studies, here an alternative method is proposed. For the purpose of meta-analysis, it will be demonstrated how basic data collection from the setting where the test is being applied may be used to tailor the selection of studies.

## 2. Method

The objective is to determine a plausible estimate for a test's performance in a specific setting given the combination of routine data from the setting and evidence from the literature. Broadly, the method relies upon first collecting actual data on the test positive rate and prevalence from the setting in question and using this to deduce a region of plausible values for the sensitivity and false positive rate for the test in the setting. This is then used to aid study selection for meta-analysis by comparing the sensitivities and false positive rates of studies with this region. This is now discussed more fully below.

### *2.1Defining a plausible region of performance for the setting*

Although the true performance of a test within a setting may not be known there are quantities which may be measured in the practice setting which allow us to define a plausible region of performance – called the *applicable region* for the test from hereon.

#### *(i) Using the test positive rate as a constraint*

Normally, within practice the test results are unverified since the true diagnosis, established by applying a reference standard, is not known at the time of testing [15]. Consequently, the numbers of true positives, false positives, false negatives and true negatives are unknown. However, these quantities must be non-negative.

In some settings the unverified results of all patients tested in a clinical population are known; for instance, the total number of positive test results and the negative test results for a particular test threshold may be counted. Importantly these may be used to constrain the possible values that the test's sensitivity, $s$ and false positive rate, $f$, may take within the practice setting.

If we define the test positive rate, $r$, as the proportion of all those tested who test positive for a particular threshold, then from these observations, the following condition may be deduced

if $f \leq s$ this implies $f \leq r$ and $s \geq r$ (1)

Thus, providing the test is able to classify subjects better than a random process, we can define the range of possible values that the false positive rate (FPR) and sensitivity may take in the (*f*, *s*) plane or ROC space (after Receiver Operating Characteristic curve) [16].

If our estimate of *r* is unbiased and precise (zero standard error), since these are mathematical constraints, the effect of imposing them is to make large areas of ROC space mathematically impossible. Although such accuracy and precision is unlikely in practice, these inequalities still serve to define the applicable region in ROC space representing the feasible values for the test's performance within the setting in question.

The size of the applicable region is determined in part by the precision of our estimate for *r*, since it affects the width of the confidence interval. Figure 1 demonstrates the effect an interval estimate for $0.15 \leq r \leq 0.3$ (bold boundary) has on the shape of the applicable region for a point estimate $r = 0.225$. The locations of the boundaries are defined by the upper and lower limits of the interval.

To summarise, the test positive rate may be derived by counting the number of patients testing positive as a proportion of the total number of patients tested within the target setting and a confidence interval estimate may be obtained readily from this. As a result, clinicians who frequently apply a particular test in practice could feasibly collect their own data to help define their own applicable region for the test.

The applicable region for the test may be refined further by incorporating knowledge of the prevalence of the target disorder in clinical practice.

*(ii) Incorporating knowledge of the local prevalence*

Suppose the prevalence of the target disorder for the setting of interest, *p* may be estimated, then, it follows from the definitions of the prevalence and sensitivity, *s* (see appendix 1) that

$$s = \frac{r}{p} - \frac{(1-p)f}{p} \qquad (2)$$

If it was possible to have perfect knowledge for $r$ and $p$ this would constrain the sensitivity and false positive rate for the test in clinical practice to a straight line. Although including uncertainty in the estimates for $r$ and $p$ widens the applicable region from a straight line to a polygonal area, the area is still narrower than when information on the prevalence is not included.

Unlike $r$, it is more difficult to estimate $p$ with both accuracy and precision. The best estimate is obtained from verifying (applying the reference standard to) all the patients in the setting, and is therefore unavailable. Accurate estimates of $p$ may be obtained from verifying a small sub-sample of patients in the setting as a form of calibration, but with an obvious loss in precision; such as, taking a swab (reference standard) from a sub-sample of patients when the test is a prediction rule for infection.

More usually, estimates of the prevalence are derived from sources outside of the setting and so risk being inaccurate for the setting in question. For example, local laboratories may have estimates of disease prevalence for the locality. A Bayesian argument could be used where the interval estimate is based on belief either from previous test results or where the empirical estimate for the prevalence is believed to be potentially biased. Thus, in a Bayesian framework, any summary estimate for the test (see 2.3) is conditioned on this interval estimate for the prevalence.

In figure 2, the applicable region shown in figure 1 is refined using an interval estimate for the prevalence, for $0.05 \leq p \leq 0.12$. For each ($r$, $p$) pair a straight line constraint (labelled 1 to 4) is generated and these are illustrated for ($r$=0.15, $p$=0.12), ($r$=0.15, $p$=0.05), ($r$=0.30, $p$=0.12), ($r$=0.30, $p$=0.05), respectively.

The dotted lines for constraints 2 and 3 are used to demonstrate the fact that these constraints are surplus or ‘dominated’ by the other two constraints (1, 4). This follows since if the ‘true’ sensitivity and FPR for the test is located to the right of 2 it will always lie to the right of 1, but not vice versa. Similarly, 3 is surplus when compared with 4.

*2.2 Study selection based on the applicable region*

In contrast to conventional meta-analysis, where studies are selected on the basis of qualitative criteria alone, here we propose to augment this process by selecting only those studies which report performances which are compatible with the setting in question. To do this requires an appropriate test procedure which incorporates the constraints defining the applicable region.

We are interested in only those studies whose 'true' sensitivity and FPR are likely to lie in the applicable region, so we may 'tailor' the selection of studies for meta-analysis to the setting in question. The test procedure used for selecting studies is dependent on whether we have knowledge of the test positive rate only or knowledge of both the test positive rate and prevalence where deriving the applicable region. The two cases are discussed below.

*(i) Knowledge of the test positive rate only*

Suppose $r$ is estimated by some confidence interval with a level of significance, $\alpha$, then we know within a classical framework that the parameter ('true' value) for the test positive rate, $\mu_r$, will be present in the interval $100(1-\alpha)\%$ of the time in the long run [17]. We will assume $\mu_r$ is contained in the interval. This is reasonable if we choose $\alpha$ so there is a high coverage probability, and suggest $\alpha$ is no more than 0.01. This will lead to a loss of precision in the estimate for $r$ which may be mitigated by ensuring a large sample size.

Let $\mu_f$ , and $\mu_s$ be the parameters for the FPR and sensitivity, respectively for the test in the setting of interest. From (2) we have $\mu_f \leq \mu_r$ and $\mu_s \geq \mu_r$ whenever $\mu_f \leq \mu_s$. Since $r_{lcl} \leq \mu_r \leq r_{ucl}$, where $r_{lcl}$ and $r_{ucl}$ are the lower and upper confidence limits for $r$, the most extreme value for $\mu_f$ lies on the boundary when $\mu_f = r_{ucl}$.

Similarly, for a study $i$ to be compatible with the setting of interest its parameter $\mu_{f,i}$ must lie in the applicable region. Its maximum permitted value is when it lies on the boundary when $\mu_{f,i} = r_{ucl}$. Thus, for studies lying to the right of this boundary we test whether the study estimate for the FPR, $\hat{f}_i$ , arises from a distribution whose parameter $\mu_{f,i}$ lies on the boundary of the applicable region. Since the FPR, $f_i$ has a binomial distribution we may calculate the probability of $f_i$ being equal to or more extreme than the study estimate $\hat{f}_i$ given that $\mu_{f,i} = r_{ucl}$.

The FPR is constant along the vertical boundary (see figure 1) so we do not need to consider the sensitivity of the study when testing hypotheses.

Hence, we suggest the following test procedure (referred to here as the '*B1 method')* for selecting studies: for study *i*, the null hypothesis $H_0$: $\mu_{f,i} = r_{ucl}$ is rejected in favour of the alternative hypothesis, $H_1$: $\mu_{f,i} > r_{ucl}$ if the tail area binomial probability $P(f_i \geq \hat{f}_i) < \beta$ for level of significance $\beta$. Studies are excluded if $H_0$ is rejected. This provides an exact approach given the assumption that $\mu_r$ lies in the confidence interval for *r*.

Note it is possible for studies lying outside of the applicable region to lie closer to the horizontal boundary, $s = r_{lcl}$ than the vertical boundary $f = r_{ucl}$. On such occasions we proceed as above but test the null hypothesis of $H_0$: $\mu_{s,i} = r_{lcl}$ versus the alternative $H_1$: $\mu_{s,i} < r_{lcl}$.

(ii) *Knowledge of test positive rate and prevalence*

Although knowledge of the prevalence within a setting does not allow us to make any assertions on the sensitivity and FPR for the setting (see appendix 1), in combination with the test positive rate we may refine the size of the applicable region for the setting.

Again we know that the parameters $\mu_p$ and $\mu_r$ for the prevalence and test positive rate respectively will lie in their 100 (1- $\alpha$)% confidence intervals, with a coverage probability of 1- $\alpha$. If we assume *p* and *r* to be independent then the coverage probability for both $\mu_p$ and $\mu_r$ being in their respective 95% confidence intervals at the same time is approximately 0.9. However, this probability is likely to be higher, since (1) may be derived from assuming *p* and *r* to be positively correlated [18]. Clearly, this coverage probability is dependent on the value of $\alpha$ chosen, but we will assume that $\mu_r$ and $\mu_p$ for the setting actually lie in their respective intervals by always choosing $\alpha \leq 0.01$.

Our aim is to select only those studies which have parameters for the sensitivity and FPR, ($\mu_{f,i}$,$\mu_{s,i}$) that lie in the applicable region. Other studies, which have ($\mu_{f,i}$, $\mu_{s,i}$) outside the applicable region, are considered implausible for the setting. Consistent with other models we will assume that for a single observation from each study *i*, the sensitivity, $s_i$ and FPR, $f_i$ are independent [11,12] and that both have binomial distributions [12].

If a study reports an estimate $(\hat{f}_i, \hat{s}_i)$ which lies in the applicable region we have no grounds for excluding it, since based on a single observation the maximum likelihood estimate (MLE) [19] for $(\mu_{f,i}, \mu_{s,i})$ must equal $(\hat{f}_i, \hat{s}_i)$. For studies which report estimates which lie outside of the applicable region we derive the MLE $(\mu_{max,f,i}, \mu_{max,s,i})$ for potential $(\mu_{f,i}, \mu_{s,i})$ pairs which are constrained to lie in the applicable region, (see Appendix 2).

Once $(\mu_{max,f,i}, \mu_{max,s,i})$ has been found, we may calculate the tail area probability, $P_i$ of the study $i$ producing an estimate $(f_i \geq \hat{f}_i, s_i \geq \hat{s}_i)$ given a parameter pair $(\mu_{max,f,i}, \mu_{max,s,i})$. Hence we have the following test procedure, ('*the B2 method*'): if we denote the applicable region by $A$, we reject the null hypothesis $H_0$: $(\mu_{f,i}, \mu_{s,i}) \in A$ in favour of $H_1$: $(\mu_{f,i}, \mu_{s,i}) \in A^c$ if $P_i < \beta$, where $A^c$ is the complement of $A$ (i.e. all points outside of the applicable region), and $\beta$ is the level of significance. The study is excluded if $H_0$ is rejected and this method provides an exact approach given the assumptions on $\mu_r$ and $\mu_p$ lying in the respective intervals for $r$ and $p$.

There are other approaches for estimating $\mu_{max,f,i}$ and $\mu_{max,s,i}$ based on assuming that the FPR and sensitivity for each study have independent but normal distributions. The estimate for $\mu_{max,f,i}$ and $\mu_{max,s,i}$ corresponds to the point on the boundary which minimises the 'statistical distance' called the Mahalanobis distance, $D_i^2$ between the boundary and the study estimate $(\hat{f}_i, \hat{s}_i)$ [20]. Here $D_i^2$ has a $\chi^2_{(2)}$ distribution [21] and so the critical value for $\beta$ =0.025 is 7.38. Thus, the '$D^2$ *method*' consists of rejecting $H_0$ if $D_i^2$ >7.38 for a particular study. This approach has the intuitive appeal of studies being selected on the basis of how 'close' they are to the applicable region. It also has the advantage that a good approximate estimate for $D_i^2$ may be calculated directly without recourse to sampling the applicable region (and therefore needing to write de novo software) as in the case of the *B2* approach. However, more refined estimates require an iterative approach which is best achieved computationally (see appendix 2).

Thus following qualitative appraisal, the process of study selection has two parts, this is summarised by the following algorithm (for the B2 method):

1. Define region in ROC space for the test in the setting by collecting data from the setting.

(i). Collect data on test positives and prevalence (if available) to calculate interval estimates for *r* and *p*.

(ii). Use upper and lower confidence limits for *r* and *p* to define region in ROC space according to relations (1) and (2) (see figure 2).

(iii). Assume the parameters $\mu_r$ and $\mu_p$ are contained in their interval estimates, then the parameters for the sensitivity, $\mu_s$ and FPR, $\mu_f$ for the test in the setting are also contained in the region – the 'applicable region'.

2. Exclude those studies which report a sensitivity and FPR that is improbable for the setting.

(i) For each primary study derive the maximum likelihood estimate (MLE) from all potential parameter pairs for the FPR and sensitivity that are constrained to lie in the applicable region – in practice the MLE will be on one of the boundaries.

(ii). Calculate the tail probability of the study estimate for the FPR and sensitivity given the MLE of the parameter pair lying on the boundary.

(iii). Exclude study if tail probability less than level of significance.

Similarly, study selection based on the 'minimum distance' ($D^2$) method follows readily with few changes to the second part of the algorithm.

*2.3 Statistical methods for aggregating the relevant studies*

Once the applicable studies have been identified, they may be aggregated using a bivariate random effects model (BRM) as previously described [12]. In effect a 'tailored' meta-analysis is conducted in which the summary estimate of the sensitivity and specificity is derived only from studies compatible with the applicable region.

The method is now illustrated by applying it to the Pap test to estimate its sensitivity and specificity when used in the NHS. All analyses were made using the statistical software R and STATA.

## 3. Results

*3.1 Case example - Screening for Cervical cancer*

Until 2003 the NHS Cervical Screening programme used the Pap test exclusively for screening women for cervical pre-cancerous changes [22]. The target group is women between the ages of 25 to 65 years, who between 25 and 49 are invited on a three yearly basis for a smear, and five yearly thereafter [23]. Negative smears result in routine recall, whereas abnormal smears lead to further investigation [24].

In 1999-2000, the NHS Cervical Screening programme reported that 169,946 smears were classed as mild dyskaryosis or worse [25]. Excluding inadequate smears, 3,675,297 smears were classed as negative or borderline [25]. To satisfy (2), the sensitivity of the Pap test in the NHS programme must be greater than $r$ =169,946/ 3,845,244 = 4.4%. Table 1 shows that within the NHS, $r$ has remained relatively constant across all thresholds for a number of years [25], suggesting that the Pap test had a consistent level of performance within the programme during this period.

We would like to maximise the coverage probability for the parameter, $\mu_r$ without a significant loss of precision. In this instance we may vastly exceed the 2.57 standard errors for a 99% confidence interval and provide an interval estimate of (4.437-4.447%) to 5 standard errors with minimal loss of precision. Essentially in the long run with repeated sampling, $\mu_r$ would lie outside such a confidence interval only once in 3.5 million [26]. We set $4.437\% \leq r \leq 4.447\%$.

The applicable region is refined by estimating the disease prevalence at a threshold of CIN 1. The prevalence of CIN 1 or greater in the UK may be estimated to be 2.2% (99% CI 2.19-2.22%) [27]. This is likely to be an underestimate as it based on those who are referred to colposcopy clinic and does not include those with negative smears. We will set a conservative range for $p$ to be $0.5\% \leq p \leq 10\%$.

One meta-analysis [28] which evaluated the accuracy of the Pap test included 68 studies at a test threshold of Low grade Squamous Intraepithelial Lesion (mild dyskaryosis in the UK system) and a reference standard threshold of CIN 1 [28]. The qualitative criteria used for selecting studies were: the study evaluated cervical cytology as a screening test; used histology or colposcopy as a reference standard; there was less than 3 months between the

cytology and the reference standard being applied; and the 2 x 2 table could be completed [28].

Overall in the review, verification bias was a potential issue with 69% of primary studies reporting verification of only a selected (non-random) sample of test negatives. Further only a quarter of studies reported blinding of the reference standard. Other factors such as patient spectrum were also poorly reported [28]. However, it is unclear what the combined effect of these was on individual study estimates.

The results of applying the different methods B1, B2 and $D^2$ are summarised in table 2. The level of significance, $\beta$ =0.025. As expected when the applicable region is defined by *r* only (B1 method) more studies (17) are included for tailored meta-analysis than when both *r* and *p* define the region. The $D^2$ method excludes one of these studies (16) and the B2 method excludes a further four. The difference between $D^2$ and B2 is that the former relies upon a normal approximation which may not be valid for some studies of small sample size.

Figure 3 illustrate the effects of applying the B2 method to meta-analysis. The dotted lines define the applicable region. The bold squares represent 12 studies reporting FPRs which are most likely to be compatible with the NHS programme; there was agreement across the three methods on these studies. The transparent triangles lying outside of the applicable region are those studies which are incompatible with the NHS. Out of 17 UK-based studies only 3 were in the applicable region. It is clear from figure 3 that conventional meta-analysis provides an estimate which would be highly improbable in the NHS given the location of the applicable region.

The median sample size for the included set of studies was 105, that is, 50% (6/12) of studies had a sample size, $n \leq 105$. To put this in context, the studies varied in sample size across ROC space: 44% (7/16) of studies had a sample size, $n \leq 105$ in the region FPR $\leq 0.1$, compared with 19% (10/52) of studies in the region FPR >0.1. Thus, the sample sizes of the included studies did not vary significantly (*p=0.39*, Fisher's exact test) from the FPR $\leq 0.1$ region which contains the applicable region.

Confining the meta-analysis to those studies which are in the applicable region has a profound effect on the summary estimates of performance for the Pap test. From conventional

meta-analysis the sensitivity and specificity are estimated to be 72.8% (95% CI: 65.8-78.8) and 75.4% (95% CI: 68.1-81.5) compared with 50.9% (35.8-66.0) and 98.0% (95% CI: 95.4-99.1) from tailored meta-analysis using the B2 method for selection. The increased specificity also accounts for the eight-fold increase in the positive likelihood ratio from 3.0 (95% CI: 2.4-3.7) to 25.6 (95% 10.1- 65.0) between the conventional and tailored meta-analysis. The effect of this is that, for a background prevalence of CIN 1 of 2.2%, the post-test probability for CIN 1 increases from 6.2% to 36.6%.

This example provides empirical evidence that basic information on the target setting may be required before applying the results of diagnostic test studies to clinical practice.

## 4. Discussion

It is the observation that a test's performance may vary between settings that makes determining when performance statistics are actually applicable to a particular setting, important. Within the traditional paradigm, if the diagnostic process is to truly become evidence-based, then this is probably one of the most important questions to answer; especially, as it affects both primary and secondary research.

When selecting primary studies for a meta-analysis the usual approach is to use qualitative criteria to decide on which studies to include. However, if meta-analyses are to produce results that are relevant to clinical practice, choosing studies based on qualitative criteria alone may be insufficient. In particular, if there is a wide variation in performances across different clinical settings, being clear on the target setting seems equally important. While this is usually made clear in the review question, the problem in defining the target setting qualitatively is that it may not accurately represent which studies are likely to be applicable.

Here, the possibility of using data collected from the target setting has been highlighted and how they may be used to introduce greater objectivity to ascertaining a test's performance in a particular setting. As a result it represents a departure from the usual methodology in meta-analysis by selecting studies on the basis of their results and not on qualitative criteria alone.

However, this should be considered in context. In general, when conducting a meta-analysis no assumption is made on where the true performance lies and therefore, there is no justification in selecting studies based on their individual results. Nonetheless the studies are still constrained by mathematical plausibility.

Similar to conventional meta-analysis potentially some of the included studies may report estimates which are either biased due to flaws in the primary study design or subject to sampling error. Whereas sampling error is largely accommodated by the models, a number of study design biases such as verification bias, participant selection bias, incorporation bias and reference standard bias are recognised and may lead to both the overstating and understating of test accuracy [29, 30, 31]. In principle part of the critical appraisal process should be aimed at mitigating these but it is not always possible.

Other setting-specific factors such as differences in patient spectrum, the positioning of the test threshold and the way the test is executed by the clinician may also explain some of the variation in the reported test accuracies. Changes in patient spectrum may manifest as a change in prevalence [5,6], whilst the latter two may be precipitated by changes in prevalence between settings [4, 32].

The difficultly is in being able to distinguish unbiased estimates from biased estimates and there seems to be no straight forward way of achieving this. Usually any differences between studies are investigated as part of a general exploration of heterogeneity and since this gives some indication of the sources of between-study variation, it should always be carried out.

Ideally, if all the relevant sources of variation were known and their individual and combined effects quantifiable, the estimate provided by each study could be modified accordingly to provide us with an unbiased estimate of the study parameter. Unfortunately, there are often multiple drivers to the accuracy of a test reported in a study and due to a lack of study-level data and the potential challenges of modelling multiple covariates, their overall combined effect on the reported accuracy is, in general, unknown.

Thus, the method developed here represents a pragmatic approach which considers the reported estimate at face-value in terms of its compatibility with the intended target setting. Providing a meta-analysis for the test in question exists and test data may be collected from

the practice setting, decisions on the applicability of the primary studies to the setting may be made. This allows us to produce a summary estimate which is at least plausible for the setting, something which conventional meta-analysis is not always able to.

There are two obvious roles for tailored meta-analysis in practice. For those diagnostic tests that are being applied in routine practice and where meta-analyses are available, there is the potential to use tailored meta-analysis to provide plausible estimates for the performance within the setting. This could improve diagnostic prediction and management decision-making.

Another use may be to aid the decision process of whether a particular test should be used at all in practice. Before a new diagnostic technology is introduced into practice, there is often substantial uncertainty on how it will actually perform in the intended setting and such decisions to implement the technology are most often based on appraisals of the literature and economic considerations. However, post-implementation there is the opportunity to re-appraise the decision to introduce the technology. By collecting setting-specific data and defining an applicable region, tailored-meta analysis could be used to re-assess the performance of the test and see if it is compatible with those estimates made prior to implementation. Ultimately this may lead to a decision on disinvestment.

The apparent theoretical advantage of constraining studies to an applicable region, to calculate summary estimates, needs to be validated empirically. This could take the form of comparing the results of a primary study in the target setting with the results of a tailored meta-analysis based on the available studies or alternatively could be done by way of a simulation study. The latter has the advantage of allowing other statistical properties of the method to be explored.

*Limitations of the analysis and potential areas of research*

The foregoing analysis has demonstrated that the summary estimate derived using only those studies which lie in the applicable region is at least feasible for the target setting which is not always the case for those estimates derived from conventional meta-analysis. Although, this follows from the mathematical constraints, uncertainty in the estimates of the test positive rate and the prevalence may result in inaccurate boundaries being drawn for the applicable region. However, this is minimised by using interval estimates with a high coverage

probability. In the case example, our estimate of the test positive rate in the NHS was to five standard errors - the sort of significance demanded by particle physicists when confirming the existence of the Higgs Boson [26]!

In tailored-meta-analysis studies are included if they are either inside or 'close enough' to the applicable region that they cannot be confidently excluded. Although it was not the case with the Pap test example used earlier, potentially, some studies of small sample sizes which are close to the applicable region could be included because uncertainty in the estimate does not allow them to be confidently excluded. However, such studies will have less influence on the summary estimate which is weighted towards studies with larger sample sizes.

A potential consequence of selecting only those studies lying in the applicable region for meta-analysis is that it is feasible there will be none to aggregate. In such instances, it is worth considering whether the data used to define *r* and *p* for the setting are reliable. If they are then it must be concluded that none of the studies accurately represents how the test performs within the target setting.

In summary, the foregoing analysis demonstrates the possibility of using quantitative data collected from clinical practice to augment the process of study selection by determining quantitatively which studies may be applicable to a setting. Such an approach suggests tailoring the meta-analysis to include only the applicable studies before deriving summary estimates.

**Acknowledgements**

We would like to thank all those who have read this or earlier drafts of this manuscript and provided comments that have helped develop and add clarity to the work. In particular, we are grateful to Graham Dunn, Aneez Esmail, Andrew Pickles, Hans Reitsma, Andy Vale, Jon Deeks, Richard Riley and Yemesi Takwoingi.

**Conflicts of interest**

None.

**Funding**

During the development of this work BHW was in receipt of a Medical Research Council fellowship (grant reference G0701649) to conduct research into diagnostic tests.

**References**

1. Pepe, MS. The statistical evaluation of medical tests for the classification and prediction. Oxford: OUP; 2003:4-34.

2. Zhou X, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: John Wiley and Sons; 2002:15-55

3. Leeflang M, Bossuyt P, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12.

4. Willis BH. Evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross sectional study. BMJ Open 2012; 2:e000746.

5. Willis BH. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. Fam Pract. 2008;25:390-396.

6. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30.

7. Koran LM. The reliability of clinical methods, data and judgement (part 1). N Engl J Med. 1975; 293:642-646.

8. Koran LM. The reliability of clinical methods, data and judgement (part 2). N Engl J Med. 1975; 293:695-701.

9. Irwig LM, Bossuyt PM, Glasziou P, Gatsonis C, Lijmer JG. The evidence base for clinical diagnosis: Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669–71.

10. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol 2011; 11:27

11. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982-90.

12. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006; 59:1331-1332.

13. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statist Med 2001; 20:2865-2884.

14. Strauss SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine: how to practice and teach it. Fourth edition. London: Churchill Livingstone; 2010:1

15. Jaeschke R, Guyatt GH, 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? JAMA. 1994; 271:389-391.

16. Tanner Jr WP, Swets JA. A decision making theory of visual detection, Psychol. Rev. 1954; 61: 401-409

17. Cox DR. Principles of statistical inference. Cambridge: CUP; 2006:8.

18. Kraemer HC. Evaluating medical tests: Objective and quantitative guidelines, London: Sage; 1992: 29-30.

19. Garthwaite PH, Jolliffe IT, Jones B. Statistical inference. London: Prentice Hill; 1995:41-70

20. Krzanowski WJ. Principles of multivariate analysis. Oxford: OUP 1988: 234.

21. Penny KI. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. Appli Statist. 1996:45(1):73-81

22. Understanding Cervical Screening, Cancerbackup, Printed by Sephen Austin England 2005, http://www.cancerscreening.nhs.uk/cervical/publications/understanding-cervical-screening.pdf

23. http://www.cancerscreening.nhs.uk/cervical/index.html

24. Willis BH, Barton P, Pearmain P, Bryan S, Hyde CJ. Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK. Health Technol Assess 2005; 9(13).

25. http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/StatisticalWorkAreas/Statisticalhealthcare/DH_4086491

26. http://blogs.scientificamerican.com/observations/2012/07/17/five-sigmawhats-that/

27. http://www.cancerscreening.nhs.uk/cervical/cervical-statistics-bulletin-2010-11.pdf

28. McCrory DC, Matchar DB, Bastian L, et al. Evaluation of Cervical Cytology. Evidence Report/Technology Assessment No. 5. (Prepared by Duke University under Contract No. 290-97-0014.) AHCPR Publication No. 99-E010. Rockville, MD: Agency for Health Care Policy and Research. February 1999.

29. Gray R, Begg CB, Greenes RA. Construction of Receiver Operating Characteristic curves when disease verification is subject to selection bias. Med Decis Making 1984; 4:151-164.

30.Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making 1987; 7:139-148.

31. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. A systematic review of sources of variation and bias in studies of diagnostic accuracy. Ann Intern Med 2004;140:189-202.

32. Egglin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996;276:1752-1755.

## Appendix 1. Derivation of the applicable region

*1.1 Using the test positive rate to define the test performances feasible in practice*

In the unverified population, all that may be known are the test results without the corresponding reference standard results. This is represented by the 2 by 2 below, where $x$ (the total number tested positive) and $y$ (the total number tested negative) are known and $a$, $b$, $c$ and $d$ are unknown. We do know that $a,b,c,d \geq 0$, otherwise negative numbers of subjects are possible.

| | | **DISEASE** | | |
|---|---|---|---|---|
| | | **Positive** | **Negative** | **Totals** |
| **TEST** | **Positive** | $a$ | $b$ | $x$ |
| | **Negative** | $c$ | $d$ | $y$ |
| | **Totals** | $a + c$ | $b + d$ | $x + y$ |

From definitions of sensitivity, $s$ and FPR, $f$` $\quad c = \dfrac{a(1-s)}{s}$ and $d = \dfrac{b(1-f)}{f}$

Also $\quad c + d = y \quad$ and $\quad x - a = b$

Hence we have $\quad y = \dfrac{a(1-s)}{s} + \dfrac{(x-a)(1-f)}{f}$

From re-arranging we have for $f \neq s$ $\quad a = \dfrac{sfy - s(1-f)x}{(f-s)}$

Now for $a \geq 0$ consider the 2 possibilities $a = 0$ and $a > 0$. Trivially if $a = 0$ we have $s = 0$

For $a > 0$

either $f > s$ and $sfy - s(1-f)x > 0 \Rightarrow f > \dfrac{x}{(x+y)}$

or $f < s$ and $sfy - s(1-f)x < 0 \Rightarrow f < \dfrac{x}{(x+y)}$

for $f = s$ $\quad f = \frac{x}{(x+y)}$

Similarly, solving for *b, c* and *d* then setting them all >0 we get the following 2 inequalities.

For $f \geq s$ we have $s \leq \frac{x}{(x+y)}$ and for $f < s$ we have $s > \frac{x}{(x+y)}$

Thus we have the results

$f < r$ and $s > r$ whenever $f < s$ (1)

$f \geq r$ and $s \leq r$ whenever $f \geq s$ (2)

where $r = \frac{x}{(x+y)}$

An alternative derivation of (1) may be obtained from assuming that the probability of a test positive and the presence of disease are correlated. This approach was used by Kraemer [18] to derive (1). Kraemer's use of the inequality was confined to single primary studies and not used to select studies for meta-analysis.

*1.2 Incorporating the prevalence of the target disorder*

Suppose the prevalence of the target disorder, *p* may be estimated, then, from the definition of prevalence, *p*, sensitivity, *s* and false positive rate *f* we have

$$p(x+y) = \frac{a}{s} \quad \text{and} \quad (1-p)(x+y) = \frac{b}{f}$$

Since $a + b = x$ $\quad sp + f(1-p) = \frac{x}{(x+y)} = r$

Hence, we have $\quad s = \frac{r}{p} - \frac{(1-p)}{p} f$ (3)

Comparing (1) and (3) it can be seen that the size of the applicable region is most dependent on having precise knowledge of *r*. Thus, if *p* is known perfectly and *r* is not known, (ie $0 \leq r$

$\leq 1$) then all values for $f$ and $s$ in ROC space are still possible. It follows from (3) that for a fixed value of $p$, allowing $r$ to vary between 0 and 1 changes only the intercept. This produces a plane of parallel lines which covers all values of ROC space. To appreciate this, consider the extreme values of $f$. For $f = 0$, $s$ has a value $r/p$, and since $0 \leq r \leq 1$ this implies that $0 \leq s \leq (1/p)$. Hence, $s$ can take on all values between 0 and 1 since $(1/p) \geq 1$.

Similarly, for $f = 1$, $s$ has a value of $(r/p) - (1-p)/p$. For values of $0 \leq r \leq 1$, this implies $-(1-p)/p \leq s \leq 1$. Hence $s$ takes on values between zero and 1 since $-(1-p)/p \leq 0$. For $f=1$, $s$ again covers all values between 0 and 1 in ROC space. Thus, without knowledge of $r$, even perfect knowledge of $p$ is unhelpful.

**Appendix 2 - Selection of studies based on the applicable region**

*2.1 Binomial B2 method*

In all three methods we make the assumption that the parameters $\mu_r$ and $\mu_p$ for the test positive rate, *r* and prevalence, *p*, respectively, lie in their corresponding interval estimates. This is only reasonable if the level of significance is chosen so that there is a high coverage probability.

For a study *i*, ($f_i$, $s_i$) are the FPR and sensitivity, $(\hat{f}_i, \hat{s}_i)$ are the study estimates and ($\mu_{f,i}$, $\mu_{s,i}$) are the study parameters. Also $n_{f,i}$ and $n_{s,i}$ are the total number without and with the target disorder. We are interested only in studies which have ($\mu_{f,i}$, $\mu_{s,i}$) pairs in the applicable region. For a study estimate $(\hat{f}_i, \hat{s}_i)$ lying outside of the applicable region, the maximum likelihood estimate (MLE) for a ($\mu_{f,i}$, $\mu_{s,i}$) pair which actually lies in the applicable region must be located on the boundary closest to $(\hat{f}_i, \hat{s}_i)$. Since $s_i$ and $f_i$ are independent with $s_i \sim Bin(n_{s,i}, \mu_{s,i})$ and $f_i \sim Bin(n_{f,i}, \mu_{f,i})$, respectively, the likelihood function, $L(\mu_{f,i}, \mu_{s,i} \mid \hat{f}_i, \hat{s}_i)$ for a study *i* is given by

$$L(\mu_{f,i}, \mu_{s,i} \mid f_i = \hat{f}_i, s_i = \hat{s}_i) = Bin(b_i \mid n_{f,i}, \mu_{f,i}) Bin(a_i \mid n_{s,i}, \mu_{s,i})$$

where $b_i$ and $a_i$ are the number of false positives and true positives for study *i* respectively. This is maximised for ($\mu_{f,i}$, $\mu_{s,i}$) pairs lying on the closest boundary to $(\hat{f}_i, \hat{s}_i)$ and provides ($\mu_{max,f,i}$, $\mu_{max,s,i}$). If this is written in terms of Log ($L(\mu_{f,i}, \mu_{s,i} \mid \hat{f}_i, \hat{s}_i)$), the maximum likelihood estimate may be obtained by finding the real root to a cubic equation. Alternatively, it may be found by sampling points on the boundary closest to $(\hat{f}_i, \hat{s}_i)$ and finding the pair ($\mu_{f,i}$, $\mu_{s,i}$) which yields the maximum value for $L(\mu_{f,i}, \mu_{s,i} \mid \hat{f}_i, \hat{s}_i)$. The latter method was used here.

The tail area probability, $P(f_i \geq \hat{f}_i, s_i \geq \hat{s}_i \mid \mu_{f,i} = \mu_{max,f,i}, \mu_{s,i} = \mu_{max,s,i})$ is given by

$$P_i = \left( \sum_{b=b_i}^{n_{f,i}} Bin(b \mid n_{f,i}, \mu_{max,f,i}) \right) \left( \sum_{a=a_i}^{n_{s,i}} Bin(a \mid n_{s,i}, \mu_{max,s,i}) \right)$$

Studies are rejected if $P_i < \beta$ for level of significance $\beta$, where $\beta$ was set to 0.025.

*2.2 Mahalanobis distance, $D^2$ method*

If we assume $s_i$ and $f_i$ to have independent normal distributions such that, $s_i \sim N(\mu_{s,i}, \sigma^2_{s,i})$ and $f_i \sim N(\mu_{f,i}, \sigma^2_{f,i})$ then the *Mahalanobis distance*, $D_i^2$ for a study $i$, may be written as

$$D_i^2 = \frac{(\hat{f}_i - \mu_{f,i})^2}{\sigma^2_{f,i}} + \frac{(\hat{s}_i - \mu_{s,i})^2}{\sigma^2_{s,i}}$$

Unlike the Euclidean distance between points this allows us to weight the variates according to the inverse of their individual variances. As such it represents a 'statistical' distance between points. The optimum ($\mu_{f,i}$ , $\mu_{s,i}$) point on the boundary, that is, the point which maximises the probability of a study estimate for a given ($\mu_{f,i}$ , $\mu_{s,i}$), is that which minimises $D_i^2$. It is straight forward to show that this corresponds to the MLE for ($\mu_{f,i}$ , $\mu_{s,i}$) since $s_i$ and $f_i$ are assumed to have independent normal distributions.

It may be shown that $D_i^2$ is minimised for a point with $\mu_{f,i}$ on the boundary when it is given by

$$\mu_{f,i} = \frac{\sigma^2_{s,i} p^2 \hat{f}_i + r(1-p)\sigma^2_{f,i} - p(1-p)\sigma^2_{f,i}\hat{s}_i}{\sigma^2_{s,i} p^2 + (1-p)^2 \sigma^2_{f,i}}$$

Similarly, $\mu_{s,i}$ is given by substituting $\mu_{f,i}$ into

$$\mu_{s,i} = \frac{r}{p} - \frac{(1-p)\mu_{f,i}}{p}$$

Where $(r,p)$ correspond to either $(r_{ucl}, p_{lcl})$ or $(r_{lcl}, p_{ucl})$ depending on which boundary we are considering. The parameter $\mu_{f,i}$ is obtained iteratively by first making initial estimates for the variances $\sigma^2_{f,i} = \frac{\hat{f}_i(1-\hat{f}_i)}{n_{f,i}}$ and $\sigma^2_{s,i} = \frac{\hat{s}_i(1-\hat{s}_i)}{n_{s,i}}$.

Once we have our first estimates $\mu_{s,i}$ and $\mu_{f,i}$ these are then substituted into the variances instead of $\hat{f}_i$ and $\hat{s}_i$ and the process continued until we have convergence.

Since $D_i^2$ here is the sum of two squared standardised normal variables it has an approximate $\chi^2_{(2)}$ distribution. Thus, for level of significance, $\beta$ =0.025, the critical value is 7.38. Studies are excluded if $D_i^2$ >7.38.

## Appendix 3 Results of inclusion/ exclusion decisions using the applicable region.

| | | | | Binomial (B1) AR (r only) | | Binomial (B2) AR (r and p) | | Mahalanobis ($D^2$) AR (r and p) | |
|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Year** | **Sens** | **FPR** | **Probability** | **Decision** | **Probability** | **Decision** | **$D^2$** | **Decision** |
| Chomet | 1987 | 33.3 | 20.0 | 0.20439 | Include | 0.0059 | Exclude | 2.97 | Include |
| Cox | 1995 | 38.0 | 4.2 | 0.62166 | Include | 1.0 | Include | 0.00 | Include |
| Davis | 1981 | 69.5 | 0.0 | 1.0 | Include | 1.0 | Include | 0.00 | Include |
| Davison | 1994 | 53.3 | 0.0 | 1.0 | Include | 1.0 | Include | 0.00 | Include |
| DiBonito | 1993 | 61.0 | 4.3 | 0.59327 | Include | 0.1777 | Include | 0.04 | Include |
| Fung | 1997 | 86.6 | 9.5 | 0.03428 | Include | 0.0033 | Exclude | 6.45 | Include |
| Giles | 1989 | 81.3 | 0.0 | 1.0 | Include | 1.0 | Include | 0.00 | Include |
| Glenthoj | 1988 | 81.6 | 0.0 | 1.0 | Include | 1.0 | Include | 0.00 | Include |
| Gonzalez | 1996 | 38.1 | 12.2 | 0.03512 | Include | 0.0030 | Exclude | 6.20 | Include |
| Gundersen | 1988 | 15.4 | 10.0 | 0.14879 | Include | 0.0152 | Exclude | 2.22 | Include |
| Jones | 1987 | 17.2 | 2.2 | 0.95960 | Include | 0.2166 | Include | 0.26 | Include |
| Morrison | 1992 | 83.3 | 0.0 | 1.0 | Include | 1.0 | Include | 0.00 | Include |
| Spitzer | 1987 | 18.2 | 4.8 | 0.56556 | Include | 0.1161 | Include | 0.01 | Include |
| Stafl | 1981 | 43.8 | 10.0 | 0.36701 | Include | 0.0257 | Include | 0.80 | Include |
| Tay | 1987 | 33.3 | 0.0 | 1.0 | Include | 0.5026 | Include | 0.13 | Include |
| Wright | 1994b | 26.7 | 1.5 | 0.9993 | Include | 0.2987 | Include | 0.08 | Include |
| Skehan | 1990 | 82.4 | 82.8 | 0 | Exclude | 0 | Exclude | 0.00 | Exclude |
| Cox | 1992 | 43.8 | 7.8 | 0.0039 | Exclude | 0.0005 | Exclude | 10.64 | Exclude |
| Coibion | 1994 | 22.0 | 15.0 | 0.0083 | Exclude | 0.0007 | Exclude | 10.76 | Exclude |
| Naslund | 1986 | 100.0 | 23.1 | 0.0182 | Exclude | 0.0000 | Exclude | 12.38 | Exclude |
| Jones | 1992 | 36.6 | 13.1 | 0.0057 | Exclude | 0.0005 | Exclude | 11.47 | Exclude |
| Korn | 1994 | 62.9 | 15.8 | 0.0064 | Exclude | 0.0004 | Exclude | 12.80 | Exclude |
| Hellberg | 1987 | 88.9 | 33.3 | 0.0266 | Include | 0.0005 | Exclude | 13.26 | Exclude |
| Frisch | 1994 | 35.0 | 27.3 | 0.0112 | Exclude | 0.0004 | Exclude | 14.03 | Exclude |
| MacCormac | 1988 | 84.6 | 5.4 | 0.0024 | Exclude | $10^{-06}$ | Exclude | 19.60 | Exclude |
| Korn | 1994 | 63.6 | 26.3 | 0.0012 | Exclude | $10^{-05}$ | Exclude | 23.30 | Exclude |
| Garutti | 1994 | 41.7 | 14.1 | $10^{-05}$ | Exclude | $10^{-06}$ | Exclude | 29.84 | Exclude |
| Kealy | 1986 | 86.0 | 12.1 | $10^{-06}$ | Exclude | $10^{-07}$ | Exclude | 34.14 | Exclude |
| Del Priore | 1995 | 80.0 | 25.0 | $10^{-05}$ | Exclude | $10^{-06}$ | Exclude | 35.64 | Exclude |
| Andrews | 1989 | 26.3 | 12.3 | $10^{-06}$ | Exclude | $10^{-08}$ | Exclude | 35.86 | Exclude |
| Herrington | 1995 | 68.8 | 38.5 | 0.0002 | Exclude | $10^{-06}$ | Exclude | 38.51 | Exclude |
| Tait | 1988 | 74.5 | 18.4 | $10^{-06}$ | Exclude | $10^{-07}$ | Exclude | 39.36 | Exclude |
| Ferris | 1998 | 37.0 | 14.5 | $10^{-07}$ | Exclude | $10^{-09}$ | Exclude | 45.32 | Exclude |
| Giles | 1988 | 53.1 | 14.9 | $10^{-07}$ | Exclude | $10^{-09}$ | Exclude | 47.03 | Exclude |
| Mayeaux | 1995 | 43.8 | 19.2 | $10^{-08}$ | Exclude | $10^{-09}$ | Exclude | 53.87 | Exclude |
| Singh | 1985 | 75.2 | 66.7 | $10^{-05}$ | Exclude | $10^{-07}$ | Exclude | 59.51 | Exclude |
| Herrington | 1995 | 68.8 | 29.7 | $10^{-07}$ | Exclude | $10^{-09}$ | Exclude | 60.91 | Exclude |
| Wheelock | 1989 | 48.1 | 18.6 | $10^{-10}$ | Exclude | $10^{-11}$ | Exclude | 70.51 | Exclude |
| Parker | 1986 | 74.5 | 15.4 | $10^{-10}$ | Exclude | $10^{-12}$ | Exclude | 72.81 | Exclude |
| Oyer | 1986 | 75.1 | 21.0 | $10^{-09}$ | Exclude | $10^{-11}$ | Exclude | 75.43 | Exclude |
| Johansen | 1979 | 90.2 | 28.3 | $10^{-10}$ | Exclude | $10^{-12}$ | Exclude | 91.35 | Exclude |
| Syrjanen | 1987 | 72.8 | 17.9 | $10^{-14}$ | Exclude | $10^{-16}$ | Exclude | 107 | Exclude |
| Koonings | 1992 | 79.5 | 29.0 | $10^{-11}$ | Exclude | $10^{-13}$ | Exclude | 109 | Exclude |
| Chomet | 1987 | 67.0 | 38.9 | $10^{-10}$ | Exclude | $10^{-12}$ | Exclude | 109 | Exclude |
| Haddad | 1988 | 86.7 | 56.3 | $10^{-09}$ | Exclude | $10^{-11}$ | Exclude | 112 | Exclude |
| Smith | 1987 | 78.0 | 41.9 | $10^{-10}$ | Exclude | $10^{-12}$ | Exclude | 113 | Exclude |
| Young | 1993 | 72.0 | 27.4 | $10^{-12}$ | Exclude | $10^{-14}$ | Exclude | 115 | Exclude |
| Nyirjesy | 1972 | 60.0 | 50.0 | $10^{-11}$ | Exclude | $10^{-13}$ | Exclude | 136 | Exclude |
| Koonings | 1992 | 69.3 | 36.4 | $10^{-13}$ | Exclude | $10^{-16}$ | Exclude | 144 | Exclude |

| Name | Year | Sens | FPR | Binomial (B1) AR (r only) Probability | Decision | Binomial (B2) AR (r and p) Probability | Decision | Mahalanobis ($D^2$) AR (r and p) $D^2$ | Decision |
|---|---|---|---|---|---|---|---|---|---|
| Walker | 1986 | 76.9 | 46.9 | $10^{-12}$ | Exclude | $10^{-14}$ | Exclude | 149 | Exclude |
| Bolick | 1998 | 85.1 | 63.6 | $10^{-14}$ | Exclude | $10^{-16}$ | Exclude | 200 | Exclude |
| Lozowski | 1982 | 93.0 | 50.0 | 0 | Exclude | 0 | Exclude | 219 | Exclude |
| Germain | 1994 | 66.7 | 21.0 | 0 | Exclude | 0 | Exclude | 220 | Exclude |
| Beeby | 1993 | 65.7 | 21.9 | 0 | Exclude | 0 | Exclude | 288 | Exclude |
| Baldauf | 1995 | 92.9 | 68.8 | 0 | Exclude | 0 | Exclude | 347 | Exclude |
| Maggi | 1989 | 76.9 | 47.8 | 0 | Exclude | 0 | Exclude | 436 | Exclude |
| Soutter | 1986 | 83.2 | 56.9 | 0 | Exclude | 0 | Exclude | 464 | Exclude |
| Bigrigg | 1990 | 96.4 | 66.0 | 0 | Exclude | 0 | Exclude | 469 | Exclude |
| Kwikkel | 1986 | 98.0 | 91.9 | 0 | Exclude | 0 | Exclude | 744 | Exclude |
| Rasbridge | 1995 | 86.6 | 49.7 | 0 | Exclude | 0 | Exclude | 872 | Exclude |
| Wetrich | 1986 | 81.2 | 33.1 | 0 | Exclude | 0 | Exclude | 928 | Exclude |
| Upadhyay | 1984 | 99.6 | 77.9 | 0 | Exclude | 0 | Exclude | 970 | Exclude |
| Fahim | 1991 | 85.1 | 38.0 | 0 | Exclude | 0 | Exclude | 1283 | Exclude |
| Melnikow | 1997 | 88.6 | 88.5 | 0 | Exclude | 0 | Exclude | 2075 | Exclude |
| Melnikow | 1997 | 98.7 | 88.5 | 0 | Exclude | 0 | Exclude | 2103 | Exclude |
| Parham | 1991 | 98.5 | 82.2 | 0 | Exclude | 0 | Exclude | 2500 | Exclude |
| Soost | 1991 | 86.6 | 65.3 | 0 | Exclude | 0 | Exclude | 6701 | Exclude |
| Jones | 1996 | 74.5 | 25.3 | 0 | Exclude | 0 | Exclude | 6871 | Exclude |

Data from McCrory et al [28]

**Notes:**
Probability = probability of study estimate given parameter pair on the boundary
Decision = decision to include or exclude study from tailored meta-analysis
AR = applicable region defined by either *r* only, or *r* and *p.*

A study was excluded if the probability was <0.025. Using the Mahalanobis distance this corresponded to a critical value of 7.38

**Table(s)**

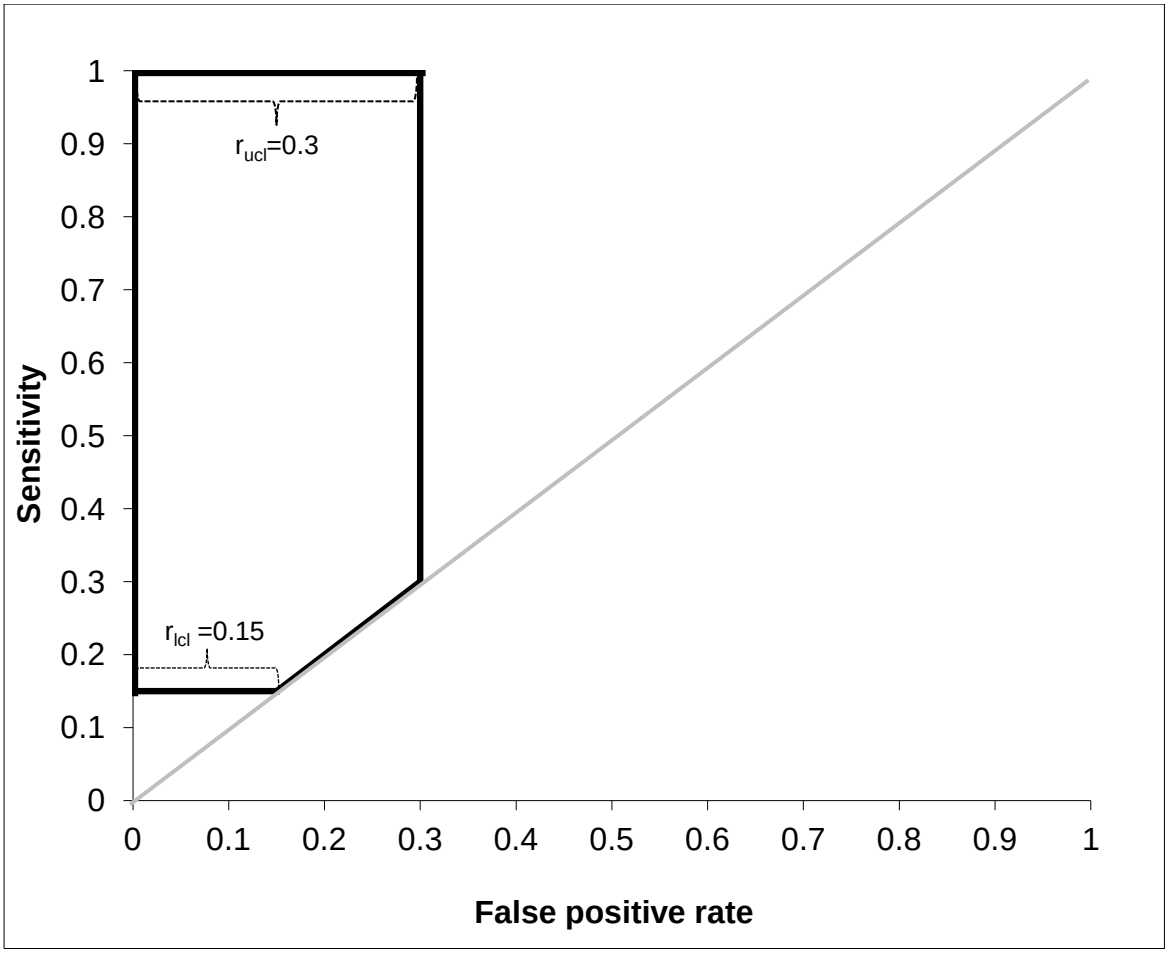


**Figure 1.** Hypothetical example demonstrating how the constraints define the applicable region (the bold continuous line).

The bold lines represent the constraint imposed by $r_{lcl} \leq r \leq r_{ucl}$, on *s* and *f* (inequality (1) in text), where $r_{lcl}$ and $r_{ucl}$ correspond to the lower and upper confidence limits for *r*. Here $r_{lcl}$ = 0.15 and $r_{ucl}$ = 0.30.

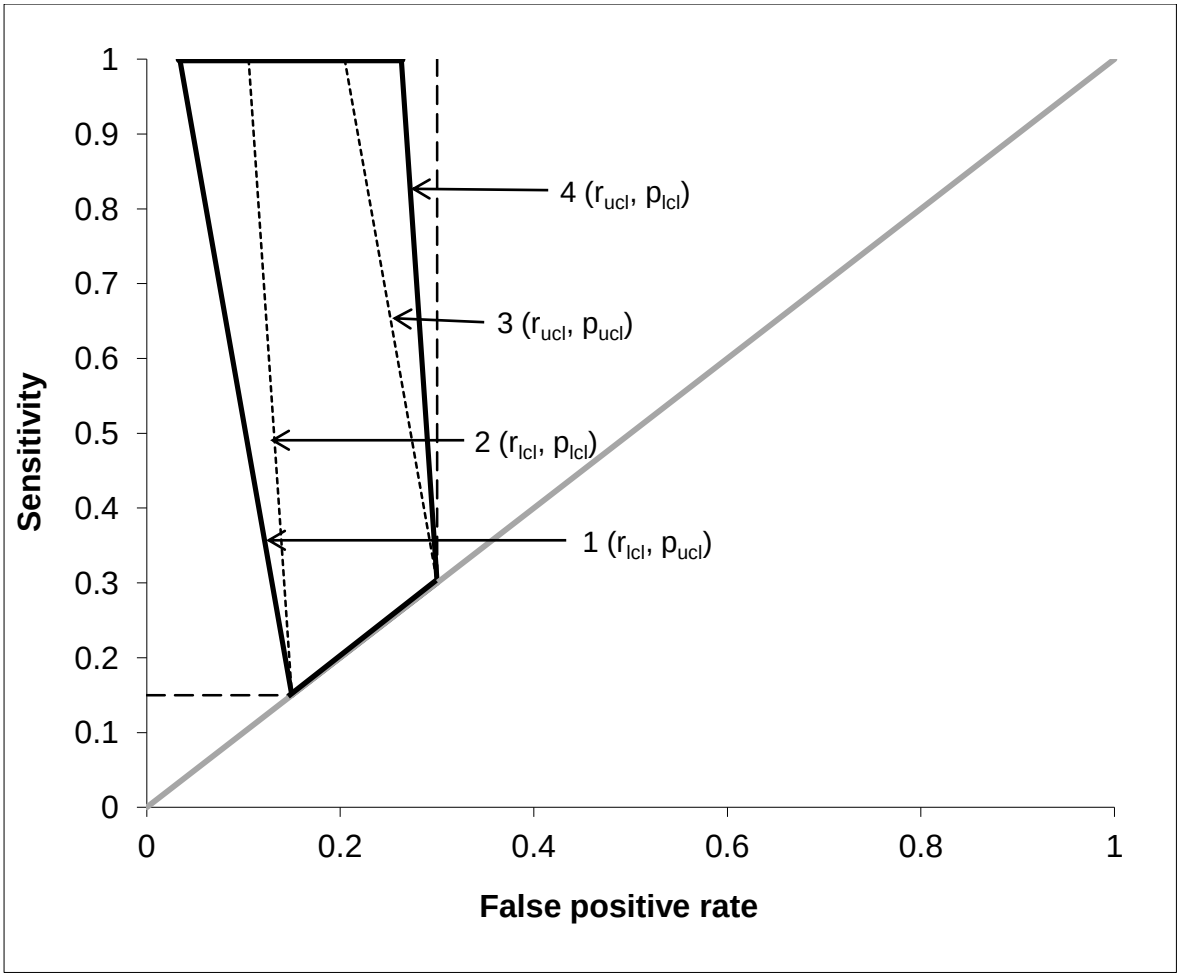


**Figure 2**. Hypothetical example demonstrating how the constraints define the applicable region (the bold continuous trapezium) after including estimates of the test positive rate, *r* and prevalence, *p.* The outer dashed lines represent the constraint imposed by $0.15 \leq r \leq 0.3$ on *s* and *f* . The lines represented by 1-4 are the result of imposing equation (2) in addition to (1) for $0.05 \leq p \leq 0.12$ and $0.15 \leq r \leq 0.3$. For example line 1 combines $s = (r/p) - ((1-p)f/p)$ for ($r = r_{lcl} = 0.15$, $p = p_{ucl} = 0.12$) and ($s \geq r$ and $f \leq r$) for ($r = r_{lcl} = 0.15$). Lines 2 and 3 are dominated by 1 and 4 and are therefore surplus (see text for explanation).

| Threshold | 1999-2000 | 2000-01 | 2001-02 | 2002-03 |
|---|---|---|---|---|
| borderline | 0.0919 | 0.0900 | 0.0861 | 0.0851 |
| mild | 0.0442 | 0.0431 | 0.0406 | 0.0406 |
| moderate | 0.0173 | 0.0166 | 0.0158 | 0.0156 |
| severe | 0.0076 | 0.0075 | 0.0071 | 0.0071 |

**Table 1.** The test positive rate, *r*, for each of the cytology thresholds in the NHS Cervical screening programmes over 4 years

| | All | AR (*r* only) Binomial (B1) | AR ( both *r* and *p*) Mahalanobis, ($D^2$) | Binomial (B2) |
|---|---|---|---|---|
| Included studies (n) | 68 | 17 | 16 | 12 |
| sensitivity | 72.8% (65.8-78.8) | 52.6% (37.9-66.9) | 49.6% (35.4-63.8) | 50.9% (35.8-66.0) |
| specificity | 75.4% (68.1-81.5) | 96.0% (93.0-97.8) | 96.6% (93.8-98.1) | 98.0% (95.4-99.1) |
| LR+ | 2.96 (2.37-3.68) | 13.21 (7.22-24.17) | 14.4 (7.4-28.2) | 25.6 (10.1-65.0) |
| LR- | 0.36 (0.30-0.44) | 0.49 (0.36-0.67) | 0.52 (0.39-0.70) | 0.50 (0.36-0.69) |

**Table 2**. Performance characteristics of the Pap test using conventional meta-analysis, the B1 method, $D^2$ method and the B2 method. Note AR = applicable region. 95% confidence intervals in brackets.

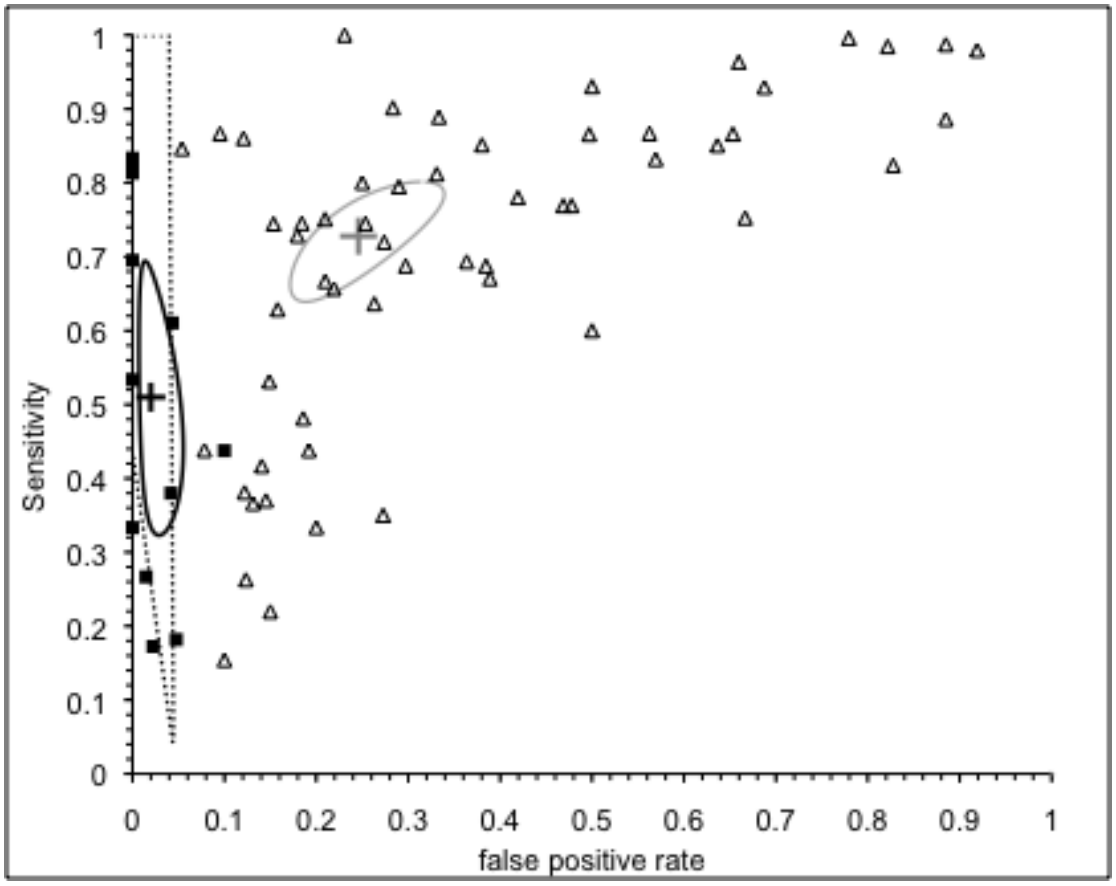


**Figure 3.** Tailored meta-analysis of the Pap test in the NHS using the B2 method.
The dotted line represents the applicable region using interval estimates for the test positive rate *r* and prevalence *p*. The bold squares are the included studies under the B2 method and the unfilled triangles are the excluded studies. The grey 'elliptical region' represents the confidence region (point estimate in centre) using conventional meta-analysis. The black ellipse is the associated confidence region enclosing the point estimate for tailored meta-analysis.

***Conflict of Interest/Financial Disclosure**

**Conflicts of interest**: None