DOI:10.1111/j.1365-2303.2008.00558.x

The development of the referral outcome diagram and an analysis of laboratory cancer detection rates in the English NHS cervical screening programme – is there an optimum level of detection of CIN 1 and CIN 2 lesions?

R. G. Blanks

Cancer Screening Evaluation Unit, Sir Richard Doll Building, Institute of Cancer Research, Cotswold Road, Sutton, Surrey, UK

Accepted for publication 27 November 2007

R. G. Blanks

The development of the referral outcome diagram and an analysis of laboratory cancer detection rates in the English NHS cervical screening programme – is there an optimum level of detection of CIN 1 and CIN 2 lesions?

Objective: To use routine annual data from the English cervical screening laboratories (KC61 returns) to evaluate individual laboratory return characteristics with particular reference to factors associated with sensitivity and specificity.

Methods: A graphical technique has been developed using data on referral to colposcopy and histological outcomes called a referral outcome (ROUT) diagram. The average grade of cervical intraepithelial neoplasia (CIN) detected (the mean CIN score, MCS) is plotted against the odds of a false-positive referral. Further analysis has been conducted to examine the relationship between the MCS and screen-detected invasive cancer rate.

Results: There are large variations in ROUT diagram positions of individual laboratories and the diagram can be used to identify laboratories for further investigation. These variations are strongly influenced by substantial differences in the rate of low-grade referrals and the MCS (and positive predictive value) are inversely related to the referral rate for low-grade cytology ($P < 0.001$). There is a strong association between high MCS values and increased screen-detected cancer rates ($P < 0.001$) particularly above an MCS of 2.2. The data can be re-formulated in terms of CIN 2 and CIN 3 only where it can be shown that the invasive cancer rate rapidly increases if the numbers of CIN 2 lesions detected drops below 50% of the number of CIN 3 lesions. Given the complexity of cervical screening this may best be viewed as a hypothesis generating observation, best tested by interventional studies.

Conclusions: The ROUT diagram represents a new and potentially interesting way of presenting annual return data. The national programme in England needs to balance the prevention of cancer against too many unnecessary referrals to colposcopy and the ROUT diagram, and associated data given in this paper may help toward this. Further research is required including examining the role of referral policy and threshold criteria in influencing low-grade referrals and the relationship between MCS and cancer detection rate.

Keywords: cervical screening, invasive cervical cancer, laboratory performance, cervical intra-epithelial neoplasia, screen-detected cancer

Introduction

Detailed information from laboratories in the English NHS cervical screening programme is provided by the annual Korner (KC61) returns. Cervical screening laboratories and their local associated screening activities (smear takers and colposcopy clinics) are difficult to evaluate for a number of reasons. Firstly, no randomised controlled trials are available with which to compare performance. Secondly, cervical screening detects a wide range of pre-invasive disease [cervical intraepithelial neoplasia (CIN) 1, 2 and 3] and there is uncertainty as to how much CIN 1 and 2 should be detected as most will regress naturally.[1] Thirdly, the

Correspondence:
Roger G Blanks, Cancer Screening Evaluation Unit,
Sir Richard Doll Building, Institute of Cancer Research,
Cotswold Road, Sutton, Surrey, SM2 5NG, UK.
Tel.: +0208 722 4008; Fax: +0208 722 4136;
E-mail: roger.blanks@icr.ac.uk

diagnostic test (colposcopy/histology) is not considered a reliable gold standard and there may be a particular lack of diagnostic consistency at the less severe end of the diagnostic spectrum.[2,3] Fourthly, the annual returns do not follow a cohort of women from invitation to histology. Fifthly, the Primary Care Trust (PCT), laboratory and colposcopy clinic annual returns are separate but not independent. Sixthly, comparisons between laboratories (and other components of the system) can be confounded by different background incidence (e.g. populations experiencing different risk factors or having different age distributions) and screening intervals, although a consistent invitation policy has recently been instigated. Finally, there can be problems with statistical stability of data from smaller laboratories and the problem that we require to interpret annual (cross-sectional) data that is best interpreted longitudinally over many years because of the potential impact of detecting earlier pre-invasive lesions on the prevalence of high-grade lesions.

The evaluation of cervical screening using annual return data can therefore often be far more complicated than experienced for example with the UK breast screening programme, which has the advantage of annual returns that follow a cohort of women from invitation to histological diagnosis, a more specific disease outcome (invasive cancer) and numerous randomised controlled trials to provide results against which to compare and evaluate performance.

From an epidemiological perspective, the ideal annual return from cervical screening laboratories would also follow a cohort of women from invitation to histology in a similar manner to that used in the NHS breast screening programme. In the absence of such return data, this paper exploits the available existing annual return information whilst recognising its limitations.

The paper uses information from the annual KC61 return (part C2), which includes a year's retrospective data from laboratories on the cytological reasons for referral and associated histological outcomes in women referred for further investigation. This return includes data which is a combination of cytology, colposcopy and histology reporting, each with its own inherent issues of false positive and false negative reporting and therefore it is important to emphasise that any summary statistics produced from the returns will reflect this complex nature. Care should therefore be taken in interpreting summary statistics, which are best used to select laboratories for further study rather than as definitive performance indicators.

The objective of the referral outcome (ROUT) diagram method is to display a substantial and comparable amount of information related to sensitivity and specificity from all laboratories in England on one diagram, in a way that allows immediate and informative comparison between those laboratories. The current purpose of the diagram is to help select laboratories (and their associated colposcopy clinics) for further investigations where the method suggests that the results indicate unusual or outlier characteristics that could merit further study. This will be best undertaken in a collective manner such that a number of laboratories are targeted for further study to examine why the return data are so different and whether they reflect true differences. If the diagram can be shown to indicate true and important differences then it could become a useful working tool in the evaluation of cervical screening. The axis variables are chosen such that they are both statistically robust using 1-year data (and therefore confidence limits, whilst useful, are not essential) and minimize confounding from factors such as varying background incidence. A similar style of analysis has previously proved effective in the NHS breast screening programme.[4]

Methods

The ROUT diagram plots the mean CIN score (MCS) against the odds of a false positive referral (OFP). The variables have been carefully chosen to give information that is as comparable as possible between laboratories and shows the laboratory return characteristics. Note that the phrase 'laboratory return characteristic' is used here in preference to 'laboratory performance measure' as the laboratory return data will be influenced by other factors, such as colposcopy clinic performance, as well as laboratory performance.

The model described in this paper considers referral leading to histology of CIN 3, CIN 2 and CIN 1 as true positive outcomes, and referral leading to 'HPV only', 'No CIN/HPV' or 'colposcopy no abnormality detected (colp NAD)' as a false-positive outcomes. Alternative models can be made, particularly where we assume that what some pathologists call CIN 1, others might call 'HPV only', and therefore the two should be grouped together. These alternative formulations including one based on only CIN 2 and CIN 3 as positive outcomes are considered in the Discussion section and Appendix.

Mean CIN score

The MCS is the average grade of CIN detected, calculated as following:

$$\mathrm{MCS} = (\mathrm{CIN3} \times 3 + \mathrm{CIN2} \times 2 + \mathrm{CIN1} \times 1)/(\mathrm{CIN3} + \mathrm{CIN2} + \mathrm{CIN1}).$$

The MCS can therefore range between 1 (only CIN 1 detected) and 3 (only CIN 3 detected). An example calculation is shown below using data from Table 1. Example:

$$\mathrm{MCS} = (245 \times 3 + 201 \times 2 + 176)/(245 + 201 + 176) = 2.11.$$

There is evidence that large lesions are more likely to be found from high-grade cytology and smaller lesions from low-grade cytology.[5,6] The assumption being made is that laboratories with high MCS scores are preferentially detecting mostly high-grade disease (reflecting a high specificity) and laboratories with low MCS scores are detecting a broader range of pre-invasive lesions (reflecting a lower specificity). Therefore, we make the assumption that laboratories detecting high proportions of CIN 1 (low MCS) are also likely to have detected most CIN 3 in their populations. Laboratories mostly detecting CIN 3 (high MCS) with much smaller proportions of CIN 1 and 2 detected are least likely to have detected all CIN 3. The opposing viewpoint to these arguments is that laboratories only detect what is available to detect and we therefore have to assume that laboratories with high MCS scores, mostly detecting CIN 3, only have populations with CIN 3 and very little CIN 1 and CIN 2. If this latter argument is true then the MCS is merely a mirror of the local population underlying disease levels rather than laboratory performance related to sensitivity/specificity trade-off. The MCS may of course to some extent reflect both these arguments, but the more it reflects the former argument then the more 'useful' it is as a measure.

There is some evidence that the MCS may not vary very much over a number of years for most laboratories. It is therefore reasonable to look for any relationship between the MCS and the screen-detected cancer rate as the screen-detected rate may be related to past levels of the MCS and therefore, if the MCS is reasonably stable, then the screen-detected cancer rate may also be related to the current MCS. Of interest is whether the screen-detected cancer rate increases with MCS or is independent of the MCS.

Table 1. Referral outcome (ROUT) diagram basic table and example data

Outcome	Designation	Weighting in mean CIN score	Example counts
CIN 3	True positive	3	245
CIN 2	True positive	2	201
CIN 1	True positive	1	176
HPV only	False positive	NA	86
No CIN/HPV	False positive	NA	132
Colp NAD	False positive	NA	45

Where additional unknown data are recorded on the part C2 return, e.g. result not known, it is assumed to have approximately the same distribution as known data.
Example: Mean CIN score = (245 × 3 + 201 × 2 + 176)/(245 + 201 + 176) = 2.11.
Example: Odds of false positive referral = (86 + 132 + 45)/(245 + 201 + 176) = 0.42.

Odds of false-positive referral

A false-positive referral is assumed to be any referral leading to a final diagnosis of 'HPV only', 'Colp NAD' or 'No CIN/No HPV' and true positive referral one resulting in a diagnosis of CIN 3, CIN 2 or CIN 1. The OFP can be calculated as:

$$\mathrm{OFP} = (\mathrm{HPV\ only} + \mathrm{No\ CIN/HPV} + \mathrm{Colp\ NAD})/(\mathrm{CIN3} + \mathrm{CIN2} + \mathrm{CIN1}).$$

Using the data in Table 1 we can calculate an example OFP. Example:

$$\mathrm{OFP} = (86 + 132 + 45)/(245 + 201 + 176) = 0.42.$$

This is interpreted as 0.42 to 1 or 42 false-positive referrals per 100 true-positive referrals. An OFP above 1 suggests that a greater proportion of referrals are false positive than true positive. Note that an odds of 1 equates to a proportion of 50% and the ROUT diagram *x*-axis, if preferred, can also be formulated in terms of percentage of referrals that are false positive. In the above example the proportion of false positive referrals would be 100 × ((86 + 132 + 45)/(86 + 132 + 45 + 245 + 201 + 176)) = 30%.

The laboratory in Table 1 would therefore be plotted at a co-ordinate of 2.11, 0.42 (or 2.11 and 30%, if the *x*-axis is the false-positive proportion). A further assumption common to the MCS and OFP is that the (usually small number of) 'results not known' are distributed in the same proportions as the known data.

Estimation of CIN rates, screen detected cancer rates and referral rates

In a further analysis, the MCS has been compared with the estimated invasive cancer detection rate per 10 000 women screened. The numerator is based on the number of stage 1A and frankly invasive (stage 1B or more) cancers detected by the laboratory and given in the KC61 part C2. The denominator (the number of women screened by the laboratory who could potentially contribute to Table C2) is unknown but can be estimated using the number of tests given in part A of the previous years return (as part C2 is retrospective data). For the whole of England (DH statistical Bulletin 2003–2004) four million smears were analysed from 3.6 million women; therefore the underlying number of women contributing smears to each laboratory can be estimated for each laboratory as 0.9 × the number of smears. This equation is used to estimate the number of women from the number of smears reported on the KC61 part A1 Grand Total of smears from all sources. The referral rate (%) to colposcopy can be estimated as 100 times the total number of women referred to colposcopy (reported on the KC61 part C2) divided by the estimated population for the appropriate year as discussed earlier. Other measures such as the referral rate for borderline/mild smears can be calculated as 100 times the number of referrals from borderline and mild smears (on part C2) divided by the number of women.

All statistical analyses have used the statistical software STATA version 8 (Stata Corporation, College Station, TX, USA).

Results

Analysis of 2004/2005 laboratory data using the ROUT diagram

Figure 1 shows the ROUT diagram for 141 laboratories using data from the 2004/2005 part C2 return. The median MCS is 2.1 and the median OFP is 0.58. The 10th to 90th percentile range for the MCS is 1.86–2.35 and for the OFP 0.31–1.05. The diagram is divided into four quadrants using an MCS value of 2 and an OFP value of 1 to make these quadrants. Figure 1 also includes an accompanying table with detailed information on four marked laboratories at the extremes of each quadrant. The ROUT diagram shows no direct relationship between MCS and OFP but laboratories at the extremes of the quadrants, (in the direction of the arrows) clearly show very different return characteristics.

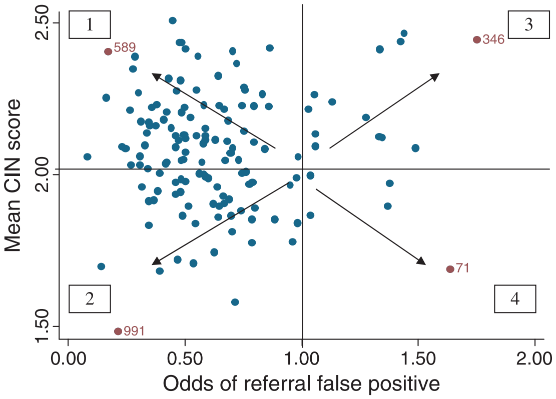

Fig 1 Accompanying table: Detailed return information for indicated outlier laboratories

Lab*	CIN 3	CIN 2	CIN 1	HPV only	No CIN/ HPV	Colp NAD	Total	Mod+ referral rate (%)	Bord/Mild referral rate (%)
589	140 (48%)	69 (24%)	39 (13%)	16 (6%)	8 (3%)	18 (6%)	290	1.8	0.6
71	39 (7%)	79 (15%)	109 (18%)	64 (11%)	87 (15%)	220 (37%)	598	0.9	3.6
346	236 (23%)	65 (6%)	71 (7%)	104 (10%)	35 (3%)	512 (50%)	1023	1.2	1.3
991	69 (12%)	88 (16%)	308 (55%)	14 (2%)	81 (14%)	3 (1%)	563	1.0	2.7

* CSEU confidential laboratory identification code

Figure 1. Referral outcome (ROUT) diagram using data from 2004/2005 with arrows in the direction of the quadrant extremes and indicating four 'outlier' laboratories.

Extreme of quadrant 1. **High MCS and low OFP** – These laboratories detect mostly CIN 3, and much less CIN 2 and CIN 1 and have a tendency to have very few false-positive referrals. It is possible that these laboratories emphasise specificity rather than sensitivity and have low referral rates from mild and borderline cytology.

Extreme of quadrant 4. **Low MCS and High OFP** – The opposite of extreme quadrant 1. These laboratories have a high proportion of CIN 1 and also a high false-positive rate. They tend to have a high proportion of referrals from borderline and mild dyskaryosis and may be recalling on only a slight suspicion and may be emphasising high sensitivity at the expense of specificity.

Extreme of Quadrant 3. **High MCS and High OFP** – These laboratories have a tendency to detect relatively little CIN 1 and 2, but have a high OFP referral. They tend to have a large numbers of colposcopy – no abnormality detected (Colp NAD) outcomes, which in some cases can be from a high number of repeat inadequate referrals, but can also be from borderline and mild referrals. A possible explanation is that biopsies are not being taken from lesions thought on colposcopy to confirm low-grade cytology.

Extreme of quadrant 2. **Low MCS and Low OFP** – These laboratories have a tendency to detect large amounts of CIN 1, but also have few false-positive referrals. These laboratories would appear to have a high sensitivity for CIN 1, but also a high specificity.

Relationship between MCS and screen-detected cancer rates

Tables 2a and b show laboratories stratified into six groups depending on MCS and relate this to cancer and CIN rates per 10 000 as well as positive predictive value (PPV) and referral rates for low and high-grade cytology.

Table 2a shows that there is a strong relationship between MCS and screen-detected (1A and 1B+) cancer detection rates, which is highly statistically significant (test of trend $P < 0.001$). The 25 laboratories in the category with the lowest MCS (group 1:

Table 2a. Relationship between the mean CIN score (MCS) and the screen-detected cancer (1a+) rates using data from 2004/2005

MCS group (n) range	Group median MCS	Est total women*	Total invasive cancers	Invasive rate per 10 000 (rate ratio)	Group median % referral rate from bord/mild	Group median % referral rate from moderate or worse	Group median PPV (%)
1 (25) <1.9	1.84	538709	88	1.63 (1.00)	2.03	1.11	71.0
2 (21) 1.9–1.99	1.96	520742	123	2.36 (1.45)	1.41	1.16	76.0
3 (29) 2.0–2.09	2.05	872122	202	2.32 (1.42)	1.44	1.13	78.1
4 (26) 2.1–2.19	2.13	537016	135	2.51 (1.54)	1.12	1.15	76.8
5 (22) 2.2–2.29	2.23	559298	173	3.09 (1.90)	1.30	1.17	81.4
6 (18) >=2.3	2.41	518871	186	3.58 (2.20)	0.93	1.19	81.6

PPV, positive-predicted value.
*Estimated as 0.9 × total smears.

Table 2b. Estimated absolute rates of CIN 1, 2 and 3 by MCS group

MCS group (*n*)	Median women	CIN 1 (rate per 10 000)	CIN 2 (rate per 10 000)	CIN 3 (rate per 10 000)	CIN 1 & 2 to CIN 3 ratio	CIN 2 to CIN 3 ratio	CIN 1 to CIN 2 & 3 ratio
1 (25)	19828	4831 (89.7)	2695 (50.0)	2698 (50.1)	2.79	1.00	0.90
2 (21)	24223	3494 (67.1)	2837 (54.5)	3037 (58.3)	2.08	0.93	0.59
3 (29)	25825	4556 (52.2)	4282 (49.1)	5306 (60.8)	1.66	0.81	0.48
4 (26)	21536	2214 (41.2)	2091 (38.9)	3320 (61.8)	1.30	0.63	0.41
5 (22)	23666	2140 (38.3)	2453 (43.9)	4243 (75.9)	1.08	0.58	0.32
6 (18)	24285	1197 (23.1)	1718 (33.1)	3981 (76.7)	0.73	0.43	0.21

MCS, mean CIN score; CIN, cervical intraepithelial neoplasia.

median 1.84) have a total screen-detected cancer rate of 1.63 per 10 000 women whereas the 18 laboratories with the highest MCS (group 6: median 2.41) have a total screen-detected cancer rate of 3.58 per 10 000 women. Table 2a also gives estimates of the median estimated referral rate (%) of the laboratories in each group for borderline or mild referrals and for moderate or worse referrals. This shows that laboratories with low MCS scores have considerably more referrals from borderline and mild smears and the inverse trend between MCS and referral rate for borderline/mild referrals is highly significant ($P < 0.001$). For example, in the lowest group (MCS group 1), an average (median) of 2.03% of women are referred from borderline/mild dyskaryosis, which is much higher than for the other groups. In contrast, the moderate or worse referral rate of 1.11% is more similar to that of the other groups. Table 2a also shows the median PPV for each group of laboratories. The PPV is defined as the proportion of referrals just from moderate or worse dyskaryosis that have histology of CIN 2 or worse. This shows that, on average, laboratories with lower PPVs tend to be associated with greater overall detection levels of CIN 1 and CIN 2 relative to CIN 3 and also with lower screen detected cancer rates ($P < 0.01$).

Table 2b includes additional information giving the estimated absolute rates of CIN 1, 2 and 3 by MCS group as well as ratios of CIN1/2 to CIN 3, CIN 2 to CIN 3 and CIN 1 to CIN 2/3. This information can be interpreted as suggesting that high rates of CIN 1 and CIN 2 detection as in MCS group 1 lead to lower rates of detection of CIN 3 and crucially of invasive cancer. Low rates of CIN 1 and CIN 2 detection as in MCS group 6 have the opposite effect and lead to higher rates of detection of CIN 3 and invasive cancer. Both CIN 1 and CIN 2 are more likely to regress than progress[1] and therefore the majority of these lesions will represent overdiagnosis as it is not known which lesions will regress and which will progress. Is there an optimum trade-off that can be suggested from the data?

To increase statistical stability 2-year data (2004/2005 and 2005/2006) are combined in Figure 2 and Table 3, which suggests increased screen-detected cancer rates are associated with a MCS above 2.2. This suggests that for a laboratory seeking to maintain a high specificity and also a high reduction in the occurrence of invasive cancers a MCS of around 2.1 to 2.2 may be optimal. Of interest here is that the observed median MCS of laboratories is currently 2.1, although the range varies from about 1.5 to 2.5. Background incidence and screening interval cannot be allowed for and it is possible, for example, that laboratories in areas of high background incidence could tend toward detecting mostly CIN 3 and in low incidence areas detecting higher proportions of CIN 1 and CIN 2. The relationship shown, could, at least theoretically, be more an association rather than necessarily causation and therefore we require some caution in our interpretation. A re-formulation of the MCS based on only CIN 2 and CIN 3 rather than CIN 1, 2 and 3 (and notated with a superscript as MCS^{23} to avoid confusion) is shown in the Appendix. It can be shown that an MCS of 2.2 roughly corresponds in practice to a MCS^{23} of 2.66, which equates to a laboratory detecting half as many CIN 2 lesions as CIN 3 lesions. Higher screen detected invasive cancer rates are therefore associated with a laboratory detecting half or less CIN 2 than CIN 3 lesions.

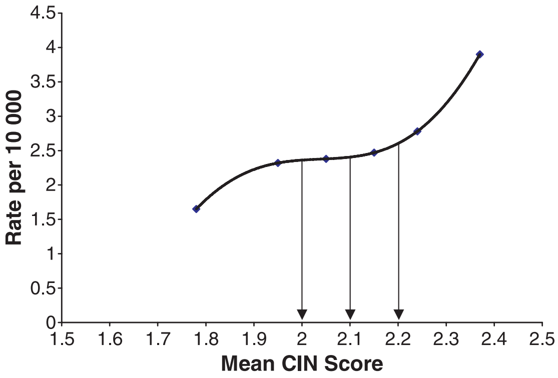

Figure 2. Screen-detected cancer rate against mean CIN score (MCS) for 2-year data (data from 2004/2005 and 2005/2006).

Discussion

The ROUT diagram has been developed to provide a simple, but informative way of examining data from many laboratories on one diagram using the current annual return information. The detailed information from the four 'outlier' laboratories from Figure 1 shows clearly how different the return data are, with lab 589 having CIN 3 as the most common outcome from referral, but with laboratories 71 and 346 having 'no abnormality detected at colposcopy' as the most common outcome. Lab 589 (extreme quadrant 1) has a particularly low low-grade (borderline/mild)

Table 3. Screen detected cancer rate per 10 000 women by MCS category for 2-year data

MCS range	No. of labs*	Mean MCS	Cancers	Estimated number of women	Rate per 10 000 (95% CI)
1.6–1.89	17	1.78	114	690 206	1.65 (1.36–1.98)
1.9–1.99	21	1.95	279	1 203 079	2.32 (2.06–2.61)
2.0–2.09	33	2.05	419	1 758 788	2.38 (2.16–2.62)
2.1–2.19	24	2.15	312	1 263 226	2.47 (2.20–2.76)
2.2–2.29	20	2.24	275	990 596	2.78 (2.46–3.12)
2.3–2.59	19	2.37	424	1 088 024	3.90 (3.54–4.29)

MCS, mean CIN score.
*134 laboratories have data for 2 years.

referral rate and may be mostly aiming to detect CIN 3 whereas the philosophy of lab 71 (extreme quadrant 4) may be to try and detect as much CIN 1, 2 and 3 as possible and has a high low-grade referral rate accordingly and a high percentage of colp NAD outcomes possibly as a consequence. Laboratory 346 (extreme quadrant 3) has low amounts of CIN 1 and 2 relative to CIN 3 but has a high proportion of colp NAD outcomes, and one possible explanation is that the laboratory may refer many women, but not do so many biopsies, if no disease is clearly evident at colposcopy. Laboratory 346 also has a high proportion of referrals from repeat inadequate smears, which is likely to be heavily reduced when LBC impacts on the return data. Laboratory 991 (extreme quadrant 2) has unusual return data with regard to the distribution of CIN 1 and 'HPV only'. At the extreme of quadrant 2, it is not clear how laboratories are able to detect large amounts of low-grade disease without also having a high number of false-positive referrals. Intuitively a laboratory detecting a high proportion of CIN 1 might also be expected to have a high number of outcomes of 'HPV only' and no disease (and tend to quadrant 4 rather than quadrant 2). Under such circumstances, the ROUT diagram might naturally be expected as a scatterplot of points from the top left curving around to the bottom right and laboratories at the extreme of quadrant 2 are therefore unexpected.

The change to a consistent screening interval across England, which will reduce confounding, and the phased introduction of LBC, which will reduce the number of referrals from repeat inadequate smears, should both make this methodology increasingly more useful. The data used in this paper are from the 2004/2005 part C2, which is a year's retrospective data from the screening year 2003/2004 during which period only three laboratories will have had women referred to colposcopy following cytology results using LBC. Because of the complexity of cervical screening, the nature of the return data and the limitations of observational studies, the results shown in this paper can at best be considered as 'strongly suggestive' rather than definitive. Nevertheless, the final judgement on an analysis such as this rests with whether this type of summary statistics approach can be shown to be useful in practice.

Several alternative formulations of the ROUT diagram have been explored. A formulation detailed in the Appendix is where only CIN 2 and CIN 3 (i.e. treatable disease) is considered as true-positive outcomes and CIN 1 is grouped together with 'HPV only' and other negative histology outcomes as false-positive referrals. The advantage of this approach is that the distinction between CIN 1 and 'HPV only' is no longer important, but the MCS is only based on CIN 2 and CIN 3 and is therefore potentially less informative because of smaller numbers. There is the issue of observer variability between the CIN 2 and CIN 3 grading, and perhaps more controversially CIN 1 is being considered as a 'false positive' referral on the basis that it is not treatable disease and most would regress. This alternative formulation still shows laboratories 71, 589 and 346 as 'outliers' and clearly shows that laboratories with a high proportion of CIN 2 relative to CIN 3 tend to have a higher odds of false positive referral. The diagram shows more evidence of curving from the top left to bottom right as might be anticipated. Laboratory 991 moves from the bottom left to the bottom right of the diagram as the distinction between CIN 1 and 'HPV only' is no longer important. It should be stressed that these variations on the ROUT diagram generally tend toward similar conclusions and that the purpose of this paper is to introduce the concept of the diagram, and its uses, rather than to be too dogmatic about the exact formulation for the derivation of the axes at the present time.

The variations in ROUT diagram positions are likely to be mostly related to different 'philosophies' or 'criteria' held by cytopathologists as to what level of low-grade disease should be referred. For low-grade disease there are two aspects to this 'philosophy of referral', firstly the 'threshold criteria' as to what cellular changes relate to cytology categories and particularly borderline changes or mild dyskaryosis, one cytopathologists borderline could be another cytopathologists negative.[7] Secondly, there is the referral policy, which relates to whether the referral to colposcopy is from only one borderline or one mild, up to a maximum of three borderlines or two milds. Therefore, a laboratory with a low threshold of what cellular changes reflect borderline or mild cytology, and which also refers on only one occurrence of borderline or mild, will have a far greater referral rate for the same population of women than a laboratory which has both a higher threshold of what cellular change reflects each category and also more conservatively only refers on three borderline smears or two mild smears. These variations are likely to explain why in Table 2a, the median referral rate from moderate or worse smears is not strongly related to the MCS (moderate or worse dyskaryosis has lower between observer variation and that there is no discretion over the referral policy), but why the referral rate from borderline/mild referrals is more strongly related to the MCS. For example, laboratories in MCS group 1 (Table 2b) with high referral rates for borderline and mild cytology detect not only far more CIN 1 than laboratories in group 6 but also more CIN 2. By detecting more CIN 1 and CIN 2, less CIN 3 may be being detected because the CIN 3 is not there, having been found at an earlier grade of CIN and before it could progress to CIN 3. Most importantly laboratories with the highest CIN 3 rates also have the highest screen-detected cancer rates and Tables 2a and 2b collectively provide evidence that laboratories that mostly detect CIN 3 with much lower rates of CIN 2 may risk an increase in the invasive cancer rate, as some lesions are not detected as CIN 3 but progress to become screen-detected invasive cancer.

It can be argued that ensuring a more even service across the country as suggested as a goal of quality assurance[8] also implies ensuring a more even low-grade referral philosophy. However, the optimum position for a laboratory on the ROUT diagram is open to debate and it is not the principal purpose of this paper to suggest a specific region to be aimed at, but really to indicate laboratories (at the extremes of the quadrants), which may have very different referral philosophies and where further investigative studies may be informative. For example, a laboratory at the extreme of quadrant 1 on the ROUT diagram with a high PPV can argue, that even if it had a higher screen-detected cancer rate, that this is justifiable in that it is maintaining the highest specificity because few women are being referred who do not have advanced pre-invasive disease. The laboratory could argue that if additional invasive cancers are being 'allowed through' they are counter balanced by a very low false-positive rate. The ROUT diagram can therefore also be used in a manner where it requires 'outlier' laboratories to discuss further their data at regional QA visits and therefore could potentially become a useful tool in this respect.

An implicit assumption made in this work is that whilst the smears analysed by laboratories will be from populations of women with different levels of background disease incidence, the relative proportions of CIN 1, 2 and 3 in those populations should be approximately the same. Therefore, if we assume that in a lower risk population there are, for example, 2% of women with CIN 1, 1% with CIN 2 and 1% with CIN 3, then we make the assumption that in a population with twice the disease risk there is 4% CIN 1, 2% CIN 2 and 2% CIN 3. With this assumption, the MCS is independent of background disease risk and all laboratories are therefore comparable. If as discussed in the Methods section the MCS merely mirrors local disease variation, then the measure will be less useful. However, this would seem unlikely, given that the MCS is related to referral rate and because local areas only having women with CIN 3 and not much CIN 1 or 2 would seem unlikely. As time progresses, laboratories detecting large amounts of CIN 1 and 2 will have less CIN 3 detectable, as fewer lesions will be allowed to progress. The effect of this, which may already be happening (laboratory 71 data with a low number of detected CIN 3 lesions may suggest this), is likely to be to widen the gap between the lowest and highest MCS and therefore will be beneficial to this analysis method. This of course suggests that an 'ideal' study would be longitudinal rather than cross-sectional and at the outset of a national screening programme cohorts of women should be subject to long term follow-up.

If we assume that the association between MCS and screen-detected cancer rate is causal, we are making two assumptions. Firstly that the MCS as measured now reflects the MCS in the past, and, secondly, that sufficient time has elapsed for some of the undetected CIN from laboratories with past high MCS scores to become some of the present screen-detected invasive cancers. There is evidence that the MCS over at least a few years is reasonably similar but poor quality data do not allow this to be tracked too many years in the past. Either undetected CIN 1 and particularly CIN 2 'get through the system' to become cancer, perhaps particularly in late or 'irregular attenders' to screening or it is necessary to detect a minimum level of CIN 2 to be sure of detecting most or all CIN 3 and it is some of the undetected CIN 3 that is most likely progressing to become invasive cancer. How quickly can these lesions progress? There is likely to be a huge variation in progression rates and although progression for the average lesion may be slow there is clear evidence that some lesions progress quickly. For example in 2004, in the UK, there were two invasive cancers reported in women aged 15–19 and 58 in women 20–24 (Source: Cancer Research UK). The progression of these lesions must have been rapid. Progression rates of moderate dyskaryosis (much of which would be CIN 2) have been estimated as 16% within 2 years and 25% within 5 years,[1] again suggesting that progression of lesions in many cases is not necessarily slow. Every cancer cannot be prevented even if all women attended screening and there is of course no 'correct' solution; harm must be balanced against benefit as in all screening programmes and a proportion of false-positive referrals is inevitable if sensitivity for high-grade CIN is maintained (and cancer rates held low). A complicating factor as suggested above is the proportion of the local population who are 'irregular attenders' in an area. A laboratory in an area with a high proportion of these women will, perhaps, have a stronger argument for detecting higher levels of lower grade disease.

It would also be useful if the KC61 part C2 returns contained further details (e.g. in an annex) on screen-detected cancer origins. There are four main categories of interest. Firstly and secondly, screen-detected cancers resulting from a woman's prevalent (first) screen and screen-detected cancers resulting from prevalent screens in a previous non-attender. These prevalent screen cancers are of particular interest in that they are clearly not preventable by the screening programme and therefore form an important subgroup. Thirdly, there are cancers occurring in women who are late attenders, e.g. occurring more than 42 months after the previous screen for a 3-year invitation policy and fourthly, cancers occurring in women screened on time or early, e.g. within 42 months for a 3-year policy. Some of the cancers occurring in the last two groups may be not only potentially preventable, but also represent those cancers that we are most interested in as performance measures. Finally some cancers detected could be 'symptomatic' and identification of any such cancers would be useful. This work would be complementary to the invasive cancer audit work currently being undertaken by the NHS cervical screening programme.[9,10] An additional factor which cannot be considered in this study, but which is of importance, is the role of interval cancers and these can only be studied in a wider setting. A further consideration for the programme would be the use of 'cohort' based returns where women or women-episodes can be used as the denominator. Using such a return measures such as the CIN 2+ detection rate and the referral rate to colposcopy would be directly measurable and this will be the subject of a further paper.

Conclusions

The NHS cervical screening programme needs to balance the prevention of cancer with the consequences of too many unnecessary referrals to colposcopy and excessive overtreatment of pre-invasive disease not destined to become invasive. It is hoped that the ROUT diagram method may help quality assurance and laboratory staff examine their own data in relation to all other laboratories. Furthermore, it has been shown that one component of the ROUT diagram, the MCS can be related to the screen-detected cancer rates and that estimates of an optimal MCS can be made that may help maximise the balance between sensitivity and specificity. Supporting information to the ROUT diagram should include the screen-detected invasive cancer rate, the PPV and some measure of the proportion of women who are irregular attenders. Further work is required.

Acknowledgment

This work was funded by the NHS Cancer Screening Programme. The Cancer Screening Evaluation Unit (CSEU) also receives funding from the Department of Health Policy Research Programme. The views expressed in this publication are those of the author and not necessarily those of the Department of Health or NHS Cancer Screening Programme.

References

1. Holowaty P, Miller AB, Rohan T, To T. Natural history of dysplasia of the uterine cervix. *J Natl Can Inst* 1999;**91**:252–8.
2. Robertson AJ, Anderson JM, Beck JS *et al.* Observer variability in histopathological reporting of cervical biopsy specimens. *J Clin Pathol* 1989;**42**:231–8.
3. Pretorius RG, Bao YP, Belinson JL, Burchette RJ, Smith JS, Qiao YL. Inappropriate gold standard bias in cervical cancer screening studies. *Int J Cancer* 2007;**121**:2218–24.
4. Blanks RG, Moss SM, Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. *J Med Screening* 2001;**8**:24–8.
5. Jarmulowicz MR, Jenkins D, Barton SE, Goodall AL, Hollingworth A, Singer A. Cytological status and lesion size: a further dimension in cervical intraepithelial neoplasia. *Br J Obstet Gynaecol* 1989;**96**:1061–6.
6. Tidbury P, Singer A, Jenkins D. CIN 3: the role of lesion size in invasion. *Br J Obstet Gynaecol* 1992;**99**:583–6.
7. The Borderline Nuclear Changes National Slide Exchange Study Group. Do borderline nuclear changes in gynaecological cytology constitute a reliable reporting category?. *Cytopathology* 2002;**13**:220–31.
8. Johnson J, Patrick J. *Achievable Standards, Benchmarks for Reporting, and Criteria for Evaluating Cervical Cytopathology*, NHSCSP publication no. 1. Sheffield: NHSCSP; 2000.
9. Sasieni P, Adams J, Cuzick J. Benefit of cervical screening at different ages: evidence from the audit of screening histories. *Br J Cancer* 2003;**89**:88–93.
10. Patrick J. *Audit of Invasive Cervical Cancers*, NHSCSP publication no. 28. NHSCSP: Sheffield; 2006.

Appendix: an alternative formulation of ROUT diagram with CIN 2 and 3 as disease positive

Disease positive: CIN 2 and 3

Disease negative: CIN 1, HPV only, No CIN/HPV and Colp NAD

MCS = (3 × CIN 3 + 2 x CIN 2)/(CIN 3 + CIN 2)

MCS >2.5 indicates more CIN 3 than CIN 2 detected

MCS <2.5 indicates more CIN 2 than CIN 3 detected

OFP = (Colp NAD + No CIN/HPV + HPV only + CIN 1)/(CIN 2+CIN 3)

In the alternative formulation of the ROUT diagram shown in Figure A1, laboratories 589, 346 and 71 remain outliers. A major change is seen in the position of laboratory 991 because there is no longer any distinction between CIN 1 and 'HPV only', both which are now considered in this formulation as 'false positive' referrals in the sense that they represent disease not requiring treatment.

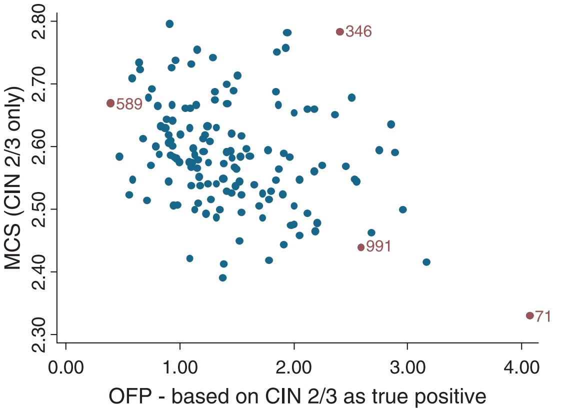

Figure A1. An alternative formulation of referral outcome (ROUT) diagram with CIN 2 and 3 as disease positive.

Table A1. Screen-detected cancer rate per 10 000 women by MCS category

MCS range	No. of labs	cancers	Estimated women	Rate per 10,000
2.3–2.52	38	228	983 447	2.32
2.53–2.57	30	158	763 034	2.07
2.58–2.63	37	225	916 688	2.45
2.64–2.8	36	296	883 588	3.35

MCS, mean CIN score.

Table A1 also shows that laboratories with the highest MCS scores in this formulation have the highest screen-detected invasive cancer rates. Correlating the main MCS in the paper (which we can term MCS^{123}) with the MCS based on CIN 2 and 3 as disease positive (which we can term MCS^{23}) then an MCS^{123} of 2.2 (the point at which the screen-detected cancer rate rapidly increases) corresponds to an MCS^{23} of 2.66, based on the regression model; $MCS^{123} = 1.72 \times MCS^{23} - 2.37$. From Table A1, it can be seen that the invasive cancer rate rapidly increases at an MCS^{23} of about 2.66. An MCS^{23} of 2.66 occurs roughly when the number of CIN 2 lesions detected is less than half the number of CIN 3 lesions detected. The hypothesis can therefore be reduced to the very simple observation that the invasive cancer rate shows a large increase when the number of detected CIN 2 lesions falls to less than half the number of detected CIN 3 lesions.