The application of Voice Risk Analysis within the Benefits System

Evaluation Report

September 2010

DWP Department for Work and Pensions

Contents

Acknowledgements

The authors would like to thank David Barr, Peter Wright (Principal Scientific Adviser) and John Viggers at the Department for Work and Pensions for their guidance and advice.

Thanks also to the Voice Risk Analysis providers at Capita and Digilog for their provision of training and engagement in the discussion over findings.

Finally, we extend our thanks to all operational staff involved in the pilot in the participating local authorities

Authors

James Ablewhite - Project Manager

Simon Lunn, Dr Jon Hyde and Sonal Patel - Departmental statisticians

Abbreviations

CCO	Customer Compliance Officer
CSA	Customer Service Agents
DWP	Department for Work and Pensions
HB	Housing Benefit
IS	Income Support
JCP	Jobcentre Plus
JSA	Jobseeker’s Allowance
LA	Local Authority
ROC	Receiver Operating Curve
VRA	Voice Risk Analysis

Abstract

It is estimated that the Department for Work and Pensions (DWP) overpays £3.1bn[1] in benefit due to fraud and error. Whilst proportionally small at 2.1% of all benefit expenditure, it represents a significant cost to the taxpayer. The Department is keen to exploit new solutions to the problem of fraud and error, and Voice Risk Analysis (VRA) is reported to have been used successfully in the private sector. The Department helped to fund trials of the technology in 24 local authorities (LAs) on the processing of new claims, in-claim reviews and reported changes of circumstance to Housing Benefit (HB) which took place between August 2008 and December 2010. This report details the evaluation of the trials and the resulting conclusions drawn.

[1] Source: ‘Fraud and Error in The Benefit System October 2008 – September 2009’

Introduction

DWP announced an intention to look at the potential for new technology to combat fraud (VRA was given as an example) in the strategy document "Reducing fraud in the benefit system - achievements and ambitions" published in October 2005.

VRA combines the measurement of levels of voice stress with behavioural analysis and intelligent scripting to enable the detection of truthful statements. It is already in use in the private sector, for example in better assessing the risk associated with insurance claims. The overall objective of the pilot was to see whether the same techniques could be successfully applied within the benefit system. This evaluation does not comment on the technological or any other aspects of VRA, only whether its use in benefits administration has proved successful.

Adopting a risk based approach to claims allows an assessment to be made of a customer's propensity to commit fraud and to then apply an appropriate level of verification of their information. Presently all customers receive the same 'check everything' approach. By determining risk those customer adjudged to be low risk can receive a lower level of scrutiny which allows faster claims processing at lower cost with less intrusion. But at the same time, by concentrating fraud and error detection and prevention work on those high risk customers that are more likely to have incorrectness within their claims, levels of fraud an error should be reduced.

VRA is a commercial product provided by Capita group. The business case for using VRA is based on an assumption that by 'fast tracking' low risk customers efficiency savings will more than pay for the VRA process. DWP funding was provided for the additional costs of piloting the process and providing management information. Whilst DWP owns benefit policy, LAs are responsible operational delivery of HB. As such, they are free to purchase VRA and in fact we are aware of nine LAs that have chosen to continue at their own expense.

Aims and objectives

Aim: To evaluate the ability of VRA to be successfully applied within the benefit system.

The pilot was designed to determine whether VRA could successfully differentiate between high and low risk and even if it could, would it let too much fraud and error in on low risk cases or wrongly target honest customers as high risk. Other indicators of success were: cost-effectiveness; suitability for application to a sufficiently large proportion of the benefit caseload and acceptability to customers and staff.

Each LA provided monthly management information that recorded the breakdown of calls by low and high risk; the outcomes of these calls; and findings from audit validation visits to confirm whether the correct risk score had been determined.

Success criteria

The criteria for success was that VRA correctly identified overpaid claims as high risk, whilst not capturing a disproportionate number of claims without overpayments in the high risk group at the same time.

The project was instigated to co-ordinate the approach and activities required across 24 LA pilot sites to test whether LAs could successfully integrate the VRA process within their benefit systems to deliver a reliable risk score. In addition the project's overriding aim was to determine whether the process could provide a risk based verification solution suitable wider use across the benefit system.

The pilot

A first phase of an evaluation programme of a risk score product utilising VRA took place between May 2007 and July 2008 and covered new claims to Income Support and Jobseeker's Allowance in Jobcentre Plus and six trials on reviews of existing benefits in LAs. The results of the evaluation and an announcement of further trials were made in a Written Ministerial Statement in March 2008. This provided the results of the first phase of trials and concluded that further research was necessary in order to evaluate the effectiveness of VRA when used to risk score benefits.

Consequently, a second phase of evaluation within 24 LAs was designed to provide a broader evidence base and a clearer indication of the VRA process's capacity to distinguish reliably between high and low risk cases; test affordability and to undertake social research. In response to parliamentary questions ministers announced that results would be available in spring 2010. The pilot work finished in December 2009.

Each participating LA chose whether to test VRA on:

- HB new claims,
- reported changes of circumstance,
- in-claim review where a periodical review of the a customers claim is conducted usually by post or a face to face interview, or
- a combination of the three.

They also chose whether to conduct the telephone call themselves or outsource the call to Capita to make the calls from their call centre. A list of the LAs and whether VRA calls were made in-house by the LA or as a managed service provided by Capita is attached at Appendix 2.

Selection criteria

Random selections of suitable clients were selected for each LA trial. VRA is a complimentary service provided in addition to traditional methods of declaring information and it is understood that a telephone based service is not suitable for all customers. Customers were deemed to be unsuitable for the trial if, for example, they did not speak English as a first language, had a disability to their hearing or cognitive function that made telephone conversation difficult or did not have access to a telephone.

A statement was read at the beginning of the call explaining that; " Before we start, I must tell you, under the Data Protection Act of 1998, that our calls are recorded and analysed using techniques and technology for the purpose of fraud prevention and detection, training and quality control, and may be reviewed later to check the details you have given." Customers were free to decline to take part in the VRA call without giving an explanation and provide information by traditional methods.

Training

Each call handler received training in the use of the technology and how to identify potential risk by analysing the behaviours exhibited by customers. The training was provided by Capita and typically lasted for four days followed by on-going mentoring throughout the pilot.

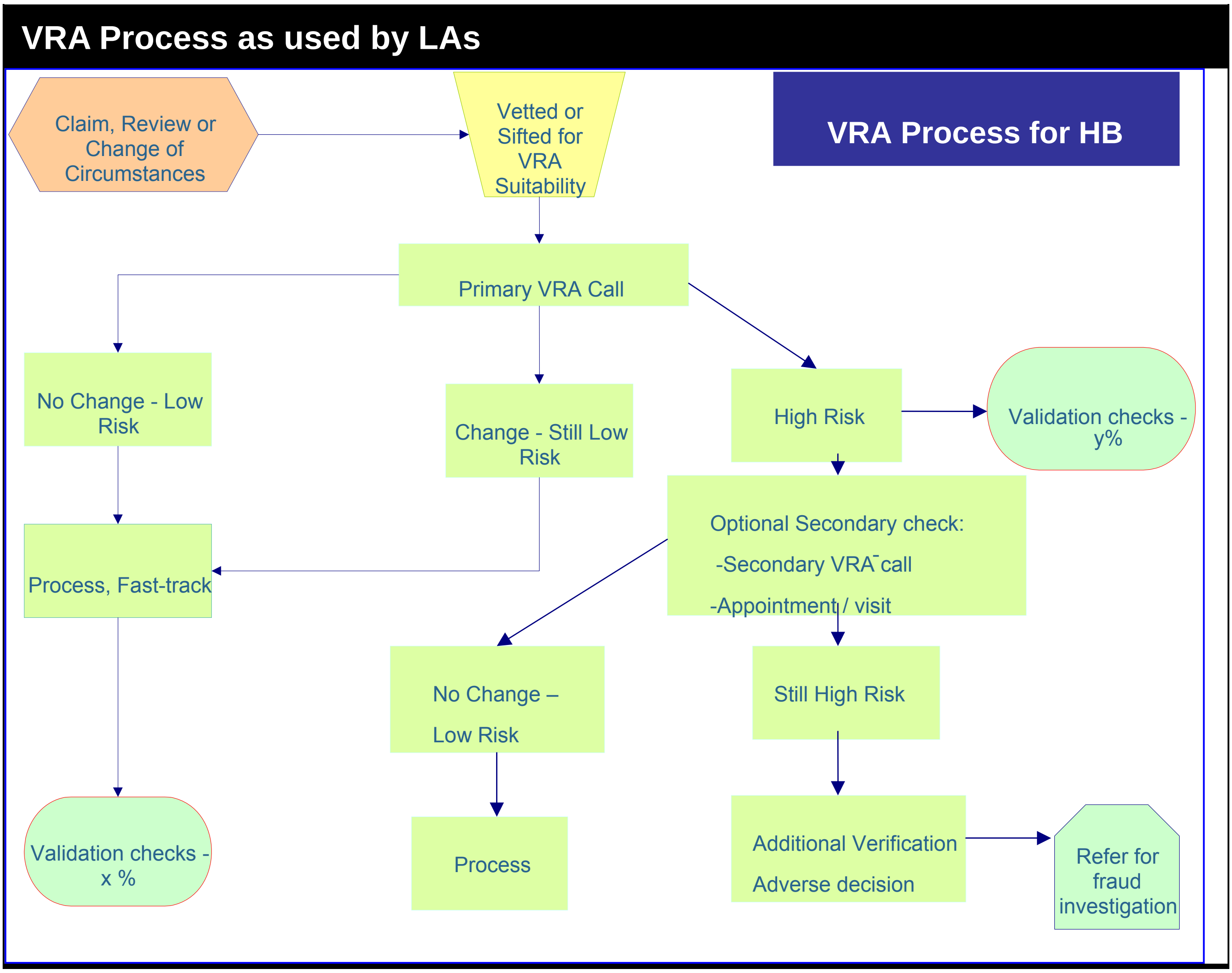
Telephony stage

Primary call: Where LAs conducted their VRA trial in-house a random selection of suitable claims were selected and forwarded to a trained VRA adviser regularly during their trial. An outbound call was made using agreed scripts, technology and behavioural analysis techniques to determine a risk score.

Secondary call: Where a high risk score was determined by the call handler the LA had an option to conduct a second telephone call a few days later in a free, ‘un-scripted’ format. An experienced member of staff investigated any areas of the original call that had been highlighted as potentially risky. This action may lead to a call that was determined as high risk in the primary call being changed to low risk as a result of the secondary call. Where the LA chose a managed service a random selection of customers were selected and forwarded to Capita's call centre for the calls to be made on their behalf.

Validation stage

LAs were required to fully validate[2] all high risk claims. The customer was required to provide written proof of all necessary information before benefit would be processed.



[2] In order to ensure the accuracy of information provided to assess a customer for benefit a validation is conducted. The customer would be required to provide written proof of information, for example; identity, finances and address, in order to assure the correctness of the benefit claim.

In order to confirm that the correct risk score was determined for those calls deemed to be of low risk a minimum of 5% were subject to full validation.

A selection of claim validations were conducted ‘blind’ where the officer conducted a full validation check of a claim without being aware it had been through the VRA process. This was to ensure against any bias from knowledge of the VRA risk score.

A key component of this approach is that both the low and high risk judgements are subjected to the same follow-up process to determine accuracy. The fact that both groups are treated equally eliminates the need for a control group of referrals that could be generated by operator judgement or a random number generator. The counter-factual for the risk indications from the process is not a separate group of referrals, but the rate of benefit change detected in the low risk group.

Social research

A social research company, IFF, have been engaged to assess the impact of VRA on customers and call handlers involved in the pilot. Their research determined:

Staff welcomed VRA and its role in enabling them to process higher volumes of claims more efficiently.

Were pleased to be able to fast-track straightforward claims and felt they had delivered a better service to the bulk of claimants as a result.

While they thought it played some role in identifying or deterring ineligible claimants, they saw its primary purpose as being a way of speeding up the process for the vast majority of claimants who had legitimate claims.

There was no evidence to suggest that VRA would put legitimate claimants off making a claim,

All audiences interviewed for this research stressed the need for traditional routes (such as a postal and face to face service) to remain open for those not wishing to or able to participate in a VRA call.

Project approach

As LAs are independent, autonomous bodies each individual pilot was responsible for conducting their testing of VRA to agreed standards supported by a memorandum of understanding. Digilog and Capita supplied the technology, training and scripting and provided the day to day project management of the pilot.

DWP provided overarching project co-ordination through the VRA Project Steering Group which included DWP’s Principal Scientific Adviser and analysts from Housing, Research Analysis Division and Benefit Performance Division as well Fraud and Error Strategy Division.

Organisation

The project co-ordination role meant the team could be kept small, consisting of a project steering group, project manager, analytical resource and DWP consultancy. Drawing together experienced staff from these areas, Digilog/Capita and LAs was particularly useful in terms of understanding this complex project.

The project conformed to DWP practice and PRINCE methodology with governance and management through the VRA Steering Group and DWP change lifecycle

Summary of major outcomes

A successful outcome was defined as:

Criteria 1: The percentage of high risk cases with overpayments should be significantly higher than the percentage of low risk cases with overpayments.

Criteria 2: The volume of fast-tracked low-risk cases should be acceptably low, or the value attached to them should be acceptably low

The Project Steering Group steered analysis to focus on criteria 1. If this were proven; further work would be required to fully establish the case for VRA.

Achievement of the objectives

Effectiveness has been determined using the area under the Receiver Operating Characteristic (ROC) curve which is a graphical plot of the sensitivity and specificity of the ability to discriminate between two potential outcomes.

The ROC curve is most frequently used when looking at the accuracy of medical diagnoses, although its application is appropriate to a variety of situations.

In deriving the curve, there are four possible outcomes which are of interest:

- True positives (in our case the prediction about finding errors in the high-risk cases are correct);
- True negatives (in our case the prediction that we wouldn’t find errors in the low-risk cases are correct);
- False positives (where we find errors in the low-risk cases);
- False negatives (where we don’t find errors in the high-risk cases).

Using these figures we are able to calculate two figures which are plotted against one another:

The True Positive Rate (TPR) is the number of true positives as a fraction of all instances of true positives and false negatives. (In our example it is the number of correctly predicted errors found in high-risk cases as a proportion of all errors actually found);

The False Positive Rate (FPR) is the number of false positives as a fraction of all instances of false positives and true negatives. (In our example it is the number of cases where no errors were found in high-risk cases as a proportion of all cases where no errors were actually found).

If the prediction and outcome are always the same the TPR is 1 and the FPR is 0, producing an area under the ROC curve of 1.00. This is the proportion of the total area contained by the curve.

A threshold of around 0.7 would normally be indicative of some significant predictive power. As the proportion of the area under the curve increases above 0.7 towards 1.0, an increasing significance can be attributed to the method of distinguishing between the two outcomes (for example, errors or no errors). A full description of the evaluation methodology is included at Appendix 1.

The results

Table 1: LA Reviews indicating whether or not any differences exist for the percentage of overpayments between high-risk and low-risk cases

	% overpaid at validation		Significant difference	AUC score	Confidence (+/-)	Overall
LA	High risk	Low risk				
Basildon	49%	4%	Yes	0.74	0.17	Strong Positive
Bromsgrove	83%	14%	Yes	0.63	0.13	Weak Positive
Northampton-shire	20%	2%	N/A	0.61	0.15	Weak Positive
Lichfield	81%	9%	Yes	0.61	0.04	Weak Positive
Harrow	50%	8%	Yes	0.61	0.08	Weak Positive
Flintshire	28%	2%	Yes	0.59	0.07	Weak Positive
Birmingham	48%	19%	Yes	0.57	0.04	Weak Positive
Aberdeen	54%	4%	Yes	0.57	0.18	No effect
Doncaster	31%	17%	No	0.58	0.26	No effect
Durham	29%	24%	No	0.52	0.04	No effect
Barking & Dagenham	51%	70%	No	0.48	0.03	No effect
Lambeth	5%	12%	No	0.46	0.03	No effect

Table 2: LA Change in Circumstance indicating whether or not any differences exist for the percentage of overpayments between high-risk and low-risk cases

	% overpaid at validation		Significant difference?	AUC score	Confidence (+/-)	Overall
LA	High risk	Low risk				
Aberdeen	75%	2%	Yes	0.76	0.08	Strong Positive
Bexley1	10%	1%	N/A	0.74	0.47	Weak Positive

Table 3: LA New claims indicating whether or not any differences exist for the percentage of overpayments between high-risk and low-risk cases

	% overpaid at validation		Significant difference?	AUC score	Confidence (+/-)	Overall
LA	High risk	Low risk				
Bromsgrove	97%	1%	Yes	0.98	0.02	Strong Positive
Walsall	34%	5%	Yes	0.72	0.03	Strong Positive
Warwick	36%	2%	Yes	0.72	0.08	Strong Positive
Flintshire	37%	4%	Yes	0.64	0.14	Weak Positive
Aberdeen	54%	3%	Yes	0.66	0.07	Weak Positive
Coventry	28%	3%	Yes	0.58	0.04	Weak Positive
Northamptonshire	7%	10%	No	0.49	0.06	No effect
Doncaster	2%	2%	No	0.51	0.25	No effect
Bristol	0%	25%	No	0.47	0.24	No effect

Limitations

The evaluation was conducted in a live environment and its principal objective was to assess the capacity of the process to identify risk in the environment of LA HB departments. It is not and was not the purpose of the evaluation to reach broad conclusions on the efficacy of the technology and process elsewhere.

VRA is used in the private sector to risk score telephone calls it is understood that the process is widely used in the insurance industry.

Summary of results

Twelve local authorities tested VRA on in-claim reviews with one strong positive and six weak positive results.

Two local authorities tested VRA on changes in circumstance with one strong positive and one weak positive result.

Nine local authorities tested VRA at new claims stage with three strong positive and three weak positive results.

Conclusion and next steps

The VRA process as tested by LAs for the DWP combined technology with intelligent questioning by specially trained call handlers supported by bespoke scripts and a process redesign. A two-stage trial was carried out between August 2008 and December 2009. The aim was to assess whether call handlers using the VRA process could correctly discriminate changes to customer's declared circumstances revealed by a full validation of information.

The evaluation was intended to determine whether VRA worked when applied to the benefit system. From our findings it is not possible to demonstrate that VRA works effectively and consistently in the benefits environment. The evidence is not compelling enough to recommend the use of VRA within DWP. At no stage did the evaluation of management information carried out by the Department explicitly consider the effectiveness of the technological aspects of VRA. However, social research evidence suggests that the scripting and training elements of the trial were successful.

The trial appeared to suggest it is operationally difficult to sufficiently monitor the success of VRA. Any risk based verification solution such as VRA would require on-going independent monitoring during live running to ensure the robustness of the process.

The failure to find compelling evidence of discrimination does not mean that the process could not work in other environments, as this is one trial conducted in a complex operational environment. The trial focussed on what an operational deployment would mean to DWP in terms of potential benefits and risks. The focus on operational deployment – an assessment of whether the VRA process 'worked' in LAs and therefore potentially have a role across DWP- meant that certain areas for exploration, such as the reliability of operators' judgements of risk, and the science behind the technology were not part of the scope of the trial. Therefore, this trial is more about whether it could work rather than how or why it works.

Trials of VRA began in 2007. Since then the department's Service Delivery Strategy has evolved to focus on pursuing benefit provision through automated service delivery provided by an on-line service rather than telephony based systems such as VRA.

No further trialling of VRA is planned and on the basis of this evaluation we cannot make any recommendations for its use within benefit processing.

Appendix 1

Voice Risk Analysis methodology – the Receiver Operating Characteristic curve

The area under the Receiver Operating Characteristic (ROC) curve is a graphical plot of the sensitivity and specificity of the ability to discriminate between two potential outcomes. The ROC curve is most frequently used when looking at the accuracy of medical diagnoses, although its application is appropriate to a variety of situations.

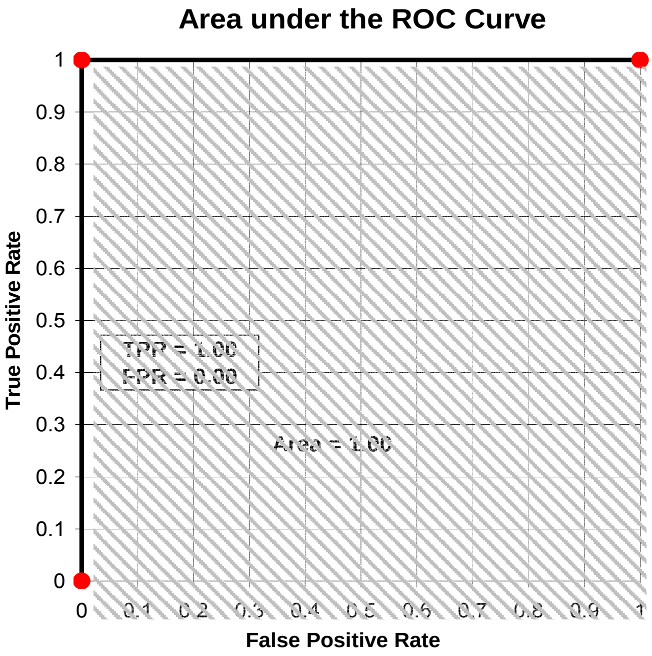
In deriving the curve, there are four possible outcomes which are of interest:

- True positives (in our case the prediction about finding errors in the high-risk cases are correct);
- True negatives (in our case the prediction that we wouldn't find errors in the low-risk cases are correct);
- False positives (where we find errors in the low-risk cases);
- False negatives (where we don't find errors in the high-risk cases).

Using these figures we are able to calculate two figures which are plotted against one another:

- The True Positive Rate (TPR) is the number of true positives as a fraction of all instances of true positives and false negatives. (In our example it is the number of correctly predicted errors found in high-risk cases as a proportion of all errors actually found);
- The False Positive Rate (FPR) is the number of false positives as a fraction of all instances of false positives and true negatives. (In our example it is the number of cases where no errors were found in high-risk cases as a proportion of all cases where no errors were actually found).

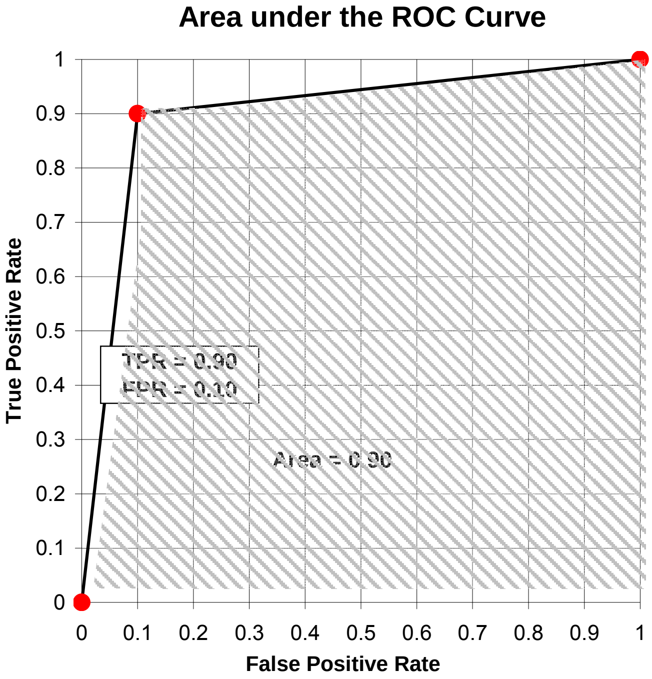
If the prediction and outcome are always the same the TPR is 1 and the FPR is 0, producing an area under the ROC curve of 1.00. This is the proportion of the total area contained by the curve. This is illustrated below:



This level of accuracy is rarely achieved and suggests a perfect ability to predict the actual outcome, with no instances of:

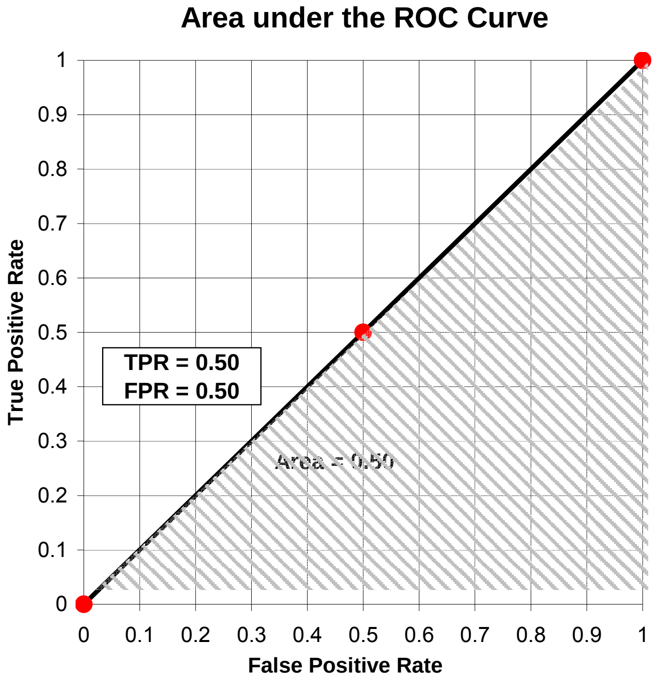
- False negatives (where errors are found amongst low-risk cases); or
- False positives (where no errors are found amongst cases classified as high-risk).

In many instances incorrect predictions will be made. This can result in the TPR being less than 1 and the FPR being greater than 0. This is illustrated below:



The proportion of the area enclosed by the curve has reduced from the previous example. The proportion of the area under the curve is now 0.90.

As the TPR reduces further and the FPR increases, the area enclosed by the curve reduces. At the point where the TPR and the FPR have the same value there is no predictive power in the method used to distinguish between different types of call. This produces a straight line where the proportion of the area under the curve is 0.50. In this case, the predictive power of the approach is no better than making the decision about whether a call is high-risk or low-risk purely at random.



If the proportion of the area under the curve becomes less than 0.50 (where the TPR becomes less than the FPR), this indicates that the classification of the call as high-risk or low-risk is worse than randomly designating the call as high-risk or low-risk.

The method of calculating the area under the ROC curve takes no account of the number of observations making up the two proportions (the TPR and the FPR). In order to get an indication of the reliability of the figures based upon the number of cases which were validated, Monte Carlo statistical techniques have been used to estimate the 95% Confidence Intervals for the TPR, the FPR and the resulting area under the curve.

A threshold of around 0.70 would normally be indicative of some significant predictive power. As the proportion of the area under the curve increases above 0.7 towards 1.0, an increasing significance can be attributed to the method of distinguishing between the two outcomes (for example, errors or no errors).The chi-square test

The chi-square test is used to test for differences in proportions for different groups of data.

In many cases differences can arise just by chance. The chi-square test is one way of trying to distinguish between differences that occur by chance and those which occur because there is a genuine difference between groups.

The chi-square test makes use of the 'Observed' values (that is, the actual figures) as well as the 'Expected' figures that would be seen if there was no difference between the different groups of data.

For example, there are two different types of claim – private tenants (Group A) and one in which there are council tenants (Group B). The claims are analysed to look at the number of errors made in the assessment of entitlement by benefit assessors. The 'Observed' results are as follows:

	Group A	Group B	TOTAL
Correct Entitlement	10	20	30
Errors	15	10	25
TOTAL	25	30	55

If there was no real difference between the two groups, then the following results would be the 'Expected' figures.

	Group A	Group B	TOTAL
Correct Entitlement	55 x (30÷55) x (25÷55) = 13.6	55 x (30÷55) x (30÷55) = 16.4	30
Errors	55 x (25÷55) x (25÷55) = 11.4	55 x (25÷55) x (30÷55) = 13.6	25
TOTAL	25	30	55

Note that the total of the rows and columns still sum to the totals seen in the initial table.

The next step would be look at the difference between the ‘Observed’ and the ‘Expected’ figures.

	Group A	Group B	TOTAL
Correct Entitlement	10 – 13.6 = -3.6	20 - 16.4 = 3.6	0
Errors	15 – 11.4 = 3.6	10 – 13.6 = -3.6	0
TOTAL	0	0	0

The differences are then squared and divided by the expected values.

	Group A	Group B	TOTAL
Correct Entitlement	(-3.6 x -3.6)÷13.6 = 1.0	(3.6 x 3.6)÷16.4 = 0.8	1.8
Errors	(3.6 x 3.6)÷11.4 = 1.2	(-3.6 x -3.6)÷13.6 = 1.0	2.1
TOTAL	2.1	1.8	3.9

The total of 3.9 is then compared with chi-square tables.

Chi-square tables show the minimum figure that the summed squared differences need to add to in order for any differences between the groups to be judged significant.

In this report, only 2 x 2 contingency tables have been used, but the method is appropriate for larger tables too.

For a 2 x 2 contingency table of the type shown above, and in the body of the report, a chi-square figure of at least 3.841 indicates that there is only a 5% chance of the differences observed between the two different groups occurring by chance. The figure of 3.9 which is arrived at above is just over this threshold, so we can be confident that there appear to be genuine differences in the proportion of assessor errors between the two groups of claims.

Different levels of confidence can be used when carrying out the chi-square test. For example, instead of using a 5% likelihood of the difference occurring by chance, we could use a 1% figure. If this were the case, the threshold for the chi-square statistic would increase from 3.841 to 6.635 – but at the same time we could be more confident about a genuine difference between the two groups.

So far we have only considered whether there are *differences* between the two groups. These differences could be in either direction (ie Group A may contain a higher proportion of errors than Group B, or Group B may contain a higher proportion of errors than Group A). Carrying out a chi-square test simply to discover if there are significant differences is sometimes referred to as a 'two-tailed' chi-square test.

In some situations there may be reasons to suspect that the difference will be observed in one particular direction. For example, we may suspect that a higher proportion of errors would be identified in the group containing private tenants (Group A) because of the additional potential for making errors associated with establishing rental liability etc. Carrying out a chi-square test in this situation is sometimes referred to as a 'one-tailed' chi-square test. Because the difference is believed to be in one specific direction, the threshold for the chi-square statistic increases.

In order to ensure that any chi-square test is legitimate, there are certain conditions that need to apply. One of these conditions is that all the 'expected' figures should be 5 or more. Using the test when there are fewer than 5 'expected' observations in any particular cell in the 2 x 2 contingency table, can produce spurious results.

Appendix 2

LAs which took part in VRA pilot	
LA	Delivery model
Aberdeen	in-house
Barking & Dagenham	managed service
Basildon	managed service
Bexley	managed service
Birmingham*	managed service
Bristol	in-house
Bromsgrove	in-house
Bury	in-house
Coventry	in-house
Derwentside Partnership* - now part of Durham	in-house
Doncaster	in-house
Eastbourne	managed service
Edinburgh*	in-house
Flintshire	in-house
Glasgow	in-house
Harrow*	in-house
Lambeth*	managed service
Lichfield	in-house
Northampton Benefit Partnership - (Corby, Kettering, Northampton & Wellingborough)	in-house
Swindon	managed service
Vale of Glamorgan	managed service
Walsall	in-house
Warwick*	in-house

Wealden*	**managed service**
Windsor & Maidenhead	**managed service**

- *LAs also took part in initial trial