Evaluation of UK NEQAS (H) Hb A_2 and related performance data

Hannah Batterbee[1], Barbara De la Salle[2], Paul McTaggart[2], Caroline Doré[3], Barbara Wild[4] and Keith Hyde[2]

[1]Haematology Department, Royal Hallamshire Hospital, Sheffield.
[2]UK National External Quality Assessment Scheme for General Haematology, Watford.
[3]MRC Clinical Trials Unit, London.
[4]UK NEQAS (H) Special Scientific Advisory Group Chair.

October 2010

Correspondence address:
Mrs B De la Salle
UK NEQAS (H), PO Box 14, Watford, WD18 0FJ
Telephone: 01923 217878
Email: haem@ukneqas.org.uk

UK NEQAS (H) Commentary on the Sheffield report

Scope

This overview encompasses the take-home messages that UK NEQAS (H) has identified from the evaluation of UK NEQAS (H) Hb A_2 data, undertaken by Mrs Hannah Batterbee from Sheffield Teaching Hospitals NHS Trust, on behalf of UK NEQAS (H), in the period July 2008 to December 2009. The report examines data from surveys of the UK NEQAS Abnormal Haemoglobins (AH) scheme during the period 2000 – 2008 for changes in methodology and evaluates participant Hb A_2 data submitted in the period 2006 – 2008.

Acknowledgement

The work was generously supported by the National Sickle Cell and Thalassaemia Screening Programme.

UK NEQAS (H) thanks Mrs Hannah Batterbee for her excellent work on this project and the Sheffield Teaching Hospitals NHS Trust for allowing her secondment to UK NEQAS (H).

Contact

In case of query, please contact Barbara De la Salle, UK NEQAS (H) Scheme Manager and Deputy Director, UK NEQAS (H), PO Box 14, Watford, WD18 0FJ.

Methodology used

During the period of the study, laboratories enrolled in the UK NEQAS (H) Abnormal Haemoglobins (AH) scheme largely changed from ion exchange column chromatography to high performance liquid chromatography (HPLC), with less than 10% of participants using column chromatography at the end of the study period (2008). Following completion of this report, a re-examination of current data has shown that just one UK clinical laboratory continues to use ion exchange column chromatography; this is a children's hospital that does not undertake antenatal screening. The most frequently used HPLC manufacturer is BioRad, followed by Tosoh, then Menarini.

Performance trends

Standard deviation (SD) and coefficient of variation (CV%) have decreased for both UK and non-UK participants indicating improvement in accuracy, however the improvement is best for UK laboratories. Since the study demonstrated a slight overall difference in performance improvement between UK and non-UK laboratories, performance analysis has focused on UK participants.

Instrument bias

Different instrument groups demonstrate bias compared to the all laboratories (or all methods) trimmed mean (ALTM). The ALTM is influenced in favour of the instrument group with the greatest numbers (in this study the BioRad group) and hence this observation should not be interpreted to indicate that any instrument is more correct than another.

In this study however the following general observations can be made concerning instrument bias:

1. Bias remains whether looking at all participants (UK and non-UK) or UK only.
2. Compared to the ALTM the Tosoh G7 instrument showed a strong positive bias; the BioRad D10, BioRad Variant Classic and BioRad V2 instruments showed a weaker positive bias.
3. Compared to the ALTM, the Menarini HA8160 instrument group shows a strong negative bias; the BioRad V2 using the Dual Kit reagent pack shows a weaker negative bias.

Reference ranges and cut off points

UK laboratories are still using widely differing reference ranges, even when using the same instrument. The source of reference ranges used is not clear.

Some UK laboratories are clearly not using the defined cut off Hb A_2 value of greater than or equal to 3.5% for the diagnosis of beta thalassaemia carrier status. This may reflect the fact that participants include laboratories from UK home countries that have not implemented the Screening Programme algorithm (Scotland, Wales, Northern Ireland), or are laboratories that do not undertake antenatal screening.

Performance assessment

There is little difference in performance assessment if this is undertaken against the ALTM (as currently used by UK NEQAS (H)) or the all laboratories (methods) median. However, if participants' results are examined by sub-method (i.e. the analyser group) a difference is seen: a greater proportion of participants show results outside the ± 2 standard deviation range and, when performance scores are calculated, more BioRad and Tosoh users have unsatisfactory performance scores.

Some proof of concept work has been done on altering the scoring algorithm. The current multiplier and truncation limits were originally chosen to give 5% of participants an unsatisfactory performance score of equal to or greater than 100: this may no longer be applicable in the light of the documented improved performance. The possibility of changing the multiplier will need to be tested using the UK NEQAS (H) scoring module and a larger amount of data, to assess the true impact on participant performance assessment. The use of alternative cut off values for performance assessment (e.g. 90 or 110 instead of 100) will also require fuller evaluation but this approach is generally undesirable as this would provide a further complication to the scoring system and the risk of confusion for participants who are enrolled in other UK NEQAS (H) schemes.

Out of consensus participant assessment of their Hb A_2 results is mostly due to the use of an 'aberrant' reference range. Transcription was the least most likely cause of an incorrect assessment. Instrument bias also has an effect.

Evaluation of interpretive comments – clinical significance

Between 2006 and 2008, 30 participants (37% in UK) failed to identify beta thalassaemia in the interpretive comment codes used for specimens with an Hb A_2 greater than 3.5%. The evaluation of free text comments where applicable is included in the UK NEQAS (H) supplementary report that accompanies each survey report; in general, approximately half of the UK laboratories that fail to identify beta thalassaemia in any form of comment in a borderline Hb A_2 specimen have returned a result below the 3.5% cut-off; the remainder have a result above the cut-off but have failed to return a correct interpretation. Although any screening programme will not be expected to identify 100% of affected individuals, especially at borderline levels, instrument bias will result in an unequal distribution of these 'misses' dependent upon the analyser used. Conversely, 42 participants (43% in UK) incorrectly identified beta thalassaemia in specimens where this did not exist: again there are issues around the generation of an incorrect Hb A_2 result or the interpretation of data from sickle cell carriers with possible co-existent alpha thalassaemia or iron deficiency. In this scenario, instrument bias against a fixed cut-off could result in over diagnosis and unnecessary partner testing, with the associated distress for the patient that this entails.

Manufacturers' feedback

This report has been reviewed by the equipment manufacturers prior to publication and comments from Tosoh Bioscience, Helena Biosciences, BioRad Laboratories and Menarini Diagnostics are included in full in appendices to the report.

Further work and discussion

UK NEQAS (H) has found this evaluation of data extremely helpful in confirming and summarising performance trends. Inevitably, this report has highlighted the need for additional work, which UK NEQAS (H) will submit to the NHS Sickle and Thalassaemia Screening Programme for consideration.

This report will provoke discussion between the developers of policy in haemoglobinopathy and thalassaemia screening and those responsible for service delivery, quality monitoring and equipment manufacture. This is welcome and should result in further improvement in the quality of laboratory services provided to patients.

Professor Keith Hyde (UK NEQAS (H) Scheme Director)
Mrs Barbara De la Salle (UK NEQAS (H) Scheme Manager)
July 2010

Evaluation of UK NEQAS (H)
Hb A_2 and related performance data

A report commissioned by the UK National External Quality Assessment Scheme for General Haematology from Sheffield Teaching Hospitals NHS Foundation Trust, on behalf of the National Sickle and Thalassaemia Screening Programme

Contents

Executive Summary 11
Introduction 13
Objectives 14
Methods 14
UK NEQAS Calculated Mean 14
Methodology 14
Method/Submethod Bias 15
Standard Deviation 15
Coefficient of Variation 15
Result Distribution 16
Normal Distribution 16
Cumulative Distribution 16
Borderline Hb A_2 Specimen 17
T-Tests 17
UK Interquartile SDs 17
Performance Scoring 18
Normal Ranges 19
Hb A_2 Assessment Codes 20
Interpretive Comment Codes 20
Results 21
Methodology 21
Method/Submethod Bias 23
Standard Deviation 26
Coefficient of Variation 27
Result Distribution 28
Normal Distribution 32
Cumulative Distribution 34
Borderline Hb A_2 Specimen 36
T-Tests 37
UK Interquartile SDs 38
Performance Scoring 40
Normal Ranges 42
Hb A_2 Assessment Codes 43
Interpretive Comment Codes 44
Discussion 46
Methodology 46
Method/Submethod Bias 46
Standard Deviation 47
Coefficient of Variation 48
Result Distribution 48
Normal Distribution 49
Cumulative Distribution 50
Borderline Hb A_2 Specimen 50
T-Tests 51
UK Interquartile SDs 52
Performance Scoring 52
Normal Ranges 54
Hb A_2 Assessment Codes 55

Interpretive Comment Codes....56
Initial Conclusions....58
Further Work....62
Acknowledgements....63
References....63
Appendices....65
Comments from the manufacturers....65
Appendix 1: Response from Tosoh Bioscience....67
Appendix 2: Response from Helena Biosciences....69
Appendix 3: Response from Bio-Rad Laboratories....71
Appendix 4: Response from Menarini Diagnostics....73

Executive Summary

The Hb A_2 project has been funded by the NHS Sickle Cell and Thalassaemia Screening Programme to look back at historic data for Hb A_2 measurement that has been gathered by UK NEQAS to look for trends in Hb A_2 quantitation and differences in interpretation between different method and analyser groups. The ultimate goal of the project is to assist with the development of more sensitive indicators for monitoring the performance of Hb A_2 measurement and to establish more evidence to evaluate the 3.5% Hb A_2 cut-off for beta thalassaemia carrier status.

The NHS Sickle Cell and Thalassaemia Screening Programme was established in 2001 and since then, there has been a change in methodology from the majority of UK NEQAS participants using column chromatography for Hb A_2 measurement to using HPLC. With this change, we have seen a concurrent reduction, and therefore improvement, in the SD and CV for Hb A_2 measurement by UK NEQAS participants which demonstrates greater consistency in results.

Investigation into method/submethod group bias and analysis of the number of participants submitting particular Hb A_2 results within specific groups has shown that the overall mean Hb A_2 value does not necessarily reflect the distribution of the data since some participant groups tend to show a positive or negative bias which may skew the overall mean. The consistent trends in Hb A_2 measurement shown by certain method and analyser groups were used to predict the pattern of results that would be submitted for a specimen with a very borderline Hb A_2 level. As predicted, the results of this specimen were consistent with previous findings in that when the level of Hb A_2 is borderline or raised, the Menarini HA8160 group has a clear negative bias and the Tosoh G7 group has a positive bias for Hb A_2 measurement.

The UK Hb A_2 data from 2006 to mid-2008 was studied to assess the differences in participants results compared to the all methods median, the method specific median and the submethod specific median. Little difference was found when comparing participant's results to the all methods median or the method specifc median, however the composition of participants submitting outlying Hb A_2 results when compared to the submethod specific median varied greatly. Similarly, the consequences of potential alterations to the performance scoring system were similar if UK NEQAS were to continue to assess participants against the method specific median or move to comparing them to the all methods median, however they were very different when a move to comparing participants results to the submethod specific median was applied.

Since it had already been established that certain methods/submethods appear to have a positive bias for Hb A_2 and others have a negative bias, it was predicted that different laboratories would have different normal reference ranges for Hb A_2. It was found, however, that even laboratories using the same method or analyser for Hb A_2 are using a wide variety of different reference ranges. It was thought that this may account for some of the differences in interpretation of the Hb A_2 results generated by participants.

When the Hb A_2 results generated are assessed by participants in terms of being low, normal, high or uncertain, the primary reasons for the submission of an 'outwith consensus' Hb A_2 interpretation appear to be due to generation of an 'outwith consensus' Hb A_2 value, application of varying normal ranges between participants or a combination of these two reasons. When the 'outwith consensus' comment codes submitted by participants between 2006 and mid-2008 were studied, the primary cause of an incorrect assessment of beta thalassaemia carrier status appeared to be the generation of an 'outwith consensus' Hb A_2 value.

There are therefore clear differences in the Hb A_2 values generated by different HPLC analysers. It has been shown that these differences are likely to result in differences in the interpretation of the Hb A_2 result which may lead to a clinically incorrect assessment of beta thalassaemia carrier status.

The primary aim of UK NEQAS (H) is to maintain and improve performance of laboratory haematology at a high level of proficiency. External quality assurance provides an objective, long term, retrospective assessment of laboratory performance and allows the improvement of performance through education. Individual laboratory performance in all UK NEQAS (H) surveys is assessed against the consensus result for the survey and a scoring system is used to help achieve this (UK NEQAS (H), 2008). As for most parameters, the Hb A_2 scoring system was devised empirically through experimentation, in conjunction with the UK NEQAS (H) Steering Committee and the National Quality Assurance Advisory Panel (NQAAP) for Haematology, and is under constant review. This project forms part of this review process.

There are a number of additional elements of the historic Hb A_2 data that could be investigated as well as specifically designing future surveys to gain more information and evidence to support the performance monitoring of Hb A_2 measurement. Before UK NEQAS make any changes to the Hb A_2 performance scoring system, however, more detailed statistical analysis of the existing data must be performed. Additional resources are required in order to perform this further data analysis.

Introduction

The NHS Sickle Cell and Thalassaemia Screening Programme was set up in 2001 to deliver the world's first linked newborn and antenatal screening programme (NHS Sickle Cell and Thalassaemia Screening Programme, 2006; http://sct.screening.nhs.uk/aboutus). The aims of the programme are to:

- Save lives through prompt identification of affected babies.
- Offer informed choice to couples expecting a baby.
- Support the development of a managed clinical care network such that people have fair access to quality services throughout England.
- Raise public awareness of the disorders and challenge stigma.

The main aim of the antenatal part of the programme is to offer sickle cell and thalassaemia screening to all eligible women and couples in a timely manner during pregnancy (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). One of the conditions that the programme is designed to screen for in pregnant mothers is the beta thalassaemia carrier state (or beta thalassaemia minor or trait) (NHS Sickle Cell and Thalassaemia Screening Programme, 2006). This is an asymptomatic condition characterised by a microcytic, hypochromic blood picture with a high red cell count and mild anaemia. The diagnosis is confirmed by a raised haemoglobin A_2 (Hb A_2) level and partner testing would be offered to establish what the potential outcomes are in the foetus. If both parents are found to be carriers of beta thalassaemia then there is a 25% chance that the child will have beta thalassaemia major (Hoffbrand *et al*., 2002). If one parent is a beta thalassaemia carrier and the other is a carrier of Hb S, $\delta\beta$-thalassaemia, Hb Lepore or Hb E then there is also a risk that the baby will inherit a serious haemoglobinopathy disorder (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a).

No further confirmatory testing for beta thalassaemia is necessary if the condition shows the typical red cell picture and the Hb A_2 is raised. If the Hb A_2 result is normal but the patient has microcytic red cell indices then the full blood count results may be attributed to iron deficiency and partner testing may or may not be offered depending on the ethnicity of the patient and the relative risk of alpha thalassaemia. It is therefore crucial that laboratories produce accurate and reliable Hb A_2 results in order to correctly diagnose beta thalassaemia carrier individuals and to prevent the misdiagnosis of other causes of red cell microcytosis. In the Handbook for Laboratories released by the NHS Sickle Cell and Thalassaemia Screening Programme, the cut-off for a normal Hb A_2 level was less than or equal to 3.5% at the time that this research was conducted. The testing algorithm employed at the time stated that if the Hb A_2 level is greater than 4.0% then there is a risk of beta thalassaemia and at 3.6 to 4%, there is a risk of beta thalassaemia if the MCH is less than 27pg (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). The testing algorithm was later altered in 2009 so that the cut-off for a normal Hb A_2 level was less than or equal to 3.4% and at 3.5 to 4%, there is a risk of beta thalassaemia if the MCH is less than 27pg (NHS Sickle Cell and Thalassaemia Screening Programme, 2009).

Objectives

The primary objectives of the Hb A_2 evaluation project are as follows:

a. To analyse UK NEQAS participant data (current and retrospective) from the Abnormal Haemoglobins scheme to identify performance trends for the measurement of Hb A_2.

b. To gather additional information from participants on work practice to inform the analysis of performance.

c. To undertake preliminary examination of the performance indicators used in the performance surveillance of laboratories providing screening and diagnostic services for haemoglobinopathy and thalassaemia disorders.

The research should allow UK NEQAS and the National Screening Programme Centre to use potentially more sensitive indicators to monitor the performance of individual laboratories and manufacturers, especially at the cut-off points for clinical decision making established in the National Screening Programme algorithm.

Methods

UK NEQAS Calculated Mean

In a number of the investigations described in this report, the UK NEQAS 'all methods mean' or 'HPLC mean' is either used or mentioned. This is the all methods or HPLC mean calculated and reported by UK NEQAS for each survey. It is generated by calculating the log of the Hb A_2 results submitted by all participants, or all participants using HPLC methods, (both UK and non-UK based) and then trimming off 10% of the results so that the 5% highest and 5% lowest results submitted are excluded. The mean of this data set is then calculated and then back transformed to give the all methods geometric mean or HPLC geometric mean.

Methodology

One of the initial aims of the project was to investigate changes in the numbers of different methods and analysers used by participants to measure Hb A_2 over the last eight years. This was achieved by looking back at the methods for which participants were registered for the Abnormal Haemoglobins scheme from the beginning of 2000 (survey 0001AH) to mid-2008 (survey 0803AH). The overall change in the methods used by participants was studied and also the changes in the methods used by UK and non-UK participants were investigated separately. The HPLC method group was then broken down for the same sets of data into specific analyser groups so that the changes in the submethods used by participants over the last eight years could also be studied.

Method/Submethod Bias

To demonstrate which methods/analysers show a clear positive or negative bias for Hb A_2 measurement, the method/submethod mean minus the UK NEQAS all methods mean was calculated for all surveys from 2006 to mid-2008. The eight most widely used methods/submethods were focussed on, i.e. those with an average number of participants during this period of greater than 10, and for this exercise, both UK and non-UK participants results were used in order to obtain an adequate number of participants (greater than 10) using each method/submethod.

In addition to comparing the calculated method/submethod mean to the UK NEQAS all methods mean, the method/submethod median for Hb A_2 was also compared to the UK NEQAS all methods mean. Also the UK and the non-UK participants were separated for this exercise.

Standard Deviation

Standard deviation (SD) is usually calculated using Equation 1 below:

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}} \qquad (1)$$

where σ is the standard deviation and $\bar{x}$ is the mean of the data set.

For this project, however, the SD values were calculated using the median of the data sets and the interquartile range. This was achieved by calculating the upper quartile and lower quartile of the data set, which can be accomplished by listing the results in ascending numerical order and selecting the value at the three-quarter (upper quartile) and one-quarter (lower quartile) positions. By subtracting the lower quartile from the upper quartile, the interquartile range is calculated and from this, a robust estimate of SD can be obtained by dividing the interquartile range by 1.349. This will be referred to as the interquartile SD.

The overall change in the interquartile SD of the Hb A_2 results submitted by UK plus non-UK participants was calculated for all surveys from 2000 to mid-2008. The results of the UK and the non-UK participants were then separated to find out if there was any difference between the interquartile SDs, and the change in interquartile SDs over the last eight years, of the results submitted by these two groups.

Coefficient of Variation

The coefficient of variation (CV) is typically calculated by expressing the SD as a percentage of the mean of the data set. Along with calculating the SD, the overall change in CV of the Hb A_2 results submitted by UK plus non-UK participants from 2000 to mid-2008 was calculated. For the purposes of this project, the CVs were calculated by dividing the interquartile SD by the median of the data set. Again, the UK and non-UK participants results were then also separated out.

Result Distribution

Result distribution graphs for Hb A_2 were generated for all surveys from 2006 to mid-2008. These graphs show the proportion of participants (UK plus non-UK) using six of the most common methods/submethods that submitted a particular Hb A_2 result for each survey specimen. They also show the overall proportion of participants that submitted a particular Hb A_2 result, which is represented by the 'All methods' line on the graphs. Only Hb A_2 values within ±3 SDs of the UK NEQAS all methods mean were included on the graphs. Both UK and non-UK participant's results were used in this analysis in order to obtain a larger number of results from participants using the six chosen methods/submethods.

All of the result distribution graphs were later redrawn but on this occasion, the results of five of the largest HPLC groups were simply compared to the overall HPLC results submitted. These HPLC result distribution graphs therefore show the proportion of participants (UK plus non-UK) using five of the most popular HPLC analysers that submitted a particular Hb A_2 result for each survey specimen from 2006 to mid-2008. They also show the proportion of participants using all HPLC submethods that submitted a particular Hb A_2 result. On this occasion, only Hb A_2 values within ±3 interquartile SDs of the overall HPLC median were included on the graphs. It was felt that using median and interquartile SD would be a more robust way of excluding outlying results from the graphs than using the UK NEQAS HPLC mean and calculating the SD on the untrimmed data.

Normal Distribution

It was proposed that by generating normal distribution curves of the data whereby the results of the different data sets would be shown as smooth curves, the overall trends within the data would be seen more clearly. Normal distribution curves were therefore generated by plotting the median of the data sets ±3 SDs based on median and interquartile range. As with the result distribution curves, the results of all methods, plus those of six of the most commonly used methods/submethods, were plotted for all surveys from 2006 to mid-2008. Again, both UK and non-UK participant's results were used in this analysis in order to obtain a larger number of results from participants using the six chosen methods/submethods.

Cumulative Distribution

Cumulative distribution histograms were plotted for the Hb A_2 results of the four specimens with varying Hb A_2 levels used as examples in the Normal Distribution part of the Results section. This was achieved by plotting the cumulative distribution using the normal distribution function in Excel. The results of all methods, plus those of six of the most commonly used methods/submethods, were plotted for these specimens. Again, both UK and non-UK participant's results were used in this analysis to obtain a larger number of results from participants using the six chosen methods/submethods and for continuity.

Borderline Hb A_2 Specimen

To test certain predictions regarding the pattern of results submitted by participants for specimens with a borderline Hb A_2 value, a specimen with a borderline level of Hb A_2 was generated by mixing the blood of a beta thalassaemia carrier donor with that of a donor with a normal Hb A_2 level. This specimen (0902AH1) was sent out in 2009 and the UK NEQAS all methods mean Hb A_2 was calculated to be 3.7%. Normal distribution curves and a cumulative distribution histogram were generated using the results of this specimen as described above.

T-Tests

In order to test the null hypothesis that there is no difference between the results submitted by UK and non-UK participants using the same method/submethod, heteroscedastic, unpaired t-tests were carried out. UK and non-UK participants results were compared from column chromatography, BioRad Variant Classic, BioRad Variant II; Beta thal, Menarini HA8160 and Tosoh G7 users for each survey between 2006 and mid-2008.

UK Interquartile SDs

The standard deviation (SD) of the various participant groups (UK only) was calculated using the median and interquartile range of the data sets for each survey from 2006 to mid-2008. For each specimen, participant's results were analysed in three different ways. They were compared to:

1) The overall median comprising all of the different methods.
2) The median of their method group where greater than or equal to 20 participants were using that method, which meant that HPLC users results were compared to the HPLC group median and for surveys 0601AH through to 0703AH, column chromatography users results were compared to the column chromatography group median. The results of participants using all other methods, e.g. electrophoresis, were compared to the all methods median.
3) The median of their submethod group where greater than or equal to 20 participants are using that particular analyser, which meant that BioRad Variant II; Beta-thal, Menarini HA8160 and Tosoh G7 users results were compared to the BioRad Variant II; Beta-thal, Menarini HA8160 and Tosoh G7 group medians respectively. The results of participants using all other HPLC analysers, e.g. BioRad Variant Classic, were compared to the HPLC method median.

Those participants whose results fell outside of a ±2SD limit (based on the appropriate median and interquartile range) using each of these three different approaches were then compared.

Performance Scoring

For assessing the performance of Hb A_2 measurement, UK NEQAS calculate each participant's analytical performance score by initially calculating their Deviation Index (*DI*) using Equation 2 below:

$$DI = \frac{R - M}{SD} \qquad (2)$$

In the current system employed by UK NEQAS:
R = laboratory result
M = trimmed mean
SD = standard deviation calculated using the trimmed mean

For the purposes of this study, the *DI* was calculated slightly differently in that:
R = laboratory result
M = median
SD = standard deviation calculated using median and interquartile range

The absolute value of the deviation index is taken (ignoring the sign) and any values greater than 3.5 are rounded down to 3.5. This is to avoid a very high DI value, e.g. due to a transcription error, having an excessive effect on the calculation. To calculate the participants Hb A_2 performance score, the deviation indices for the last 6 specimens are added together and the total is multiplied by 9. This figure was established empirically by experimentation in around 2004 in order to obtain approximately 5% of participants as having Persistent Unsatisfactory Performance. A participant is currently defined as having Persistent Unsatisfactory Performance if their score is greater than or equal to 100.

Using a system based on the current UK NEQAS scoring system of comparing participants results to the method specific median (where there are greater than or equal to 20 participants using that particular method), adding up the deviation indices of the last 6 specimens and multiplying by 9, a total of 9 separate scores were calculated for each UK participant using the results of surveys 0601AH through to 0803AH. This was then repeated using the all methods median and submethod specific median (where there are greater than or equal to 20 participants using that particular submethod) in the calculation and the number of participants that qualify as having Persistent Unsatisfactory Performance using each of these three different approaches was assessed.

To experiment with altering the definition of Persistent Unsatisfactory Performance, the performance score cut-off that would identify 5% of participants as Persistent Unsatisfactory Performers was calculated. This was achieved by arranging participants in order of ascending performance score and the score of the target participant 95% of the way up the list was noted as the target score. This was carried out for the 9 scores generated on the specimens from 2006 to mid-2008. The target scores were then averaged and rounded up or down to create a new number that could be used to obtain 5% of participants as Persistent Unsatisfactory Performers. This was carried out using the all

methods median, method specific median and submethod specific median separately within the performance scoring equation.

A more standard way of achieving 5% of participants as Persistent Unsatisfactory Performers was experimented with by altering the figure by which the sum of the previous 6 deviation indices is multiplied by. To mathematically calculate what this figure should be, the participants were listed in ascending order of the sum of their previous 6 deviation indices (Σ6DI). To work out where the cut-off should be in order to obtain 5% of participants as Persistent Unsatisfactory Performers (i.e. with scores greater than or equal to 100), the number of participants was multiplied by 0.95 so that, for example, if there were 160 participants, the cut-off should be at the 152nd in the list. To calculate the multiplier required to generate a performance score of 100 for this target participant (e.g. 152nd), Equation 3 below was used:

$$\frac{100}{\sum 6DI} \qquad (3)$$

This was carried out for the 9 scores generated on the specimens from 2006 to mid-2008 and then averaged to the nearest whole number. Again, this was carried out using the all methods median, method specific median and submethod specific median separately within the performance scoring system. In the method and submethod comparisons, column chromatography was no longer classed as a separate method group from survey 0704AH onwards as the group then comprised of less than 20 participants in the UK. Column chromatography results were therefore compared to the all methods median from survey 0704AH onwards.

The multiplier required to give equal proportions of persistent unsatisfactory performers within each method/submethod group was also calculated separately using the all methods median, method specific median and submethod specific median. This was achieved by using the same method as described above whereby the target score of the participant 95% of the way down the ascending Σ6DI list was identified and then the required multiplier calculated using Equation 3. For the BioRad Variant II; Beta-thal, Menarini HA8160 and Tosoh G7 groups, this was carried out for the 9 scores generated on the specimens from 2006 to mid-2008. For the Column Chromatography group, this was only performed on the 5 scores generated from specimens 0601AH1 to 0703AH4 as from survey 0704AH onwards, this group comprised of less than 20 participants in the UK.

Normal Ranges

Early in 2009, an additional sheet was sent out with survey 0902AH which asked participants to state their method and analyser for Hb A_2 measurement along with their normal reference range. Altogether, 172 UK and 85 non-UK participants returned this information. Of the remaining 19 UK and 21 non-UK participants, normal range information of 13 UK and 7 non-UK laboratories was able to be obtained as they had given this information on a recent survey return sheet in 2008. No normal range information was able to be obtained for 4 UK and

10 non-UK participants but the 4 UK laboratories were all found to be those of commercial manufacturers and therefore they do not interpret patient's results using a reference range. The minimum and maximum levels of the normal ranges obtained from UK and non-UK participants were plotted along with the midpoint of the range.

Hb A_2 Assessment Codes

For each specimen where a Hb A_2 result is requested, UK NEQAS participants are asked for an assessment of the Hb A_2 result that they have generated which involves ticking one of four boxes labelled low, normal, high or uncertain (UK NEQAS (H), 2008). The results given by UK plus non-UK participants for the Hb A_2 assessment were studied and were classed as 'consensus results' if greater than 85% of participants gave that answer. Where results were 'outwith consensus' but specimen quality was given as unsatisfactory, these results were not included. The Hb A_2 assessment codes for 62 different specimens between 2000 and 2008 were studied and the 20 participants with greater than or equal to 3 'outwith consensus' assessment results in the 23 specimens between 2006 and mid-2008 were investigated further. The Hb A_2 results that these 20 participants submitted were studied along with their normal ranges. This allowed the potential cause of the 'outwith consensus' assessment code to be hypothesised. The causes of these 'outwith consensus' assessments were divided into:

- generation of an 'outwith consensus' Hb A_2 result by the laboratory.
- the use of a normal range for Hb A_2 which is different from the majority of other participants.
- a transcription error in which the participant simply ticked the wrong box by mistake or an interpretation error in which the individual who filled out the UK NEQAS return form incorrectly interpreted the result.
- a combination of two or more of the above three causes.

The methodology employed for Hb A_2 measurement by these 20 participants with at least 3 'outwith consensus' assessment results between 2006 and mid-2008 was also investigated.

Interpretive Comment Codes

The interpretive comment codes that participants attach to their specimen results were studied for all specimens sent out by UK NEQAS from 2006 to mid-2008 with the exception of survey number 0604AH, as the required data was unavailable for this survey. After listing all of the 'outwith consensus' comment codes submitted by participants for these surveys, the comments that are relevant to beta thalassaemia were investigated further. Participants were identified that had used comments which state 'no evidence of thalassaemia' or 'no evidence of beta thalassaemia' for specimens which were from beta thalassaemia carrier individuals. Also, instances were participants had used comments inappropriately suggesting that beta thalassaemia was present were also identified. As with the Hb A_2 assessment codes, the 'outwith consensus' comment codes submitted by participants between 2006 and mid-2008 were studied in more detail, along with their normal ranges, in order to hypothesise what may have caused the incorrect comment code submission.

Results

Methodology

Figure 1 and Figure 2 below show how the number of UK and non-UK participants using different methods for Hb A_2 measurement has changed from 2000 to mid-2008.

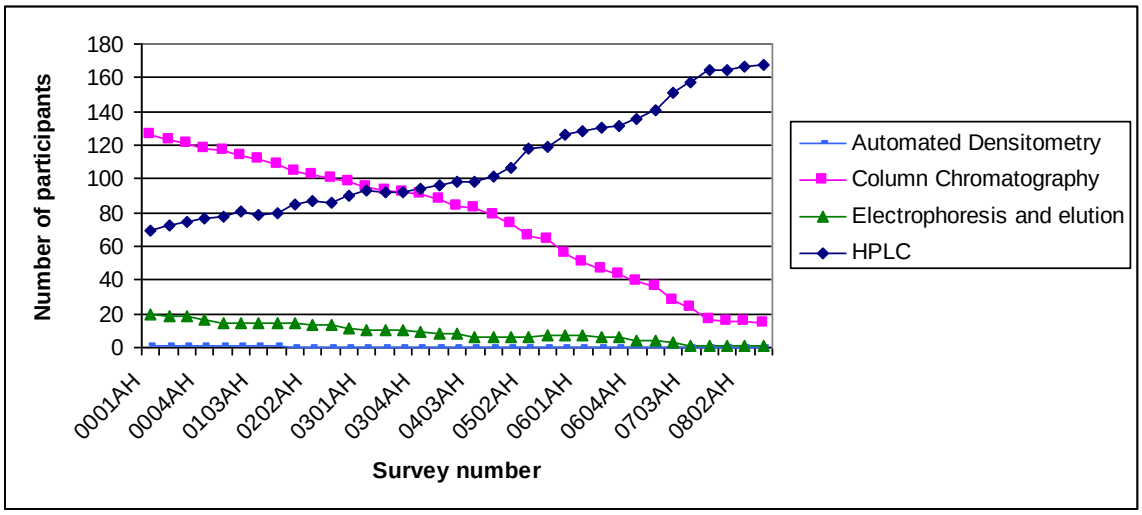

Figure 1: Change in the methods used by UK participants to measure Hb A_2 from 2000 to 2008.

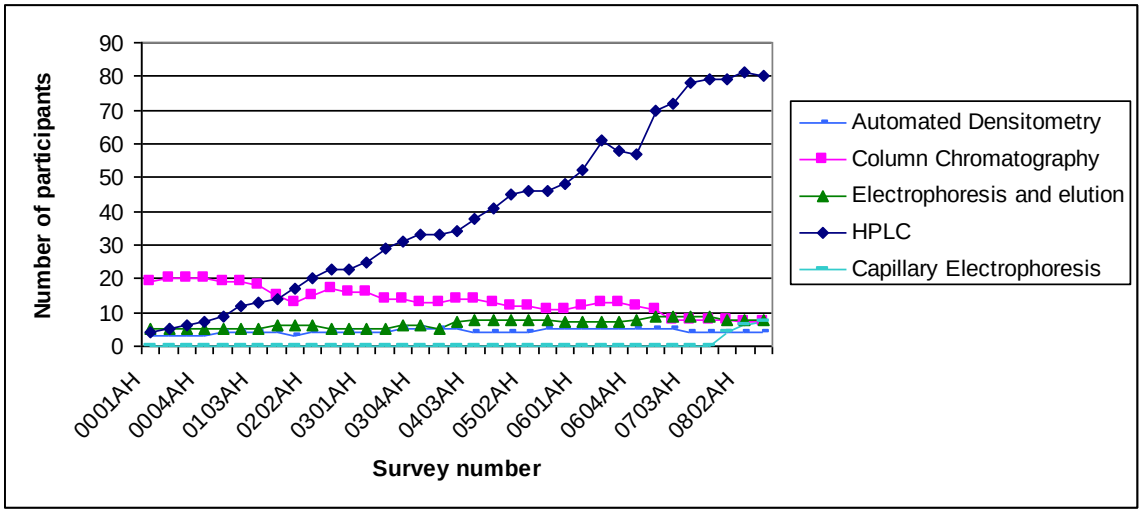

Figure 2: Change in the methods used by non-UK participants to measure Hb A_2 from 2000 to 2008.

Figure 1 shows that at the beginning of 2000, 58% of participants of the Abnormal Haemoglobins scheme were using column chromatography for Hb A_2 measurement and only 32% were using HPLC. By mid-2008, however, just 8% of participants were using column chromatography for Hb A_2 measurement and 92% were using HPLC. Figure 2 shows that there has been a reduction in the number of non-UK labs using column chromatography since 2000 but the main driving force behind the change in methodology of non-UK participants has been due to new labs joining the scheme which mostly use HPLC.

The HPLC users were then broken down into submethod, or analyser, groups as shown in Figure 3 and Figure 4 below. Note that the Misc. HPLC group comprises of all labs that UK NEQAS knew are using HPLC but where they did not know which analyser they were using.

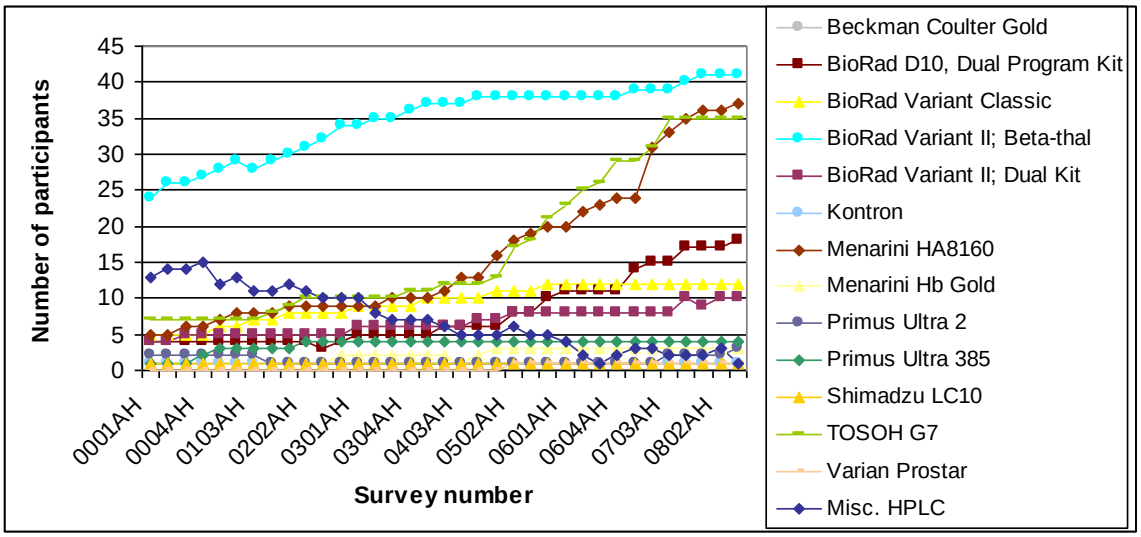

Figure 3: Change in the HPLC analysers used by UK participants to measure Hb A_2 from 2000 to 2008.

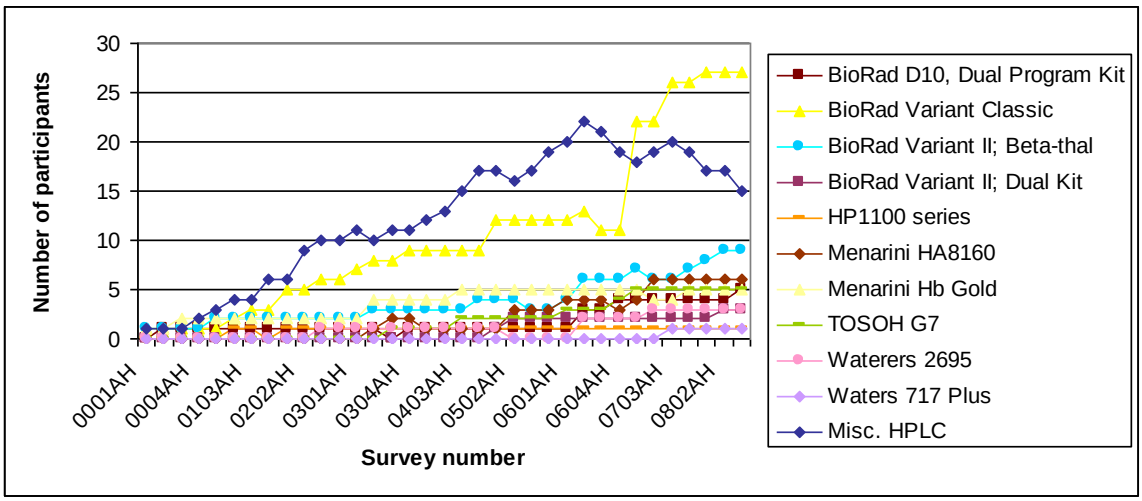

Figure 4: Change in the HPLC analysers used by non-UK participants to measure Hb A_2 from 2000 to 2008.

Figure 3 shows since the use of HPLC for Hb A_2 measurement has become more popular, the market leaders in the UK have been the BioRad Variant II; Beta-thal, the Menarini HA8160 and the Tosoh G7, followed by the other BioRad analysers. Figure 4 shows that amongst non-UK participants, the BioRad Variant Classic is by far the most popular HPLC analyser.

Method/Submethod Bias

Figure 5 shows the results of the calculated method/submethod mean minus the UK NEQAS all methods mean for the eight most widely used methods/submethods. It was generated using the data from both UK and non-UK participants.

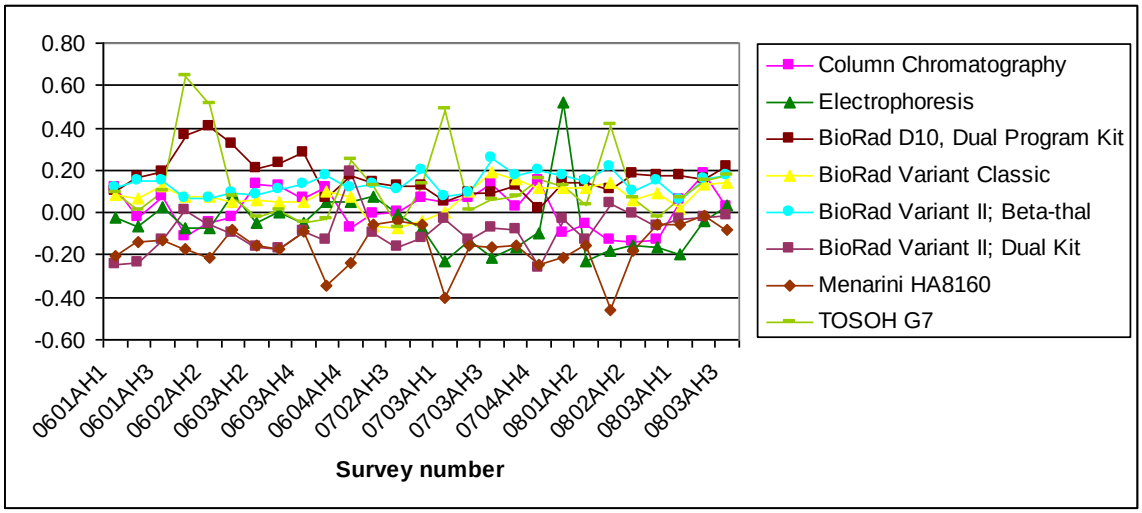

Figure 5: Method/submethod mean minus the all methods mean for the eight most widely used methods/submethods from 2006 to 2008.

Figure 5 shows that both the column chromatography and the electrophoresis method means are generally fairly close to the all methods mean and that neither method shows a clear positive or negative bias relative to the all methods mean. Of the HPLC analysers, the graph shows that the submethod means of the BioRad D10 Dual Program Kit, BioRad Variant Classic, BioRad Variant II; Beta-thal and the Tosoh G7 all tended to show a positive bias relative to the all methods mean, whereas the BioRad Variant II; Dual Kit and the Menarini HA8160 tended to show a negative bias.

In addition to comparing the calculated method/submethod mean to the UK NEQAS all methods mean, the method/submethod median for Hb A_2 was also compared to the UK NEQAS all methods mean for six of the most widely used methods/submethods. To generate Figure 6 through to Figure 11, the UK and the non-UK participants were separated and the calculated group mean and median compared to the UK NEQAS all methods mean for six of the most widely used methods/submethods.

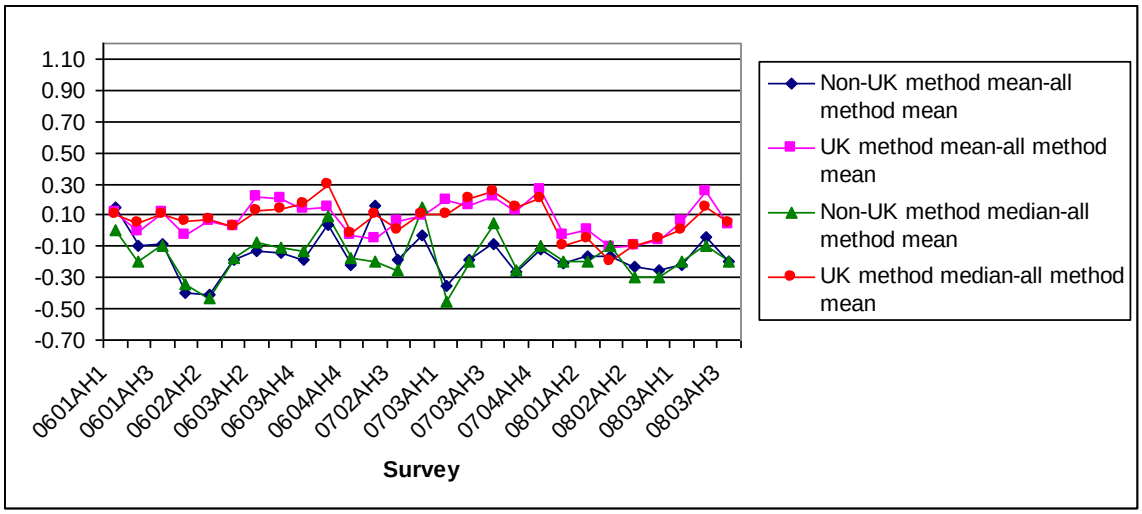

Figure 6: How the mean and median of the UK and non-UK column chromatography groups results compare to the all methods mean from 2006 to 2008.

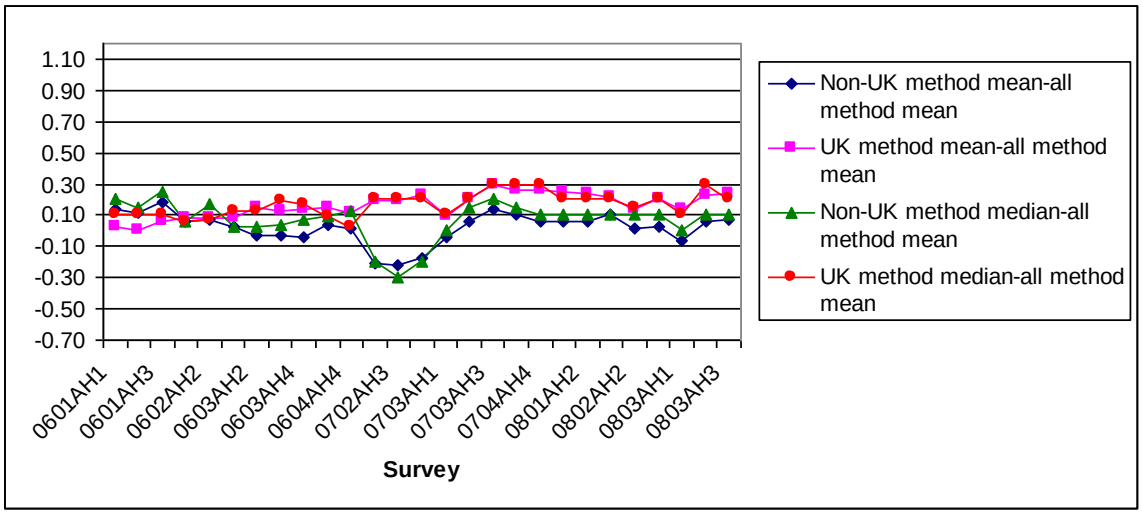

Figure 7: How the mean and median of the UK and non-UK BioRad Variant Classic groups results compare to the all methods mean from 2006 to 2008.

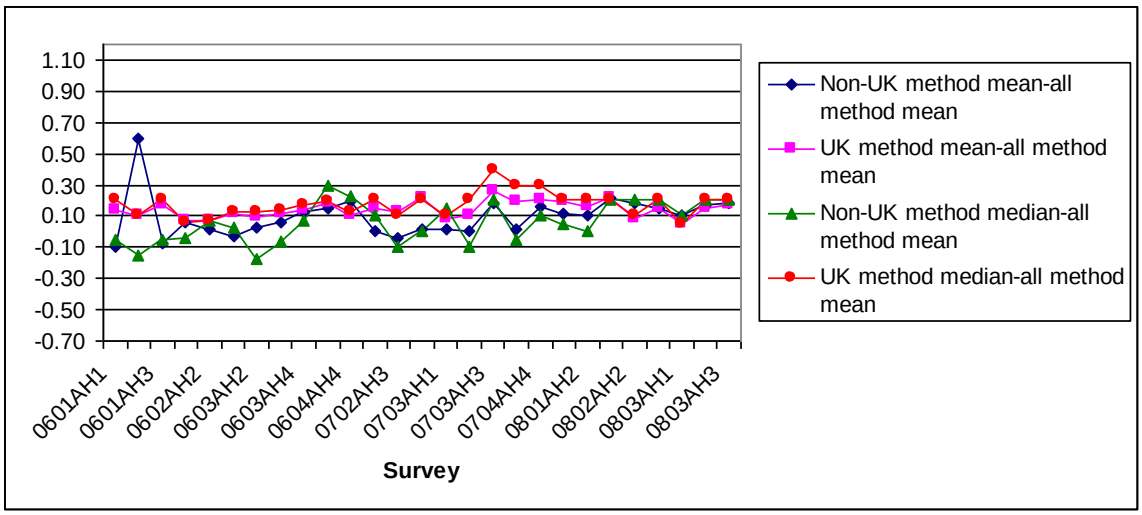

Figure 8: How the mean and median of the UK and non-UK BioRad Variant II; Beta Thal groups results compare to the all methods mean from 2006 to 2008.

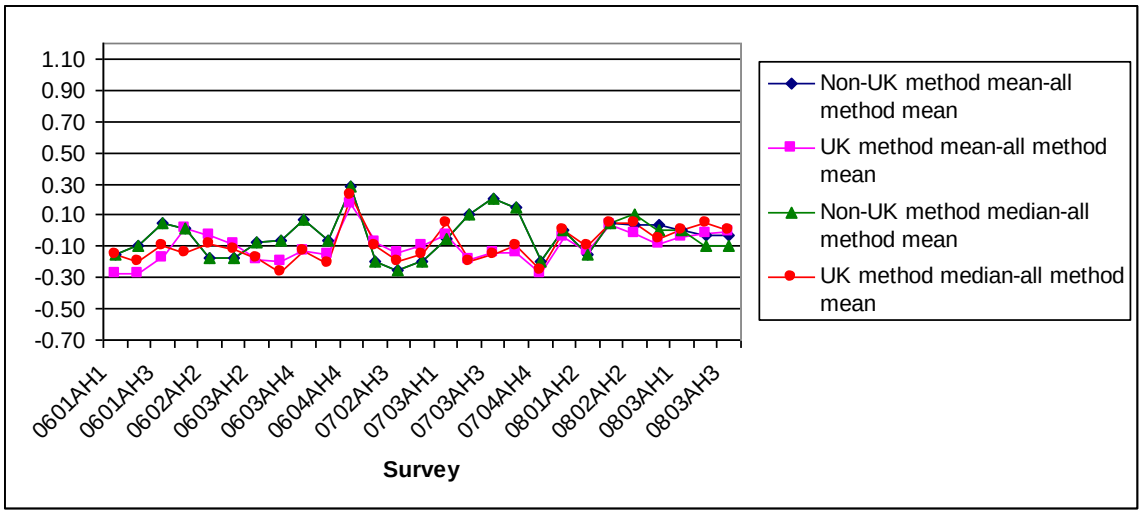

Figure 9: How the mean and median of the UK and non-UK BioRad Variant II; Dual Kit groups results compare to the all methods mean from 2006 to 2008.

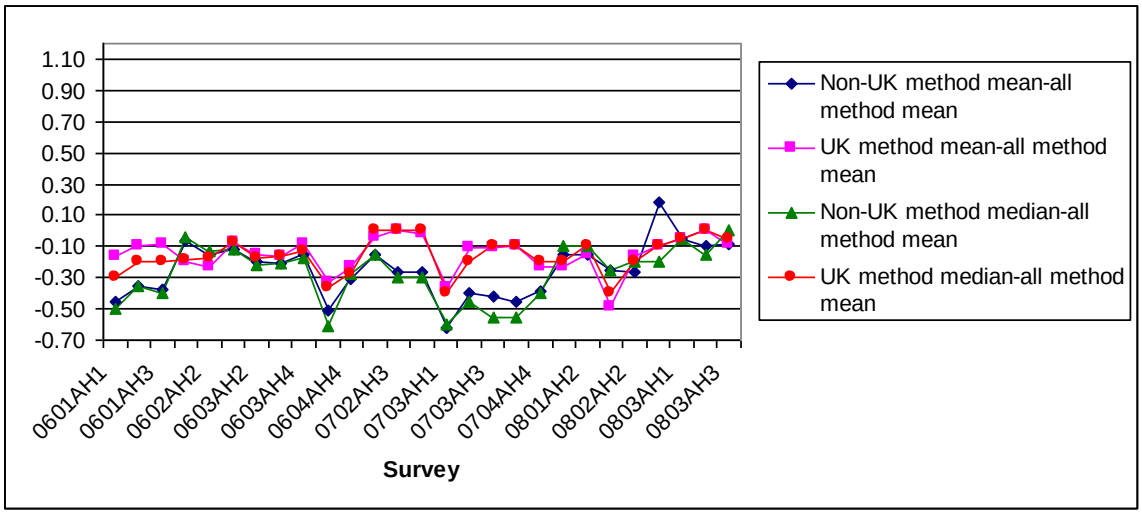

Figure 10: How the mean and median of the UK and non-UK Menarini HA8160 groups results compare to the all methods mean from 2006 to 2008.

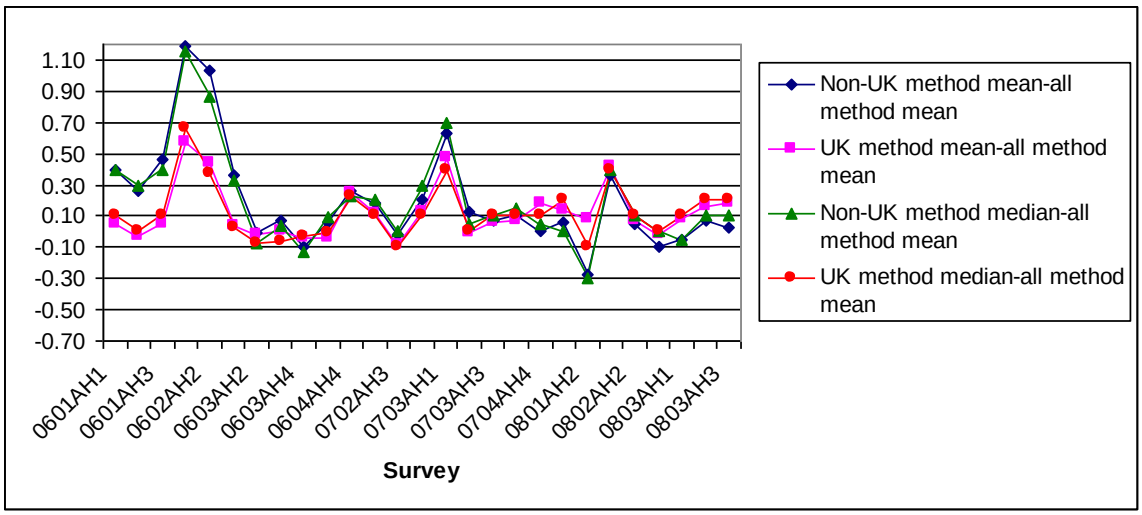

Figure 11: How the mean and median of the UK and non-UK Tosoh G7 groups results compare to the all methods mean from 2006 to 2008.

Figure 6 through to Figure 11 indicate that the BioRad Variant Classic and BioRad Variant II; Beta-thal tended to show a slight positive bias relative to the UK NEQAS all methods mean, whereas the BioRad Variant II; Dual Kit tended to show a slight negative bias. Also, the Tosoh G7 appeared to show a fairly strong positive bias for Hb A_2 whereas the Menarini HA 8160 demonstrated a fairly consistent negative bias.

Figure 6 appears to show that separating the UK and non-UK participant's results does have a significant effect on method mean and median for column chromatography as the UK group shows a slight positive bias relative to the all methods mean whereas the non-UK group shows a small negative bias. Similarly, Figure 7 shows that the UK method mean and median for the BioRad Variant Classic is higher relative to the UK NEQAS all methods mean than the non-UK group. Figure 8 through to Figure 11 do not appear to show any major differences between the UK and non-UK method mean and median for any of the other methods/submethods.

Standard Deviation

Figure 12 shows the overall change in SD of the Hb A_2 results submitted by UK plus non-UK participants from 2000 to 2008. The SDs were calculated using the median and interquartile range of the data sets.

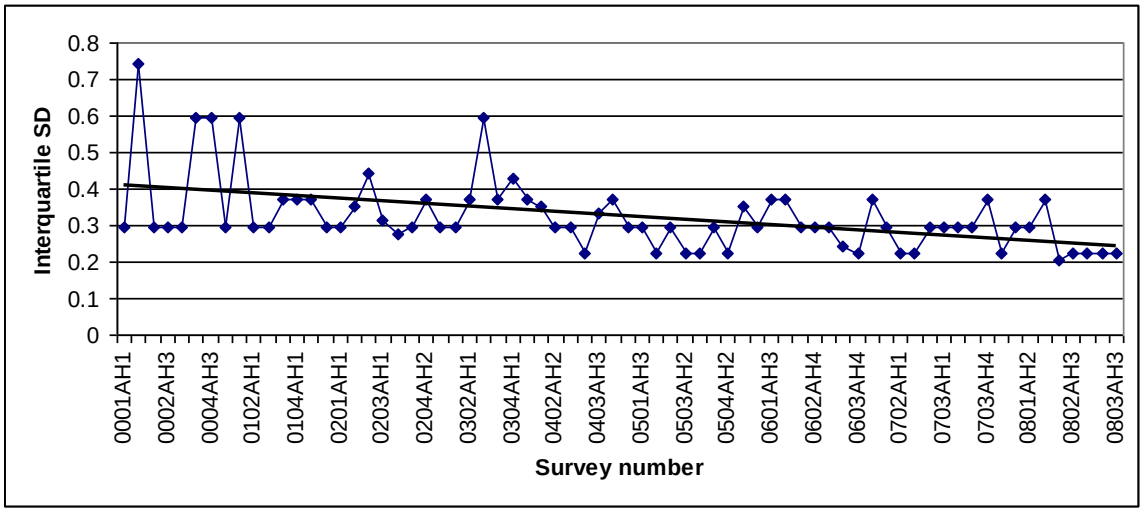

Figure 12: Total change in SD (based on median and interquartile range) of participants using all of the different method/submethod results from 2000 to 2008.

In Figure 12, the trendline shows that the SD of UK plus non-UK Hb A2 results has reduced from around 0.41 to 0.25, which shows that there has been an improvement in the SD of Hb A_2 measurement over the eight year period. The results of the UK and non-UK participants were then separated as shown in Figure 13.

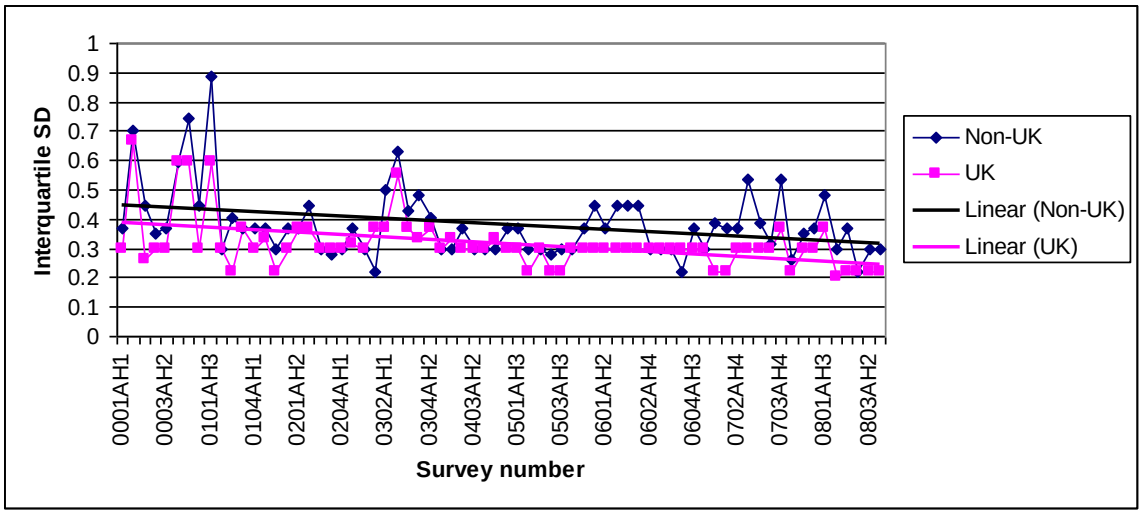

Figure 13: Change in SD (based on median and interquartile range) of the Hb A_2 results submitted by UK and non-UK participants from 2000 to 2008.

In Figure 13, the trendlines show that the SD of the UK data set is consistently around 0.07 lower than that of the non-UK data set but that the improvement in SD is similar for both groups.

Coefficient of Variation

The overall change in the CV of the Hb A_2 results submitted by UK plus non-UK participants is shown in Figure 14. These CVs were calculated by dividing the interquartile SD by the median of the data sets.

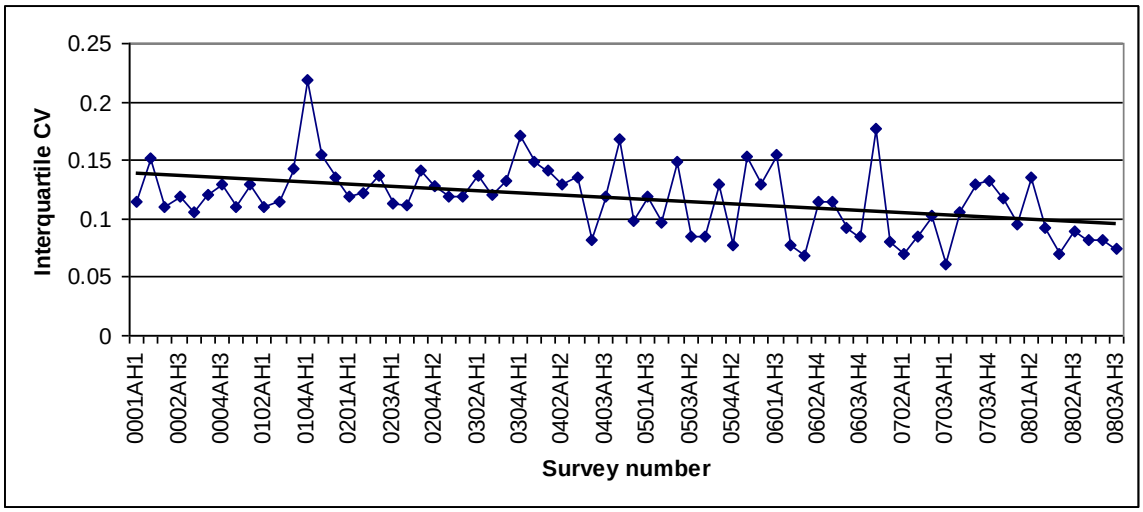

Figure 14: Total change in CV (based on median and interquartile range) of participants using all of the different method/submethod results from 2000 to 2008.

As with the overall SDs shown in Figure 12, Figure 14 shows an overall reduction in the CV of Hb A_2 measurement since 2000. There was found to be a similar improvement in the CV of results submitted by UK compared to non-UK participants, as shown in Figure 15.

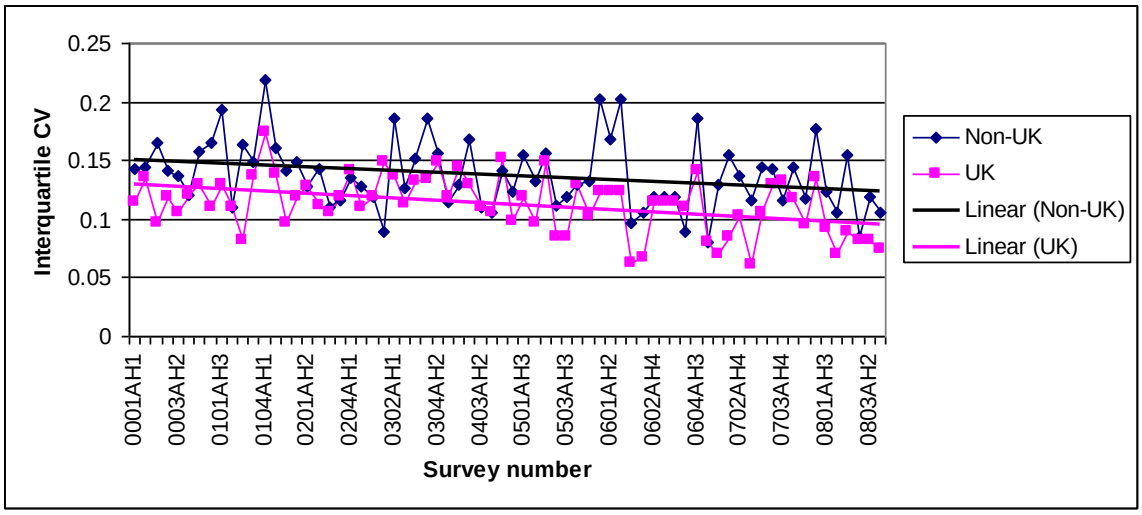

Figure 15: Change in CV (based on median and interquartile range) of the Hb A_2 results submitted by UK and non-UK participants from 2000 to 2008.

In Figure 15, the trendlines show that the improvement in CV is similar for both the UK and non-UK participant groups, although slightly better for the UK group.

Result Distribution

Three examples of the result distribution graphs generated for all surveys from 2006 to mid-2008 are shown in Figure 16 though to Figure 18. The 'All methods' lines on the graphs show the overall proportion of participants that submitted a particular Hb A_2 result. The other lines show the proportion of participants (UK plus non-UK) using six of the most common methods/submethods that submitted a particular Hb A_2 result for each specimen.

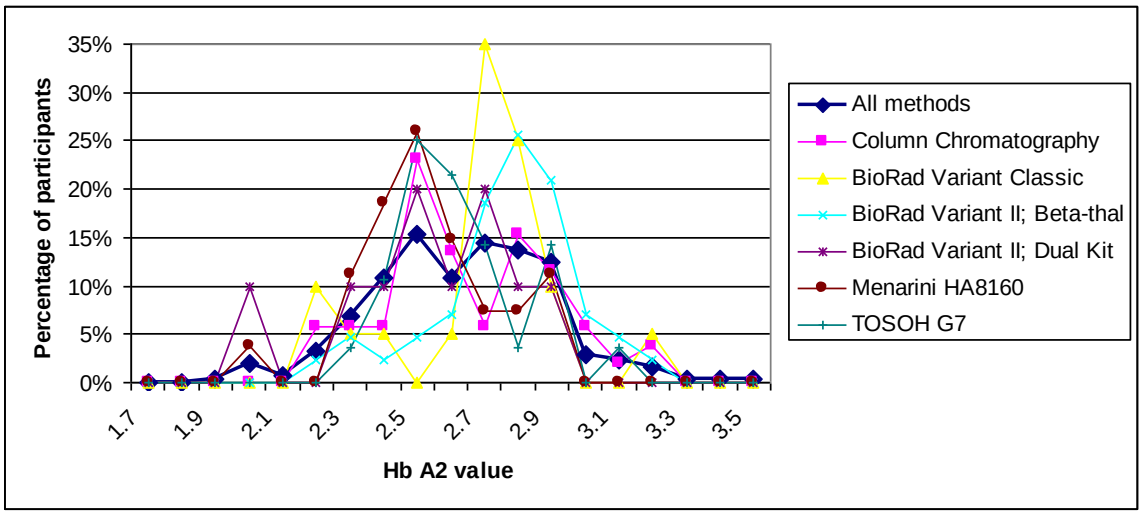

Figure 16: Distribution of Hb A_2 results submitted by participants for specimen 0603AH4 using six of the most common methods/submethods relative to the results submitted by all participants, represented by the 'All methods' line. The UK NEQAS all methods mean for this specimen was 2.6%.

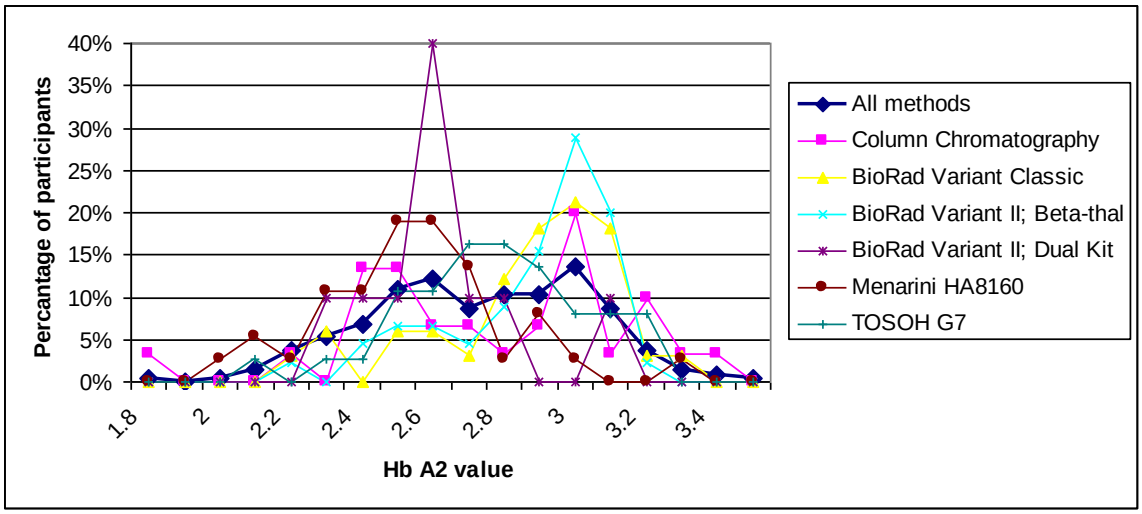

Figure 17: Distribution of Hb A_2 results submitted by participants for specimen 0703AH4 using six of the most common methods/submethods relative to the results submitted by all participants, represented by the 'All methods' line. The UK NEQAS all methods mean for this specimen was 2.7%.

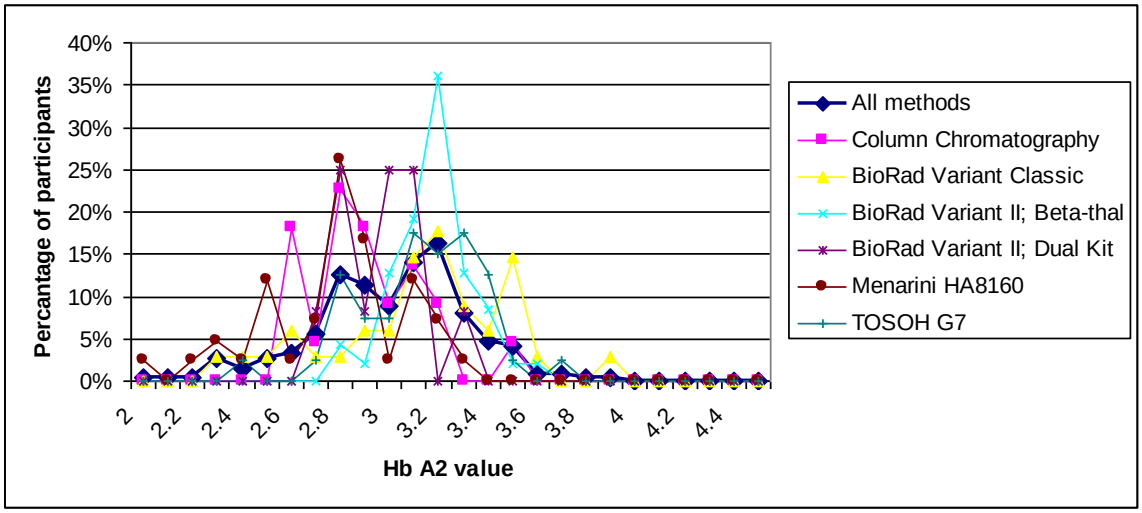

Figure 18: Distribution of Hb A_2 results submitted by participants for specimen 0801AH1 using six of the most common methods/submethods relative to the results submitted by all participants, represented by the 'All methods' line. The UK NEQAS all methods mean for this specimen was 3.0%.

The UK NEQAS all methods mean Hb A_2 values for the three specimens used to generate Figure 16 through Figure 18 were actually submitted as the result by far fewer participants than surrounding Hb A_2 values. This can be seen by dual peaks of the 'All methods' lines which reflect the high proportion of participants which use methods/submethods that tend to show a positive or negative bias. Of the six main method/submethod groups, the BioRad Variant II; Dual Kit and the Menarini HA8160 tended to show a negative bias relative to the all methods mean and comprise a significant proportion of the participants that cause the initial peak of the 'All methods' dual peaks. The BioRad Variant Classic, BioRad Variant II; Beta-thal and Tosoh G7 tended to show a positive bias relative to the all methods mean and are largely responsible for the second 'All methods' peak. The 'Column Chromatography' lines on the result distribution graphs tended to follow a similar pattern to the 'All methods' lines, with some participants

submitting relatively high Hb A_2 values and others submitting relatively low results.

Figure 19 and Figure 20 are examples of result distribution graphs generated for specimens that had borderline and high UK NEQAS all methods means respectively.

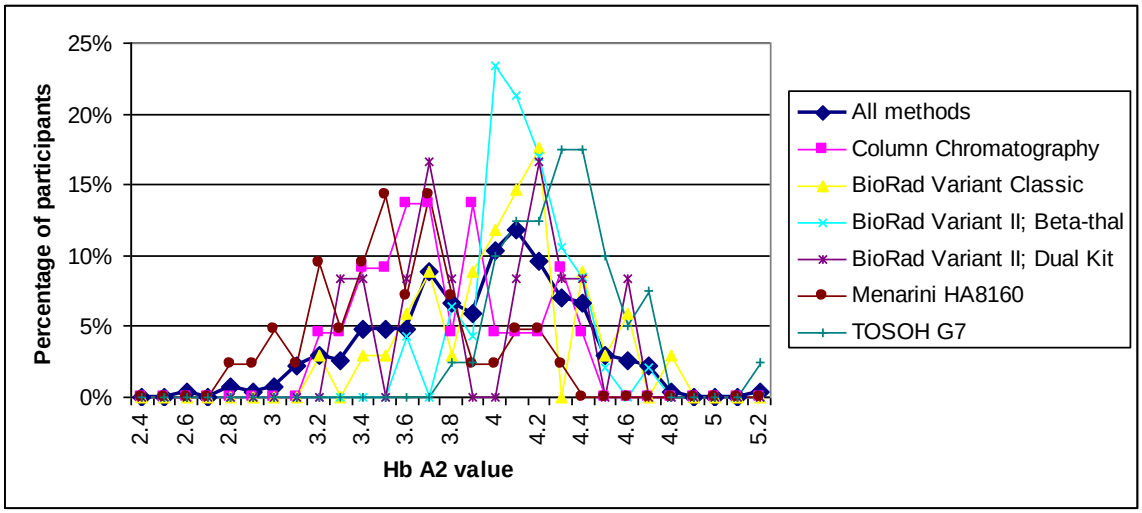

Figure 19: Distribution of Hb A_2 results submitted by participants for specimen 0801AH3 using six of the most common methods/submethods relative to the results submitted by all participants, represented by the 'All methods' line. The UK NEQAS all methods mean for this specimen was at a borderline level of 3.9%.

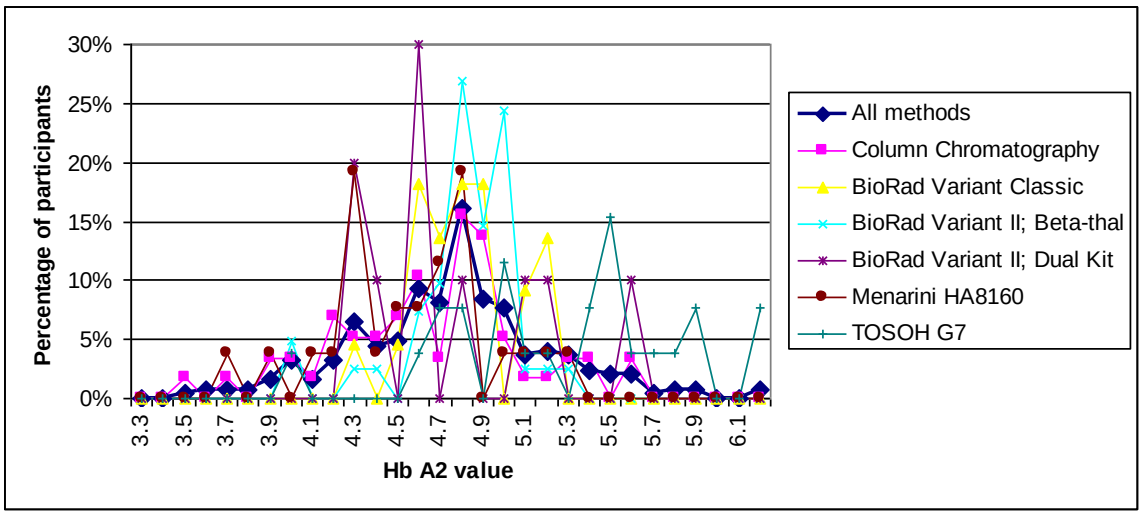

Figure 20: Distribution of Hb A_2 results submitted by participants for specimen 0602AH1 using six of the most common methods/submethods relative to the results submitted by all participants, represented by the 'All methods' line. The UK NEQAS all methods mean for this specimen was at a high level of 4.7%.

The result distribution graphs of the specimens with borderline and high Hb A_2 levels shown in Figure 19 and Figure 20 show similar trends to those of the previous three figures, except that when the Hb A_2 level is borderline or raised, the negative bias of the Menarini HA8160 and the positive bias of the Tosoh G7 groups become even more apparent.

Figure 21 and Figure 22 show the HPLC result distribution graphs for the specimens with borderline and high levels of Hb A_2 shown in Figure 19 and Figure 20. The 'All HPLC' lines on the graphs show the proportion of participants using any HPLC submethod that submitted a particular Hb A_2 result. The other lines show the proportion of participants (UK plus non-UK) using five of the most popular HPLC analysers that submitted a particular Hb A_2 result for each specimen.

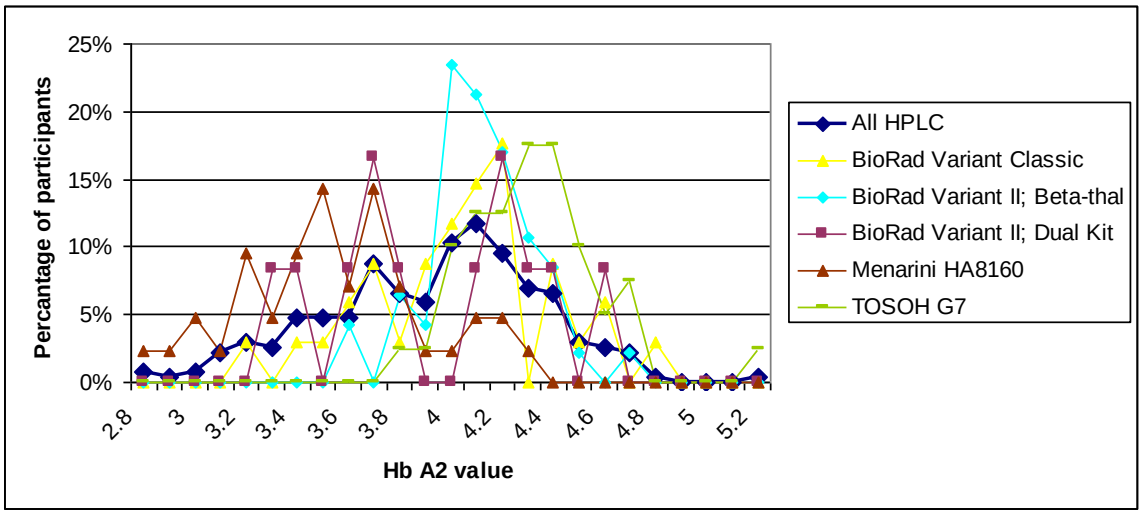

Figure 21: Distribution of Hb A_2 results submitted by participants for specimen 0801AH3 using five of the most common HPLC analysers relative to the results submitted by all HPLC users, represented by the 'All HPLC' line. The UK NEQAS HPLC mean for this specimen was at a borderline level of 3.9%.

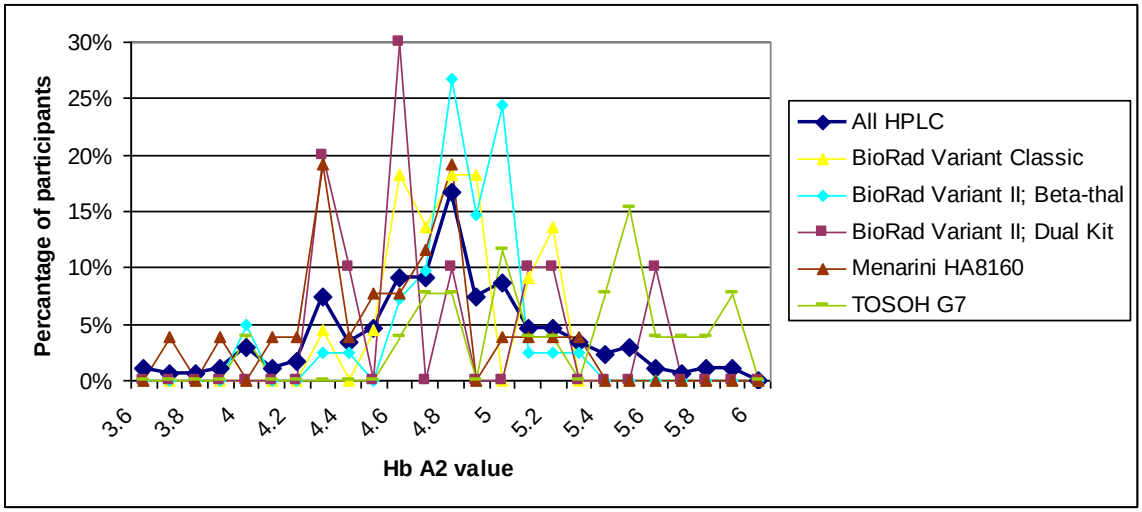

Figure 22: Distribution of Hb A_2 results submitted by participants for specimen 0602AH1 using five of the most common HPLC analysers relative to the results submitted by all HPLC users, represented by the 'All HPLC' line. The UK NEQAS HPLC mean for this specimen was at a high level of 4.8%.

The HPLC result distribution graphs shown in Figure 21 and Figure 22 obviously show the same trends as the result distribution graphs generated for the same specimens, however these trends are possibly even more apparent when studying the HPLC results in isolation.

Normal Distribution

Normal distribution curves were generated by plotting the median of the data sets ±3 SDs based on interquartile ranges. The results of all methods, plus those of six of the largest method/submethod groups, were plotted for all surveys from 2006 to mid-2008. The normal distribution curves generated from the results of four of these specimens, each with a very different Hb A_2 level, are shown in Figure 23 through to Figure 26. It should be noted that the area under each of the normal distribution curves is the same. Since the top of the peak is the mean and the limits are ±3 SDs, the taller and narrower the curve is, the smaller the SD for that particular data set.

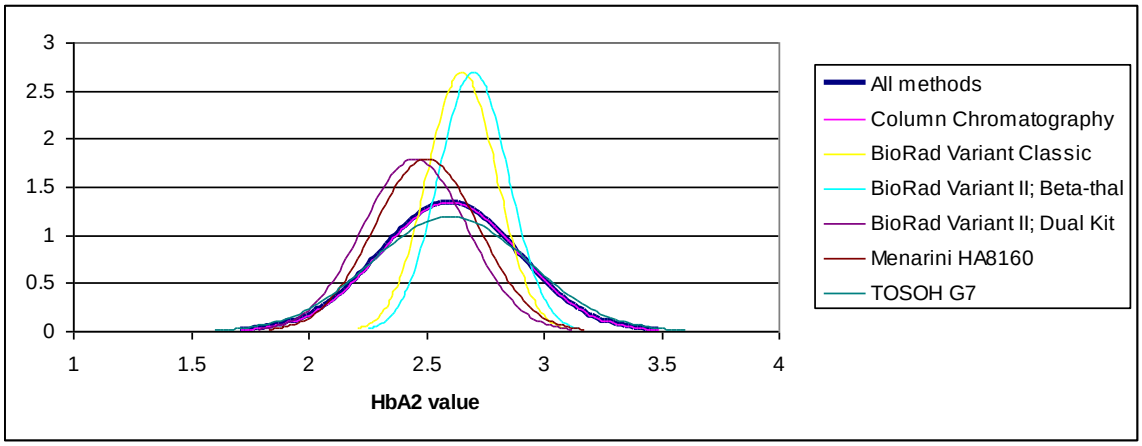

Figure 23: Normal distribution of Hb A_2 results for specimen 0602AH4 for six of the most common methods/submethods and for all methods (highlighted in bold). The curves are the median of the data sets ±3SDs (based on interquartile ranges). The UK NEQAS all methods mean for this specimen was 2.6%.

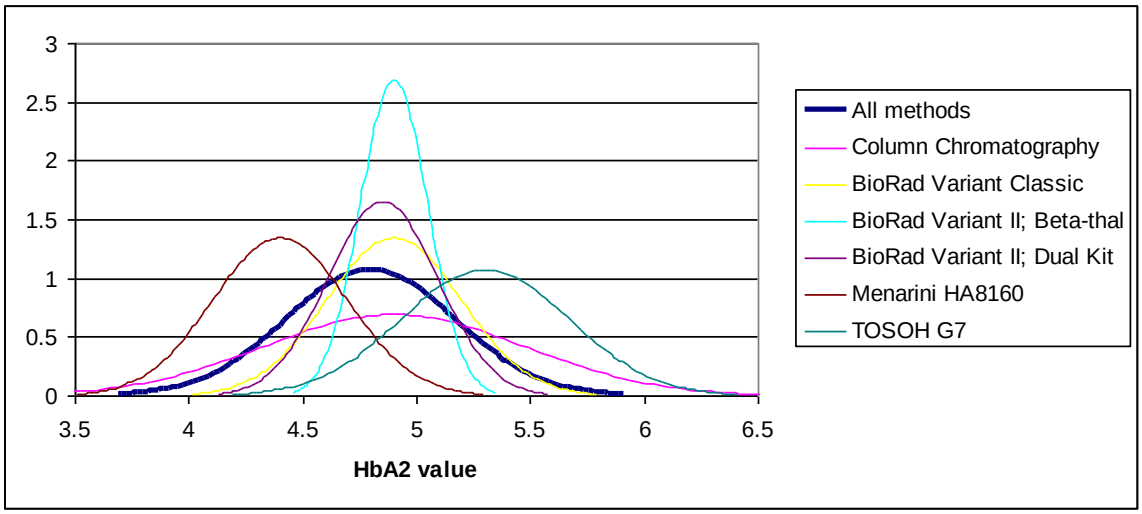

Figure 24: Normal distribution of Hb A_2 results for specimen 0703AH1 for six of the most common methods/submethods and for all methods (highlighted in bold). The curves are the median of the data sets ±3SDs (based on interquartile ranges). The UK NEQAS all methods mean for this specimen was 4.8%.

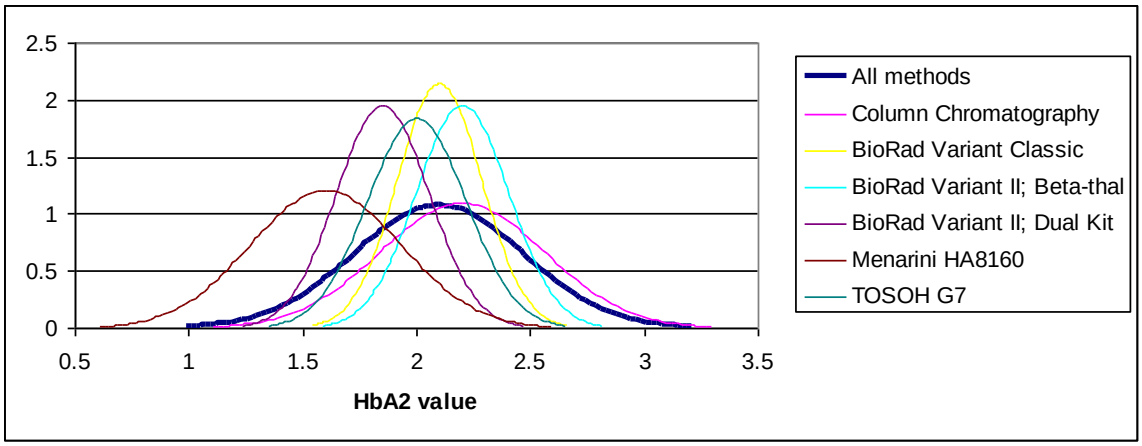

Figure 25: Normal distribution of Hb A_2 results for specimen 0604AH3 for six of the most common methods/submethods and for all methods (highlighted in bold). The curves are the median of the data sets ±3SDs (based on interquartile ranges). The UK NEQAS all methods mean for this specimen was 2.0%.

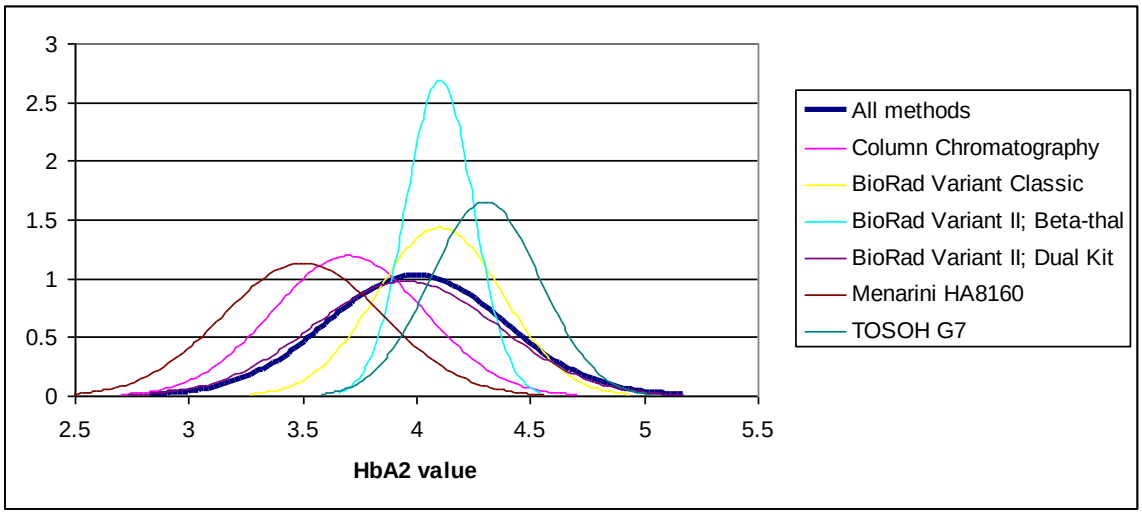

Figure 26: Normal distribution of Hb A_2 results for specimen 0801AH3 for six of the most common methods/submethods and for all methods (highlighted in bold). The curves are the median of the data sets ±3SDs (based on interquartile ranges). The UK NEQAS all methods mean for this specimen was 3.9%.

Figure 23 is an example of the type of pattern seen when the overall mean Hb A_2 is normal (in this case, 2.6%). Here, you can clearly see that the mean of the BioRad Variant II; Dual Kit and the Menarini HA8160 results are lower than the all methods mean, that the mean of the BioRad Variant Classic, BioRad Variant II; Beta-thal and the Tosoh G7 results are higher than the all methods mean and that the column chromatography results follow a similar pattern to the all methods data. Figure 24 is an example of a normal distribution curve generated when the all methods mean Hb A_2 result is high (in this case, 4.8%). In such cases, the Menarini HA8160 data tends to sit further left on the graph than the other data sets due to its negative bias and the Tosoh G7 data sits much further to the right due to it having a strong positive bias when the overall Hb A_2 result is raised. Figure 25 is an example of the typical normal distribution curves generated when the all methods mean Hb A_2 result is relatively low (in this case, 2.0%). The most obvious and consistent trend seen when the Hb A_2 result is low is the fact that the Menarini HA8160 data sits much further to the left of the graph than the other

data sets as Menarini HA8160 users generally show an even stronger negative bias when the overall mean Hb A_2 result is low. Finally, Figure 26 is an example of a normal distribution curve generated when the all methods mean Hb A_2 result is borderline (in this case, 3.9%). This graph shows the same trends as the previous graphs in that the most noticeable features are that the Menarini HA8160 curve is positioned to the left of the all methods curve and the Tosoh G7 curve is positioned to the right.

Cumulative Distribution

Cumulative distribution histograms were plotted for the Hb A_2 results of the four specimens with varying Hb A_2 levels used as examples in the Normal Distribution section above. The results of all methods, plus those of six of the most commonly used methods/submethods, were plotted as shown in Figure 27 through to Figure 30.

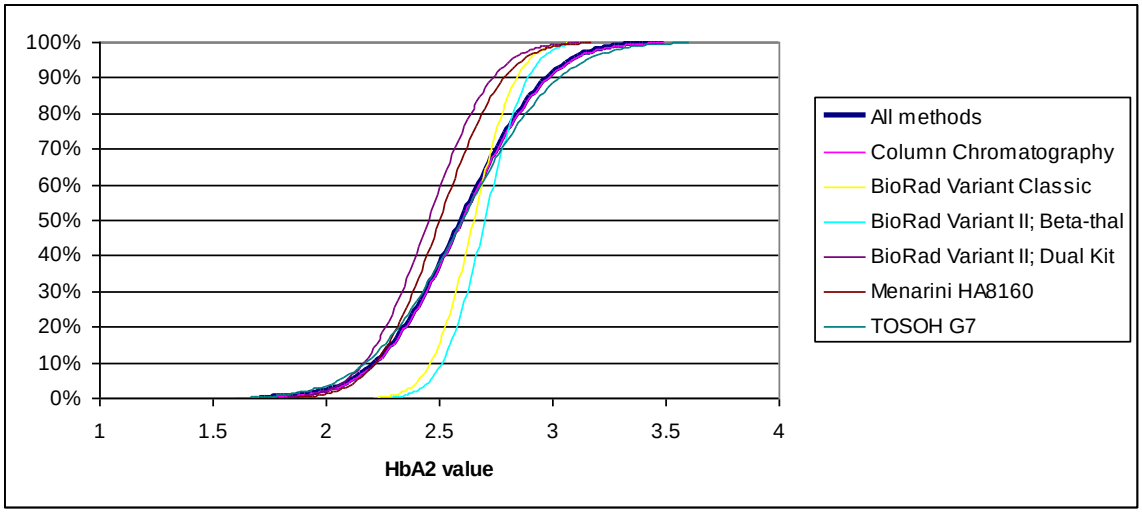

Figure 27: Cumulative distribution histogram of Hb A_2 results for specimen 0602AH4 for six of the most common methods/submethods and for all methods (highlighted in bold). The UK NEQAS all methods mean for this specimen was 2.6%.

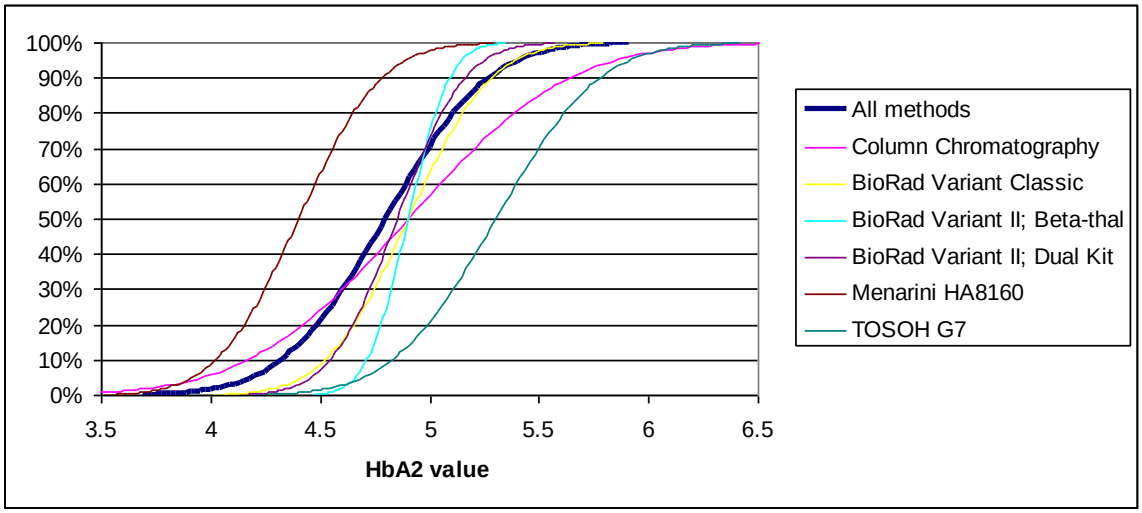

Figure 28: Cumulative distribution histogram of Hb A_2 results for specimen 0703AH1 for six of the most common methods/submethods and for all methods (highlighted in bold).

The UK NEQAS all methods mean for this specimen was 4.8%.

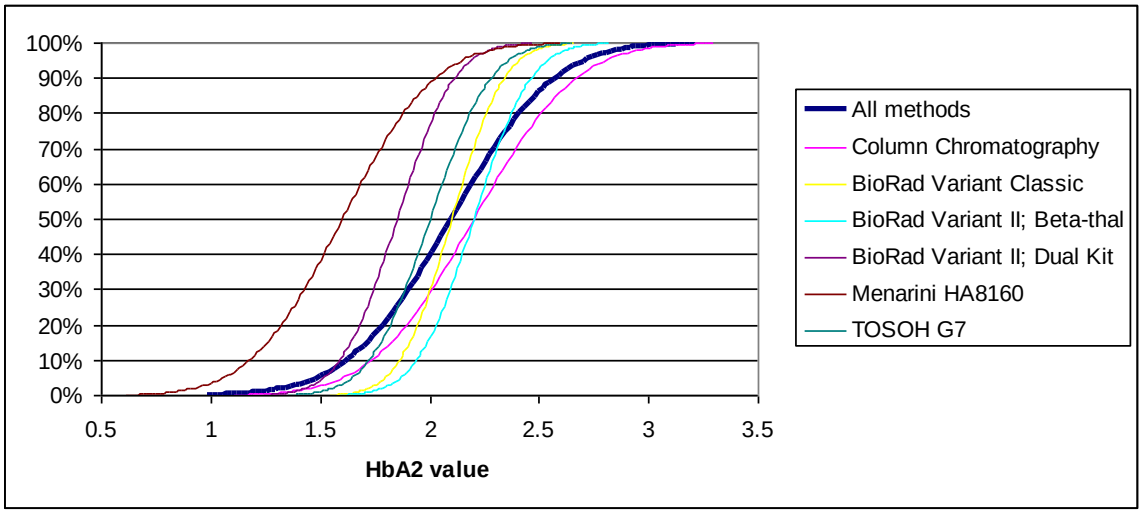

Figure 29: Cumulative distribution histogram of Hb A_2 results for specimen 0604AH3 for six of the most common methods/submethods and for all methods (highlighted in bold). The UK NEQAS all methods mean for this specimen was 2.0%.

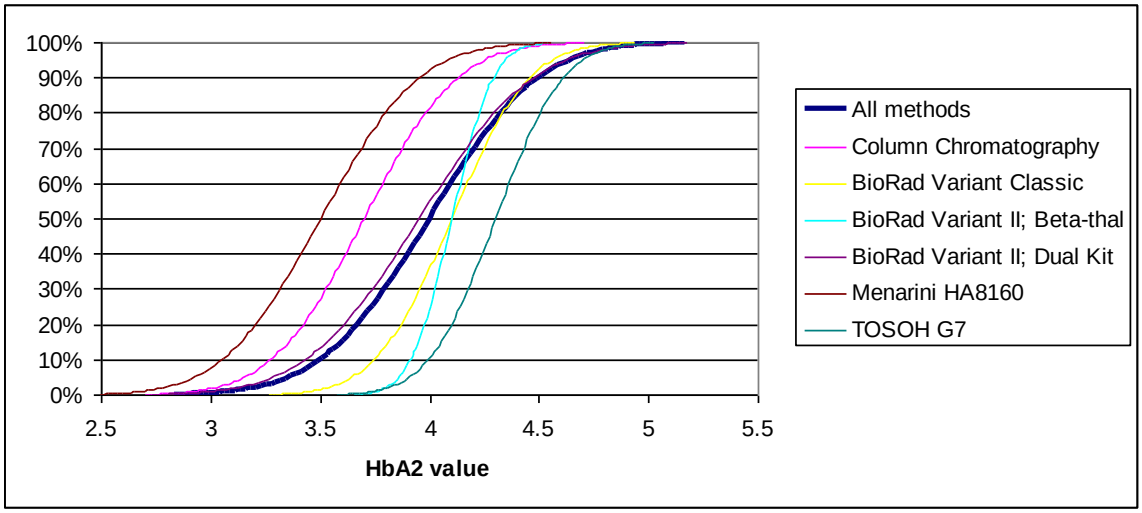

Figure 30: Cumulative distribution histogram of Hb A_2 results for specimen 0801AH3 for six of the most common methods/submethods and for all methods (highlighted in bold). The UK NEQAS all methods mean for this specimen was 3.9%.

Figure 27 is an example of the type of pattern seen when the overall UK NEQAS mean Hb A_2 is normal (in this case, 2.6%). As with the normal distribution curves for this specimen, you can clearly see that the results given by the BioRad Variant II; Dual Kit and the Menarini HA8160 groups are lower than the overall (all methods) results and that the results submitted by the BioRad Variant Classic and the BioRad Variant II; Beta-thal are generally slightly higher than the results of other method/submethod groups. Figure 28 is an example of a cumulative distribution histogram generated when the UK NEQAS all methods mean HbA_2 result is high (in this case, 4.8%). This shows that the Menarini HA8160 results sit further left on the graph than the other data sets due to its negative bias and the Tosoh G7 results sit much further to the right due to it having a strong positive bias when the overall Hb A_2 result is raised. Figure 29 is an example of a cumulative distribution histogram generated when the UK NEQAS all methods mean HbA_2 result is relatively low (in this case, 2.0%). Again, the Menarini

HA8160 data sits much further to the left of the graph than the other data sets which is consistent with previous findings that Menarini HA8160 users generally show an even stronger negative bias when the overall mean HbA_2 result is low. Finally, Figure 30 is an example of a cumulative distribution histogram generated when the UK NEQAS all methods mean HbA_2 result is borderline (in this case, 3.9%). This histogram shows the same trends as the previous ones in that the most noticeable features are that the Menarini HA8160 curve is positioned to the left of the other methods/submethods and the Tosoh G7 curve is positioned to the right.

Borderline Hb A_2 Specimen

A specimen with a borderline Hb A_2 level was sent out in 2009 and the UK NEQAS all methods mean Hb A_2 for this specimen was calculated to be 3.7%. Normal distribution curves and a cumulative distribution histogram were generated using the results of this specimen as shown in Figure 31 and Figure 32.

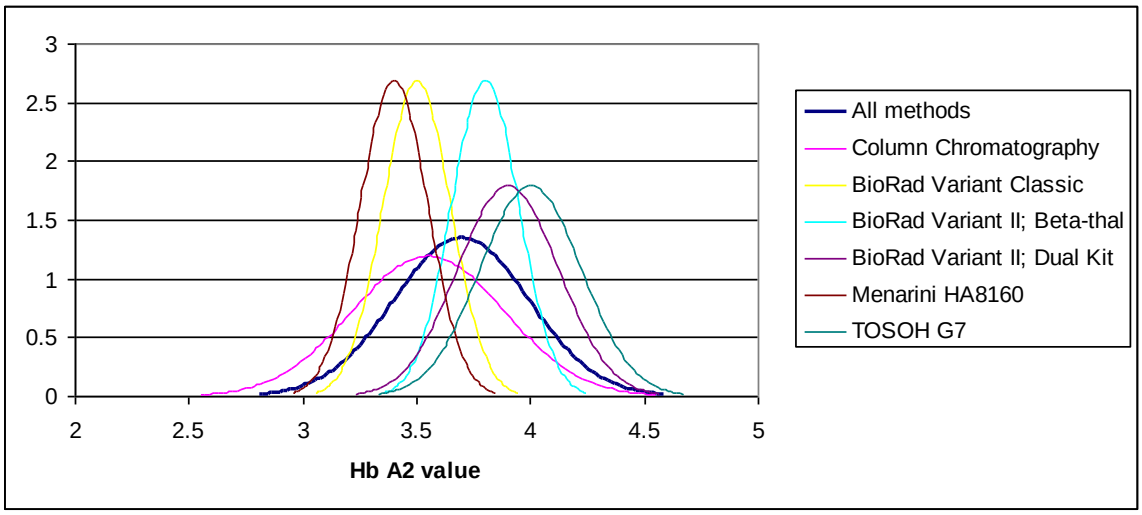

Figure 31: Normal distribution of Hb A_2 results for specimen 0902AH1 for six of the most common methods/submethods and for all methods (highlighted in bold). The curves are the median of the data sets ±3SDs (based on interquartile ranges). The UK NEQAS all methods mean for this specimen was 3.7%.

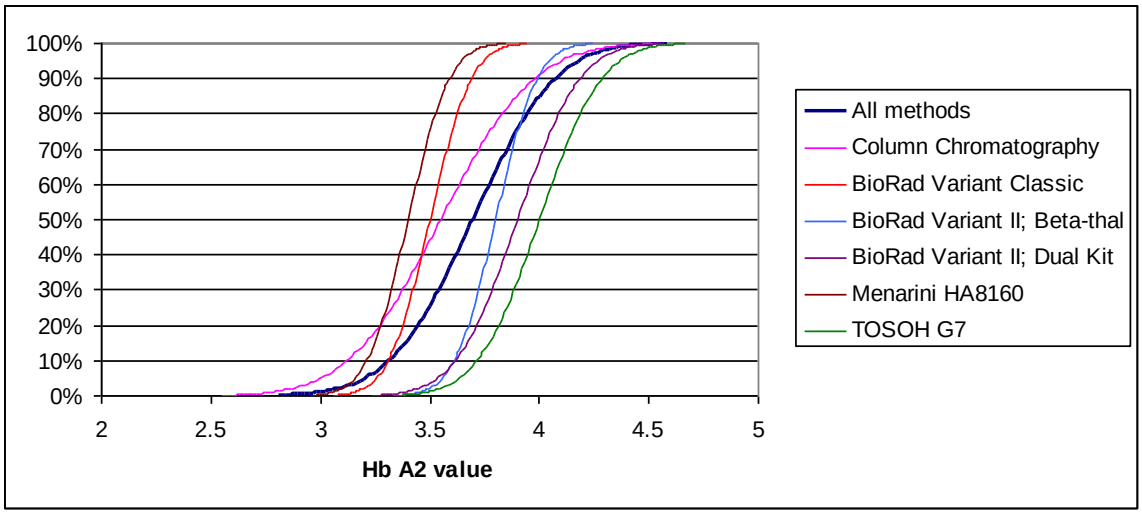

Figure 32: Cumulative distribution histogram of Hb A_2 results for specimen 0902AH1 for six of the most common methods/submethods and for all methods (highlighted in bold). The UK NEQAS all methods mean for this specimen was 3.7%.

The results shown in Figure 31 and Figure 32 are consistent with previous findings that the Menarini HA8160 has a negative bias relative to other methods and the Tosoh G7 shows a positive bias when the Hb A_2 level is borderline/raised.

At the time of this survey, there were 270 participants registered for the abnormal haemoglobins scheme, 175 of which were UK based. When UK NEQAS examined the interpretive comment codes submitted by participants, they found that 64 of the UK laboratories did not use the beta thalassaemia carrier code for this specimen, although 16 out of the 64 did note the possibility of beta thalassaemia carrier status in a free text comment. 26 of the 64 UK participants returned an Hb A_2 measurement of 3.4% or less. 22 of the 64 laboratories returned an Hb A_2 measurement of between 3.5 and 3.8%, but neither used the beta thalassaemia carrier code nor mentioned beta thalassaemia carrier status in free text.

28 of the 64 UK laboratories (44%) that did not use the beta thalassaemia carrier interpretive code were Menarini HA8160 users, despite the fact that only 22% of the UK participants were using this particular instrument at that time. 24 of these 28 participants returned an Hb A_2 measurement of 3.4% or less and the remaining 4 submitted a result of 3.5%.

T-Tests

If the null hypothesis of no difference between the results submitted by UK and non-UK participants was true, then we would expect about 5% of the t-tests to generate a P value of less than 0.05 by chance. Heteroscedastic, unpaired t-tests were used to compare the UK and non-UK participants results from column chromatography, BioRad Variant Classic, BioRad Variant II; Beta thal, Menarini HA8160 and Tosoh G7 users for each survey between 2006 and mid-2008. Of the 135 t-tests carried out, however, 1 in 5 (27) gave a value of less than 0.05. Of the tests performed to compare the results of UK and non-UK column

chromatography users, 1 in 2 (13 out of 27) gave a P value of less than 0.05 and of those performed to compare the results submitted by UK and non-UK participants using the different HPLC submethods, 1 in 8 (14 out of 108) gave a P value of less than 0.05.

UK Interquartile SDs

The standard deviation (SD) of the various participant groups (UK only) was calculated using the median and interquartile range of the data sets for each survey from 2006 to mid-2008. For each specimen, participant's results were analysed in three different ways. They were compared to:

1) The overall median comprising all of the different methods.
2) The median of their method group where greater than or equal to 20 participants were using that method.
3) The median of their submethod group where greater than or equal to 20 participants are using that particular analyser.

Those participants whose results fell outside of a ±2SD limit (based on the appropriate median and interquartile range) using each of these three different approaches were then compared, as shown in Figure 33 and Figure 34.

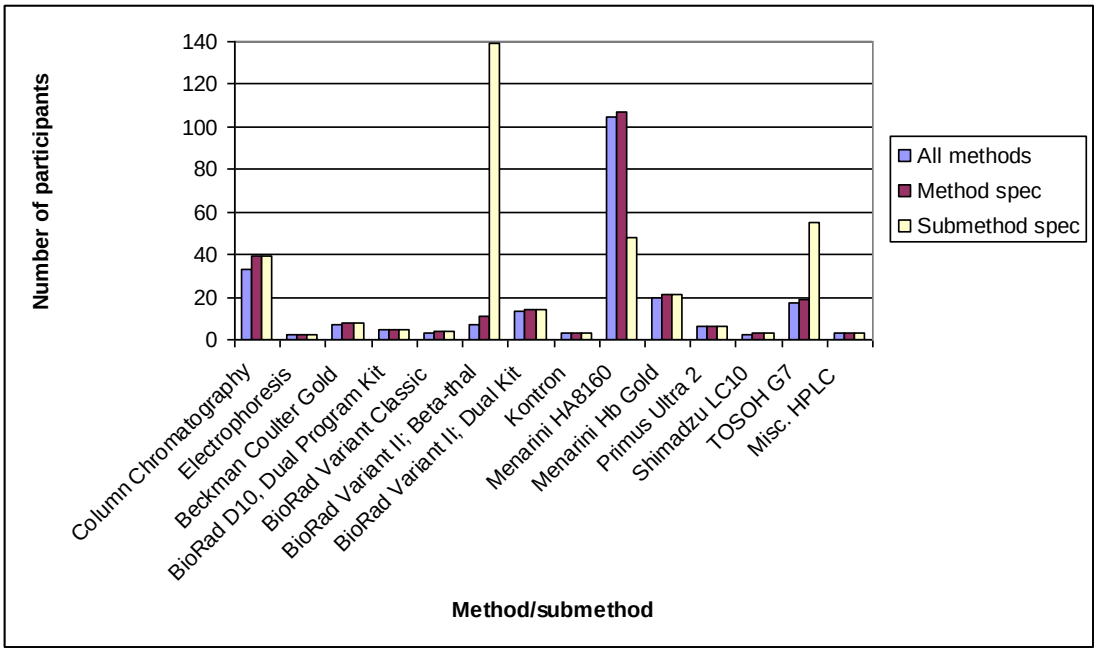

Figure 33: Difference in the composition of participants with results greater than 2SDs below the median depending on whether their results were compared to the all methods median, method specific median or submethod specific median.

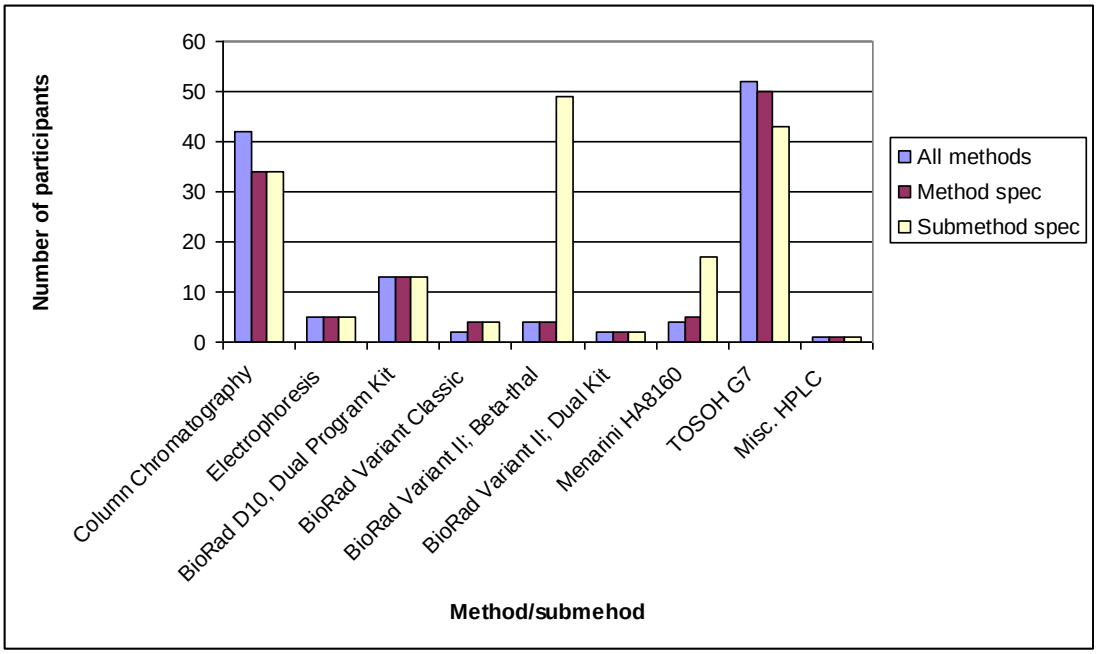

Figure 34: Difference in the composition of participants with results greater than 2SDs above the median depending on whether their results were compared to the all methods median, method specific median or submethod specific median.

As you can see from Figure 33 and Figure 34, the composition of participants with results greater than 2SDs above or below the median varies very little between results being compared to the all methods median and to the method specific median. When participants results are compared to the submethod specific median, however, the composition of those outside the ±2SD limit changes quite dramatically. The submethod median comparison only applies to the BioRad Variant II; Beta-thal, Menarini HA8160 and Tosoh G7 groups as only these submethods had greater than or equal to 20 participants in the UK from 2006 onwards. Figure 33 and Figure 34 show, however, that this method of analysing participants results causes approximately a fourteen-fold increase in the number of BioRad Variant II; Beta-thal users that are greater than ±2SDs from the median. It also increases the number of Menarini HA8160 users that are greater than 2SDs above the median but reduces the number that are greater than 2SDs below the median with an average overall 41% reduction in the number of participants greater than ±2SDs from the median when compared to the other methods of comparing the results. The Tosoh G7 group predictably shows the opposite trend to that of the Menarini HA8160 in that this way of comparing results reduces the number of Tosoh G7 users that are greater than 2SDs above the median but increases the number that are greater than 2SDs below the median with an average overall 42% increase in the number of participants greater than ±2SDs from the median.

Overall, when participants results were compared to the all methods median, 351 participants (7%) were found to give answers that were outside of the ±2SD range in the surveys between 2006 and mid-2008 (226 participants were found to give answers that were more than 2SDs below the median and 125 participants were found to give answers that were more than 2SDs above the median). When participants results were compared to the method specific median, similarly 363

participants (7%) were found to give answers that were outside of the ±2SD range between 2006 and mid-2008 (245 participants were found to give answers that were more than 2SDs below the median and 118 participants were found to give answers that were more than 2SDs above the median). When, however, participants results were compared to the submethod specific median, a much higher 518 participants (10.5%) were found to give answers that were outside of the ±2SD range (350 participants were found to give answers that were more than 2SDs below the median and 168 participants were found to give answers that were more than 2SDs above the median).

Performance Scoring

Using a system based on the current UK NEQAS scoring system of comparing participants results to the method specific median, adding up the deviation indices of the last 6 specimens and multiplying by 9, a total of 9 separate scores were calculated for each UK participant using the results of surveys 0601AH through to 0803AH. This was then repeated using the all methods median and submethod specific median in the calculation so that effects of comparing participant's results to the all methods median, method specific median and submethod specific median could be compared as shown in Figure 35.

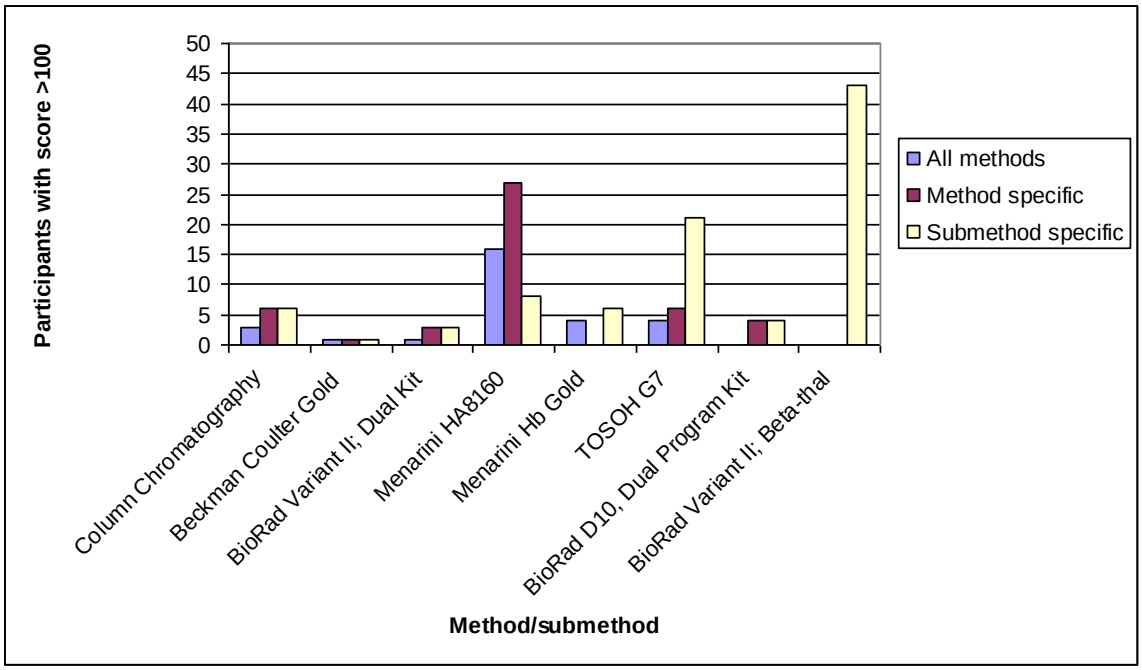

Figure 35: Difference in the composition of UK participants with performance scores greater than or equal to 100 depending on whether their results were compared to the all methods median, method specific median or submethod specific median.

Except for the Menarini HA8160 group, the number of participants with scores greater than 100 appears to vary very little between when results are compared to the all methods median and to the method specific median. For the Menarini HA8160 group, fewer users have a score greater than or equal to 100 when compared to the all methods median than when compared to the HPLC group median. When participants results are compared to the submethod specific median, the composition of those with scores greater than 100 changes quite

dramatically in the submethod groups with greater than or equal to 20 participants (BioRad Variant II; Beta-thal, Menarini HA8160 and Tosoh G7). Comparing participants results to the submethod median has the overall effect of reducing the number of Menarini HA8160 users with scores greater than or equal to 100 but increases the number of Tosoh G7 and BioRad Variant II: Beta-thal users with scores greater than or equal to 100.

Overall, it was found that in the 9 scores generated from the results of 0601AH to 0803AH, an average of 2.0% of participants had scores greater than or equal to 100 when comparing results to the all methods median, 3.3% had scores greater than or equal to 100 when comparing results to the method specific median and 6.4% had scores greater than or equal to 100 when comparing results to the submethod specific median.

One of the aims of this study was to try to modify the scoring system so that 5% of participants are identified as persistent unsatisfactory performers. Calculations were therefore performed to estimate the cut-off for Persistent Unsatisfactory Performance (PUP) required in order to achieve this. The current scoring philosophy employed by UK NEQAS(H) is to use a performance score cut-off of greater than or equal to 100 to define persistent unsatisfactory performance and that the multiplier in the performance scoring equation can be altered to adjust the proportion of participants that score above this cut-off. However it was found that when comparing participant's results to the all methods median or the method specific median, a lower PUP score cut-off of greater than or equal to 90 would be an alternative way of identifying around 5% of participants as above the cut-off. When comparing participant's results to the submethod specific median, however, a higher PUP score cut-off of greater than or equal to 110 would be needed to identify around 5% of participants as above the cut-off.

Experiments were also performed to look into the more typical way of adjusting the UK NEQAS scoring system by altering the figure by which the sum of the previous 6 deviation indices is multiplied by in the performance scoring calculation to achieve 5% of participants as having a PUP score greater than or equal to 100. Based on an average of 9 sets of Σ6DI from 2006 to mid-2008, it was calculated that the Σ6DI should be multiplied by 10 (to the nearest whole number) when comparing participants results to either the all methods median or the method specific median in order to obtain 5% of participants with scores greater than or equal to 100. When comparing participants results to the submethod specific median, however, it was found that the Σ6DI should continue to be multiplied by 9 (to the nearest whole number) in order to obtain 5% of participants with scores greater than or equal to 100.

Table 1 shows the multipliers that would have to be applied to the performance scoring system of each method/submethod group in order to obtain equal proportions (5%) of persistent unsatisfactory performers within each group. If UK NEQAS chose to compare participant's results to the all methods median, the third column of the table gives the multiplier that would have to be applied to the scoring system of those method/submethod groups with greater than or equal to 20 participants in order to obtain around 5% persistent unsatisfactory performers within each group. The fourth column shows the multiplier that would have to be

applied to the scoring system of each group to obtain around 5% persistent unsatisfactory performers within each group if participants results were to remain being compared to the method specific median. If UK NEQAS chose to compare participant's results to the submethod specific median, the fifth column shows the multiplier that would have to be applied to the scoring system of each group to obtain around 5% persistent unsatisfactory performers within each group. The required multipliers were calculated to the nearest whole number as UK NEQAS had decided that they would only ever use whole number multipliers in the performance scoring calculation.

Method/ Submethod	Average number of UK participants	Multiplier required for all methods comparison	Multiplier required for method specific comparison	Multiplier required for submethod specific comparison
BioRad Variant II; Beta-thal	39	16	16	7
Menarini HA8160	29	9	9	10
Tosoh G7	31	11	11	9
Column Chromatography	38	10	10	10

Table 1: The different multipliers required in the performance score calculation in order to identify 5% of UK participants within each method/submethod group as persistent unsatisfactory performers depending on whether participant's results are compared to the all methods, method specific or submethod specific median.

Normal Ranges

The minimum and maximum levels of the normal ranges for Hb A_2 used by UK and non-UK participants were plotted along with the midpoint of the range. The reference ranges used by UK participants are shown in Figure 36. The midpoint markers are colour coded so that the ranges of participants using the same method/submethod have the same colour midpoint marker.

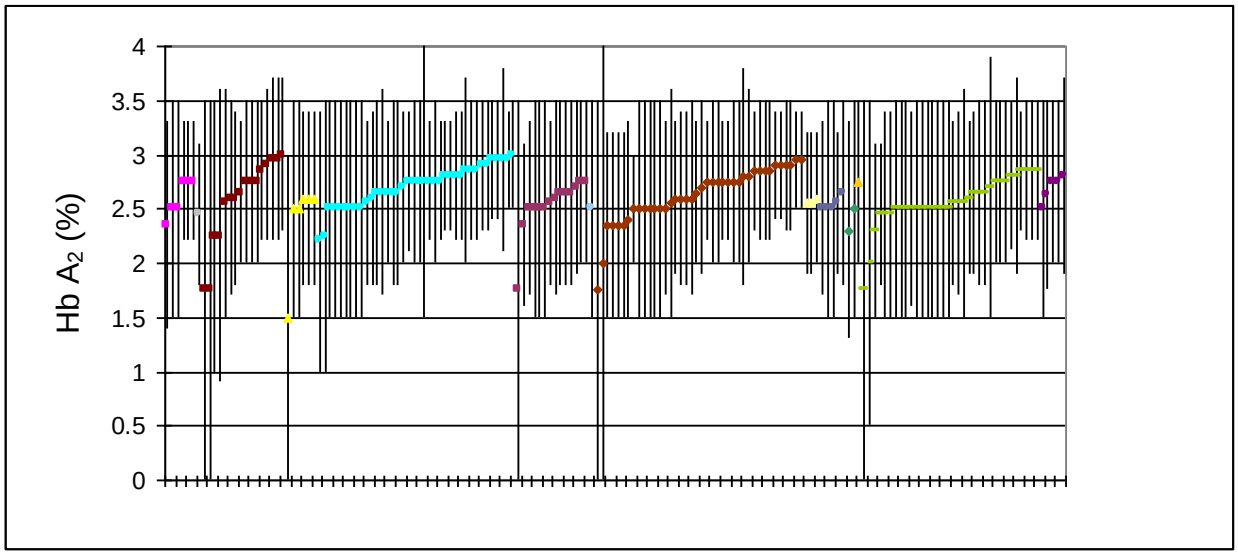

Figure 36: The Hb A_2 normal reference ranges used by UK participants.

Comparison of the normal ranges for Hb A_2 used by UK participants revealed differences in the reference ranges used by participants using the same methods/submethods as well as those using different ones.

Hb A_2 Assessment Codes

The Hb A_2 assessment codes submitted for 62 different specimens between 2000 and 2008 were studied and the 20 participants with 3 or more 'outwith consensus' assessment results in the 23 specimens between 2006 and mid-2008 were investigated further. It was found that 9 of these participants were UK based and 11 were non-UK labs.

Figure 37 shows the likely causes of the 'outwith consensus' assessment codes submitted by the 20 participants that gave 3 or more 'outwith consensus' assessment results between 2006 and mid-2008. It shows that 32% of these 'outwith consensus' comments appear to be due to generation on an 'outwith consensus' Hb A_2 result by the laboratory, 37% are likely to be due to the use of a normal range that differs from most other participants and just 5% appear to have arisen from transcription or assessment errors. It was thought that 27% of the comments were likely to be 'outwith consensus' due to a combination of factors, most of which were due to both generation of an 'outwith consensus' Hb A_2 result and the use of a normal range that differed to most other participants.

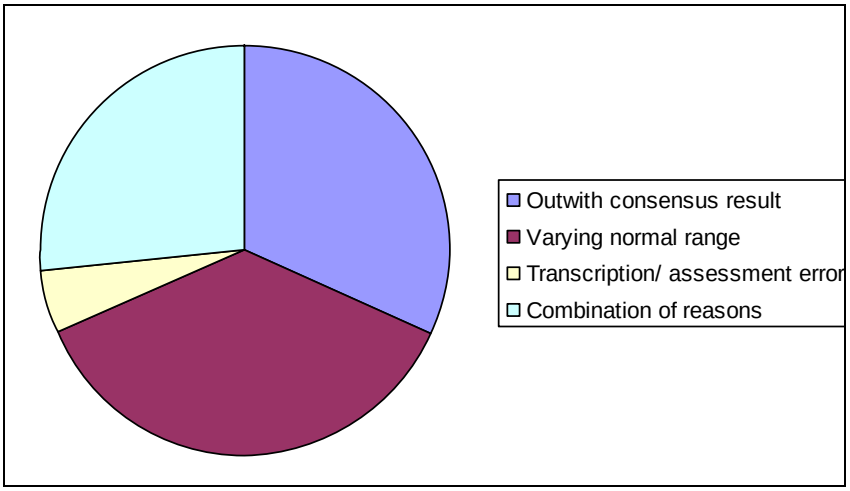

Figure 37: Likely causes of the 'outwith consensus' Hb A_2 assessment codes submitted by participants that have given 3 or more such results from 2006 to mid-2008.

Figure 38 shows the methodology used by the participants that submitted 3 or more 'outwith consensus' Hb A_2 interpretation results between 2006 and mid-2008. The largest group of these participants were found to use a Menarini HA8160 HPLC analyser for Hb A_2 measurement.

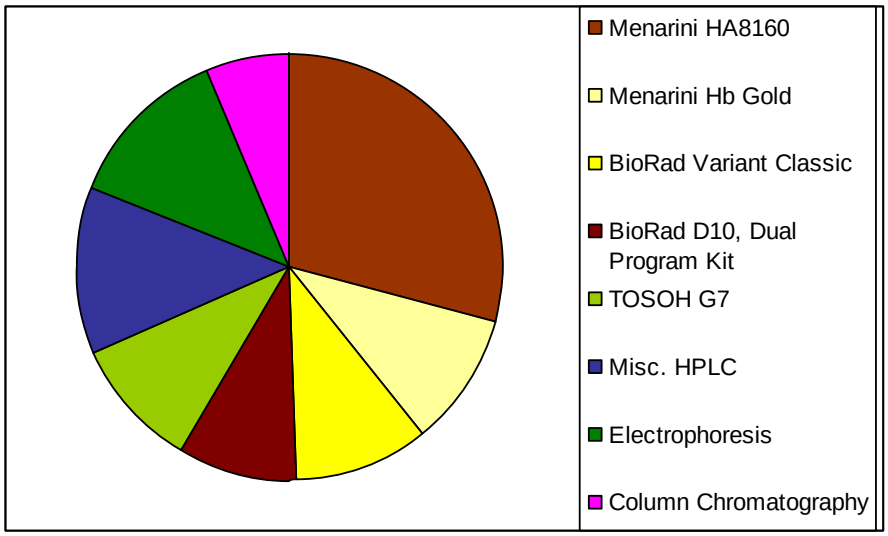

Figure 38: Methods/submethods used to generate the 3 or more 'outwith consensus' Hb A_2 interpretation results submitted by the 20 participants from 2006 to mid-2008.

Interpretive Comment Codes

All of the 'outwith consensus' comment codes submitted by participants for almost all surveys from 2006 to mid-2008 were listed and the comments that are relevant to beta thalassaemia were investigated further. 30 participants were found to have used comments which state 'no evidence of thalassaemia' or 'no evidence of beta thalassaemia' for specimens which were from beta thalassaemia carrier individuals. As with the Hb A_2 assessment codes, the likely causes of the 'outwith consensus' comment codes were hypothesised and the results of this investigation are shown in Figure 39.

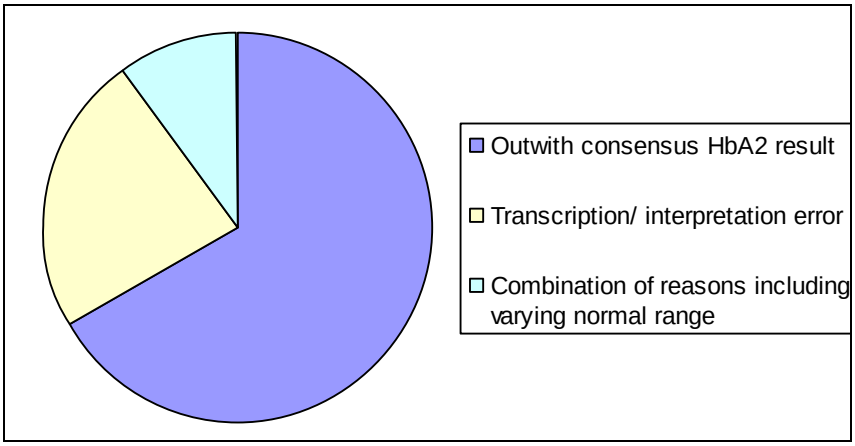

Figure 39: Likely reasons behind why participants used comment codes which ruled out thalassaemia in specimens from individuals with beta thalassaemia.

The generation of Hb A_2 results below the 3.5% cut off appeared to be responsible for why beta thalassaemia was missed by the majority of the participants (67%) that submitted comments stating that there was 'no evidence of (beta) thalassaemia' for specimens from beta thalassaemia carrier individuals. In 23% of cases where a 'no evidence of (beta) thalassaemia' comment was used incorrectly, a transcription or result interpretation error appeared to be the likely cause and in one case, it seem apparent that the error was due to a

transposition between UK NEQAS specimen results. The remaining 3 cases (10%) appeared to be due to a combination of reasons including varying normal ranges used by participants. It was also noted that of these 30 laboratories that used codes which suggested that they missed the fact that the specimen was from a beta thalassaemia carrier, 37% were UK labs and 63% were based outside of the UK.

It was also found that between 2006 and mid-2008, 42 participants had used comments incorrectly suggesting beta thalassaemia in specimens that were not from beta thalassaemia carriers and 21 (50%) of these specimens were from sickle cell carriers. Of these 21 sickle cell carrier specimens, 11 were actually at risk of coexisting alpha thalassaemia rather than beta thalassaemia.

For the 21 participants that incorrectly used the beta thalassaemia comment code in the absence of Hb S, we again tried to establish if the incorrect comment was submitted due to the generation of an 'outwith consensus' Hb A_2 result, the use of a different normal range to most other participants or a transcription or interpretation error. The findings are shown in Figure 40.

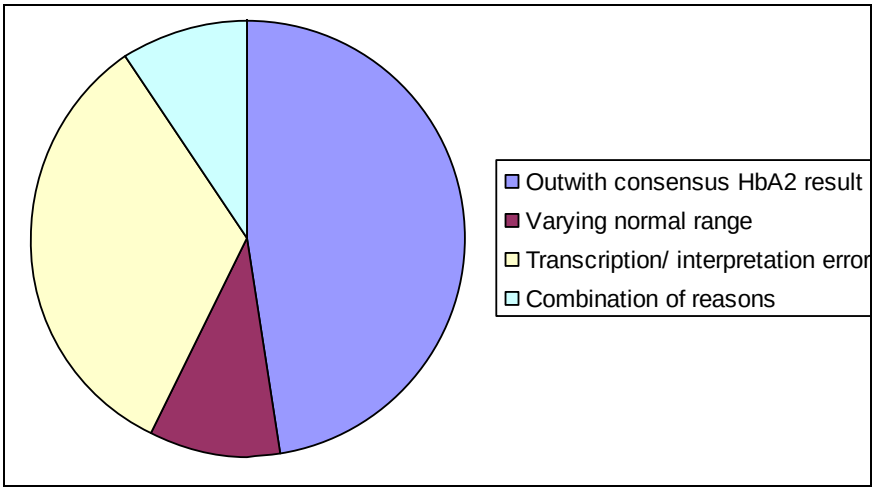

Figure 40: Likely reasons behind why participants used comment codes which suggested beta thalassaemia in specimens from individuals who did not carry beta thalassaemia or Hb S.

The generation of falsely high Hb A_2 results again appeared to be responsible for why beta thalassaemia was suggested by the majority of the participants (48%) that submitted comments incorrectly stating that the specimen was from a beta thalassaemia carrier individual. Transcription or result interpretation errors also appeared to be the likely cause of a large proportion of the incorrect beta thalassaemia carrier comments (33%). The use of varying normal ranges between participants and a combination of reasons including the generation of an 'outwith consensus' Hb A_2 result also appeared to account for a number of these comments in specimens which were not from carriers of beta thalassaemia or Hb S.

Of the 42 participants that incorrectly concluded that specimens were from beta thalassaemia carriers, 43% were UK-based and 57% were non-UK labs.

Discussion

Methodology

Figure 1 clearly shows how a large number of laboratories in the UK have changed from using column chromatography to HPLC for measuring Hb A_2 over the last eight years. Figure 2 shows that although there has been a reduction in the number of non-UK labs using column chromatography since 2000, the main driving force behind the change in methodology of non-UK participants has been due to new labs joining the scheme which mostly use HPLC. Also, a small number of non-UK participants measure Hb A_2 using automated densitometry or electrophoresis and elution.

Figure 3 shows that the most commonly used HPLC instruments in the UK for Hb A_2 measurement are currently the BioRad Variant II; Beta-thal, the Menarini HA8160 and the Tosoh G7, followed by the other BioRad analysers. Figure 4 shows that amongst non-UK participants, the BioRad Variant Classic is by far the most popular HPLC analyser.

Method/Submethod Bias

The results shown in Figure 5 were meant to demonstrate which particular methods and submethods have a positive or negative bias relative to the UK NEQAS all methods mean. By calculating the group mean minus the UK NEQAS all methods mean (at 0), the graph was designed to show how far the means of the different method/analyser groups results were above or below the UK NEQAS all methods mean.

Figure 5 shows that the submethod means of the BioRad D10 Dual Program Kit, BioRad Variant Classic, BioRad Variant II; Beta-thal and the Tosoh G7 all tended to show a positive bias relative to the all methods mean, whereas the BioRad Variant II; Dual Kit and the Menarini HA8160 tended to show a negative bias. This way of analysing the data, however, is of limited use as the method/submethod group data sets were not trimmed and therefore the calculated mean values may have been skewed by outlying results.

To try to see what effect this potential skewing of the mean has, as well as comparing the method/submethod mean to the UK NEQAS all methods mean, the method/submethod median for Hb A_2 was also compared to the UK NEQAS all methods mean. Using the median of the data sets should prevent outlying results being a problem and is a better way of assessing data sets that are not normally distributed. The UK and the non-UK participants were separated to investigate what effect, if any, separating participants in this way would have on the overall performance of the different groups. The results of six of the most widely used methods/submethods from 2006 to mid-2008 were analysed and again, the BioRad Variant Classic and BioRad Variant II; Beta-thal tended to show a slight positive bias relative to the UK NEQAS all methods mean, whereas the BioRad Variant II; Dual Kit tended to show a slight negative bias. Also, the Tosoh G7 tended to show a fairly strong positive bias for Hb A_2 whereas the Menarini HA 8160 demonstrated a fairly consistent negative bias.

Figure 6 and Figure 7 appear to show that the UK method mean and median for column chromatography and the BioRad Variant Classic are higher relative to the UK NEQAS all methods mean than the non-UK groups. These differences are likely to be due to the relatively large number of non-UK participants using these methods for measuring Hb A_2 and therefore the discrepancies are probably due to the size of group rather than to geographical differences. Figure 8 through to Figure 11 do not appear to show any major differences between the UK and non-UK method mean and median for any of the other methods/submethods.

Standard Deviation

Standard deviation (SD) is used to measure the spread of a data set. The standard deviation is high when the data set contains a large spread of results about the mean and is low when all of the results are close to the mean. Due to changes in methodology and improvements in technology over the last eight years, we predicted that the overall SD of the Hb A_2 results submitted by UK plus non-UK participants would have dropped since 2000.

Following a discussion with the Statistician that was involved with this Hb A_2 project, it was determined that the data sets should not be assumed to be normally distributed. Often, the Hb A_2 data sets contain spurious results submitted by UK NEQAS participants and when analysing the mean of the data, these results can skew the overall mean which makes this an unreliable way of assessing the performance of participant groups. There are several ways of dealing with data that is not normally distributed:

1. Non-parametric data analysis can be performed, e.g. using the Kruskal-Wallis or the Mann-Whitney-Wilcoxon test.
2. The data could be trimmed and then the log of all of the individual results calculated. The mean of the logged results could then be used to assess the performance of the various participant groups. This would involve using the trimmed SD equation which is different to the normal SD equation (Downton's estimate).
3. The median, rather than the mean, of the data sets could be used and the SD could be estimated using percentiles (quartiles).

Due to the limitations of the computer software available for carrying out the data analysis, options 1 and 2 would have been difficult. The third option, which involved using the median of the data sets and estimating SDs based on the interquartile ranges, was therefore proposed as an alternative, and statistically more sound way, of calculating the SD for these types of data. The SDs were therefore calculated using the median and interquartile range.

Figure 12 shows that our prediction was correct in that there has been an overall reduction and therefore improvement in the SD of Hb A_2 measurement since 2000. When the results of the UK and the non-UK participants were separated, Figure 13 shows that the reduction in SD of UK and non-UK participants is similar. The SD of the UK data set is consistently lower than that of the non-UK data set but the improvement in SD is similar for both groups.

Coefficient of Variation

Coefficient of variation (CV) is typically calculated by dividing the SD by the mean and it is useful for comparing the degree of variation between data sets when the means are drastically different from each other (www.investopedia.com/terms/c/coefficientofvariation.asp). Although the means of the UK NEQAS Hb A_2 data are not wildly different to each other, the CVs of the various data sets were calculated in addition to SD so that the results of the two different statistical parameters could be compared.

As explained in the Standard Deviation section above, it was considered important to eliminate any spurious results submitted by UK NEQAS participants which could potentially skew the SD. The way in which this was achieved was by calculating the SD using the median and the interquartile range of the data set. This method was therefore also adopted to look at the changes in CV by calculating the CV by dividing the SD (based on median and interquartile range) by the median of the data set.

It was found that the overall CV of the Hb A_2 results submitted by UK plus non-UK participants followed a very similar pattern to the overall SD in that there has been an overall reduction in the CV of Hb A_2 measurement since 2000. It was also found that the reduction in the CV of Hb A_2 measurement by UK and non-UK participants was similar over the eight year period, although slightly better in the UK group. The CV of the UK data set, however, is consistently lower than that of the non-UK group.

Overall, the CV values generated follow similar trends to the SD values for the Hb A_2 data sets. This is likely to be because the means of the different data sets do not differ substantially from one another.

Result Distribution

Result distribution graphs were generated in order to see the proportion of participants that submitted a particular Hb A_2 result using six of the most common methods/submethods, relative to the overall proportion of participants that submitted that particular Hb A_2 result. They also allow us to see where the results submitted by the larger participant groups lie relative to the UK NEQAS all methods mean for Hb A_2.

Altogether, 27 specimens were sent out to UK NEQAS participants for Hb A_2 measurement between 2006 and mid-2008. The three result distribution graphs shown in Figure 16 though Figure 18 are typical examples of the overall pattern seen with the majority of these surveys. The ‘All methods’ lines have dual peaks which reflect the high proportion of participants which use methods/submethods that tend to show a positive or negative bias, as shown by the results of the six main method/submethod groups that are also marked on the graphs.

The UK NEQAS all methods mean Hb A_2 values for the three specimens used to generate Figure 16 through Figure 18 were actually submitted as the result by far

fewer participants than surrounding Hb A_2 values. This shows that care should be taken when comparing results to the UK NEQAS all methods mean, as this mean is simply the average of the results submitted by participant groups that generally show a positive or negative bias relative to the overall mean and does not necessarily reflect the distribution of the data, particularly if one particular group is larger than the others and skews the overall mean.

Of the six main method/submethod groups, the BioRad Variant II; Dual Kit and the Menarini HA8160 tended to show a negative bias relative to the all methods mean and the BioRad Variant Classic, BioRad Variant II; Beta-thal and Tosoh G7 tended to show a positive bias relative to the all methods mean. The 'Column Chromatography' lines on the result distribution graphs tended to follow a similar pattern to the 'All methods' lines, with some participants submitting relatively high Hb A_2 values and others submitting relatively low results. The respective positive and negative bias of the Tosoh G7 and Menarini HA8160 groups are even more apparent when the specimen contains a borderline or high level of Hb A_2, as shown in Figure 19 and Figure 20. This indicates that certain participant groups, e.g. Menarini HA8160 users, are more likely to categorise specimens with borderline Hb A_2 levels as normal when most other participants would class them as beta thalassaemia carriers when applying the National Screening Programme cut-off for Hb A_2 of 3.5%.

When the result distribution graphs were redrawn to show the results of five of the largest HPLC groups compared to the overall HPLC results submitted, these HPLC result distribution graphs look very similar to the result distribution graphs generated using the results of all methods. This is likely to be due to the fact that HPLC users constitute the majority of scheme participants. The trends in the results submitted by the different participant groups are obviously the same in both types of graph but may be slightly clearer to see when studying the HPLC results in isolation.

Normal Distribution

The graphs in the Result Distribution part of the Results section which were produced from looking at the distribution of Hb A_2 results submitted by participants show some interesting trends, however the lines generated on these graphs are not smooth curves. It was proposed that by generating normal distribution curves of the data, the overall trends within the data would be seen more clearly.

As stated earlier in this report, spurious Hb A_2 results can skew the mean of the data sets which makes the mean, and SDs calculated using the mean, unreliable for comparing the performance of participant groups. The normal distribution curves were therefore plotted using the median of the data sets ±3 SDs based on interquartile range. The results showed that the Menarini HA8160 median is consistently lower than the all methods median, especially when the Hb A_2 level is unusually low or high. The median of the Tosoh G7 data sets are generally similar to the all methods median, except when the Hb A_2 percentage is high, in which case the Tosoh G7 group shows a positive bias relative to other methods and submethods.

It is at the borderline levels of Hb A_2 that these biases can become clinically significant. For example, Figure 26 shows that greater than 50% of Menarini users would be expected to obtain a Hb A_2 value of less than 3.5% on this particular specimen from a beta thalassaemia carrier. The fact that this group may miss a beta thalassaemia carrier individual if the Hb A_2 level is borderline is the most significant finding with this method of analysing the data. The Tosoh G7 group on the other hand, are more likely to obtain a Hb A_2 result of greater than 4.0% and therefore class an individual with a borderline Hb A_2 level but normal red cell indices as a beta thalassaemia carrier due to their positive bias when the Hb A_2 level is raised. This is less concerning but is still significant due to the possibility of causing potentially unnecessary worry for pregnant ladies with a borderline Hb A_2 level.

Cumulative Distribution

The production of cumulative distribution histograms was suggested in order to allow the proportion of participants that submitted a Hb A_2 result above or below a certain level to be seen very clearly, without assuming that the data follow a normal distribution. They also show how the results submitted by six of the largest method/submethod groups differ from one another and from the overall results submitted. As expected, the overall findings were the same as in the Normal Distribution section in that the Menarini HA8160 showed a negative bias relative to other methods of Hb A_2 measurement and the Tosoh G7 showed a positive bias when the Hb A_2 level is raised.

Figure 30 shows at this borderline Hb A_2 level of 3.9%, greater than 50% of Menarini users would be expected to obtain a Hb A_2 value of less than 3.5% on this specimen whereas all participants using either a BioRad Variant II; Beta-thal or a Tosoh G7 gave Hb A_2 results greater than 3.5%. This difference in the results generated by the different HPLC systems is likely to result in discrepancies in the interpretation of the results of specimens with borderline Hb A_2 levels which in turn may lead to inconsistencies in the assessment of beta thalassaemia carrier status and the requirement for further testing.

Borderline Hb A_2 Specimen

Having analysed historic Hb A_2 data gathered by UK NEQAS, a prediction was made about the pattern of results that would be seen for the specimen with a borderline level of Hb A_2 that was arranged to be sent out in 2009. It was predicted that at a borderline Hb A_2 level of 3.5-4%, the Menarini HA8160 participant group would show a negative bias relative to other methods of Hb A_2 measurement and the Tosoh G7 group would show a positive bias. When the results from this specimen were collated, the UK NEQAS all methods mean was calculated to be 3.7%.

The normal distribution curves and cumulative distribution histogram shown in Figure 31 and Figure 32 clearly demonstrate that the prediction was correct. Overall, 21% of participants gave a Hb A_2 result below the 3.5% cut-off for this specimen. Within the Menarini HA8160 group, however, 59% of participants registered as using this analyser gave a result below the cut-off. Within the

Tosoh G7 group, however, 100% of participants gave a result of 3.5% or above. This means that over half of the Menarini users may concluded that the Hb A_2 result on this specimen is normal whereas all of the Tosoh G7 users are likely to have classed the result as abnormal and concluded that the patient is at risk of beta thalassaemia if they are applying the 3.5% cut-off.

When the interpretive comment codes submitted by participants were studied, it was found that 64 of the UK laboratories did not use the beta thalassaemia carrier comment code, although 16/64 did note the possibility of beta thalassaemia carrier status in a free text comment. UK NEQAS reported that of these 64 laboratories, 26 returned an Hb A_2 measurement of 3.4% or less and 22 returned a result of between 3.5 – 3.8%, but neither used the beta thalassaemia carrier code nor mentioned beta thalassaemia carrier status in free text.

It was also interesting to note that 28 of the 64 UK laboratories (44%) that did not use the beta thalassaemia carrier interpretive code were Menarini HA8160 users, despite the fact that only 22% of the UK participants were using this particular instrument. 24 of these 28 participants returned an Hb A_2 measurement of 3.4% or less and the remaining 4 submitted a result of 3.5%. The cut-off for beta thalassaemia set by the National Sickle Cell and Thalassaemia Screening Programme changed from an Hb A_2 result of greater than 3.5% to greater than **and including** 3.5% (with microcytic red cell indices) in late 2009. Since this survey was sent out in early 2009, all 28 of the Menarini HA8160 users failed to produce a Hb A_2 result above the cut-off for beta thalassaemia at that time, despite the fact that the majority of participants obtained a result above the cut-off and that the consensus interpretive comment was that the specimen was from a beta thalassaemia carrier.

A potentially interesting piece of further work would be to analyse the Hb A_2 assessment codes submitted by the different participant groups for this specimen to assess the composition of participants that have classed their result for this specimen as low, normal, high and uncertain. It would also be interesting to study the interpretive comment codes submitted in more detail to investigate the composition of participants that used the ‘no evidence of (beta) thalassaemia’ comments and the comments that were submitted relating to partner testing. These investigations may reflect the different ranges and cut-offs used by participants for the interpretation of Hb A_2 results and would give more information about the proportions of participants that are following the testing algorithms set by the National Sickle Cell and Thalassaemia Screening Programme.

T-Tests

The results of the t-tests suggest that there are, in fact, systematic differences in the results submitted by UK and non-UK participants using the same method/submethod. For this reason, only UK data was used for the following UK Interquartile SDs and Performance Scoring sections.

UK Interquartile SDs

One of the big questions that has arisen during this project is whether participants results should be compared to:

- the overall all methods mean/median
- the mean/median of their particular method group
- the mean/median of their particular submethod group

The standard deviation (SD) of the various participant groups (UK only) was therefore calculated using the median and interquartile range of the data sets for each survey from 2006 to mid-2008. For each specimen, participant's results were analysed in three different ways. They were compared to:

1) The overall median comprising all of the different methods.
2) The median of their method group where at least 20 participants were using that method.
3) The median of their submethod group where at least 20 participants are using that particular analyser.

Those participants whose results fell outside of a ±2SD limit (based on the appropriate median and interquartile range) using each of the different approaches were compared. The composition of participants with results greater than 2SDs above or below the median was found to vary very little between results being compared to the all methods median and to the method specific median. Currently, UK NEQAS compare participants results to the method specific mean where there are at least 20 participants using that particular method. When participants results are compared to the submethod specific median, however, the composition of those whose results fall outside the ±2SD limit changes quite dramatically. Far more BioRad Variant II; Beta-thal and Tosoh G7 users are greater than ±2SDs from the median using this method of analysis, and far less Menarini HA8160 users are greater than ±2SDs from the median.

Overall, the analysis showed that whether participants results are compared to the all methods median or the method specific median, a similar proportion (7%) were found to have results that were outside of the ±2SD range between 2006 and mid-2008. When, however, participants results were compared to the submethod specific median, a much higher proportion (10.5%) of participants were found to have submitted results that were outside of the ±2SD range.

Performance Scoring

As mentioned previously, UK NEQAS currently compare participants results for Hb A_2 to the method specific mean where there are at least 20 participants using that particular method. Where there are less than 20 participants using that method, the results are compared to the all methods mean. In this study, the performance scoring system was analysed in a number of different ways in order to help address the question of whether participant's results should be compared to:

- the overall all methods mean/median
- the mean/median of their particular method group
- the mean/median of their particular submethod group

As would probably be expected, the difference in the participants with analytical performance scores greater than or equal to 100 using each of these three different approaches follows a similar pattern to the difference in the participants with results greater than ±2SDs from the relevant median shown in the UK Interquartile SDs section. The UK NEQAS definition of Persistent Unsatisfactory Performance is having an analytical performance score of greater than or equal to 100. This scoring philosophy allows a long term retrospective measure of performance. On the whole, the number of participants with scores greater than or equal to 100 appears to vary very little between when results are compared to the all methods median and to the method specific median. For the Menarini HA8160 group, however, fewer users have a score greater than or equal to 100 when compared to the all methods median than when compared to the HPLC group median. When participant's results were compared to the submethod specific median, fewer Menarini HA8160 users are identified as having performance scores greater than or equal to 100 but more Tosoh G7 and BioRad Variant II: Beta-thal users have scores of 100 or more.

One of the aims of this study was to try to modify the scoring system so that 5% of participants are identified as persistent unsatisfactory performers and that equal proportions of persistent unsatisfactory performers are identified in each method/submethod group. One of the ways in which this could potentially be achieved would be to change the current definition of Persistent Unsatisfactory Performance (PUP). It was found that when comparing participant's results to the all methods median or the method specific median, a lower PUP score cut-off of greater than or equal to 90 would identify around 5% of participants as above the cut-off. When comparing participants results to the submethod specific median, however, a higher PUP score cut-off of greater than or equal to 110 would be needed to identify around 5% of participants as above the cut-off. An analytical performance score of greater than or equal to 100 is used by UK NEQAS as the definition of Persistent Unsatisfactory Performance in a number of it's different schemes, however, therefore altering this score would mean that the Hb A_2 scoring system is no longer in keeping with others schemes. Also, it may appear unfair to apply different cut-offs to different participant groups in order to obtain equal numbers of persistent unsatisfactory performers within each method/submethod group, plus having different cut-offs for different participant groups does not fit with the NHS Sickle Cell and Thalassaemia Screening Programme's single cut-off for Hb A_2. It was therefore decided that rather than altering the PUP score cut-off, a better approach would be to alter the analytical performance scoring system.

The UK NEQAS scoring system was devised so that the multiplier in the performance scoring equation could be altered in order to obtain the desired proportion of participants with scores greater than or equal to 100. The figure by which the sum of the previous 6 deviation indices is currently multiplied by in the Hb A_2 scoring system is 9. The multiplier that should be used in the analytical performance scoring system in order to obtain around 5% of participants as persistent unsatisfactory performers will vary depending on whether participants results are being compared to the all methods median, method specific median or submethod specific median. Based on the data gathered between 2006 and mid-2008, it has been calculated that, to the nearest whole number, the number

that the previous 6 deviation indices should be multiplied by to obtain around 5% of UK participants as persistent unsatisfactory performers is 10 if comparing participants results to either the all methods or method specific median and 9 if comparing results to the submethod specific median. This approach does not, however, allow us to obtain equal proportions of persistent unsatisfactory performers within each method/submethod group.

Table 1 shows that the multipliers required within the scoring system to obtain around 5% persistent unsatisfactory performers within each method/submethod group is the same whether participants continue to be compared to the method specific median or whether they are compared to the overall, all methods median. If, however, UK NEQAS chose to start comparing participants results to the submethod specific median, then the multipliers required to obtain around 5% persistent unsatisfactory performers within each method/submethod group changes for each of the three submethod groups with greater than or equal to 20 participants. The idea of using different multipliers in the performance score calculation for different participants groups, however, is inconsistent with the application of a single Hb A_2 cut-off for beta thalassaemia carrier status. It may also be considered unfair, and inconsistent with good EQA practice, to use different scoring systems for different participant groups.

The idea of trying to obtain equal proportions of persistent unsatisfactory performers within each method/submethod group also has a major disadvantage in that a group that has very tight SDs and CVs for Hb A_2 measurement would have the same proportion of participants being identified as persistent unsatisfactory performers as a group that has very wide SDs and CVs. Therefore the composition of the participants identified as persistent unsatisfactory performers would not reflect the differences in SD and CV of the different participant groups.

Normal Ranges

Comparison of the normal ranges used by UK and non-UK participants revealed that laboratories are using a wide variety of different reference ranges for Hb A_2. This was not surprising since other work carried out during this project had shown that certain methods/submethods appear to have a positive bias for Hb A_2 and others have a negative bias. What was surprising, however, was the difference in the reference ranges used by participants using the same method/submethod. The ranges used vary greatly in both the UK and elsewhere and there does not appear to be any clear trend in participants using particular methods/submethods generally having higher or lower reference ranges than those who use other methods/submethods. It also appears that some laboratories in the UK are not applying the 3.5% cut-off for a normal Hb A_2 level as set by the NHS Sickle Cell and Thalassaemia Screening Programme, however these laboratories are not necessarily performing antenatal screening and may be located within the UK but outside England where the NHS Sickle Cell and Thalassaemia Screening Programme guidelines are not applicable.

It would be very interesting to establish where laboratories obtain their normal reference ranges from. If laboratories were mostly using the ranges

recommended by instrument manufacturers, we would have seen the same range being used by the majority of participants using the same method or submethod for Hb A_2 measurement, therefore this explanation appears unlikely. If participants were mostly using ranges quoted within particular references, then a number of different references must be being referred to in order to account for the inconsistencies seen. It is likely that some participants have established their own reference ranges by measuring the Hb A_2 within the blood samples of a number of healthy volunteers, however it would be interesting to know how and when these ranges were established. Due to the ethical issues surrounding testing specimens from volunteers and the associated problems of obtaining an adequate number of specimens from a representative pool of donors, establishing reference ranges 'in house' is becoming increasingly difficult. Some participants may therefore be using historic 'in house' derived ranges that were not established using their current method/submethod for Hb A_2 measurement. A combination of the sources of reference ranges described above are likely to account for the variation seen in those used by the different UK NEQAS participants.

Hb A_2 Assessment Codes

For each specimen where an Hb A_2 result is requested, UK NEQAS participants are asked to assess the Hb A_2 results that they generate in terms of their normal reference range (UK NEQAS (H), 2008). This involves ticking one of four boxes labelled low, normal, high or uncertain. The results given by participants for the Hb A_2 assessment were studied and were classed as 'consensus results' if greater than 85% of participants gave that answer. Specimens that had been sent out with a borderline Hb A_2 level which failed to produce a 'consensus result' for the Hb A_2 assessment were therefore excluded from this part of the project. Between 2006 and mid-2008, 20 participants were found to have submitted 3 or more 'outwith consensus' assessment results. Since the ratio of UK to non-UK participants between 2006 and mid-2008 was almost exactly 2:1, you would expect roughly 13 of these 20 labs to be UK based and around 7 to be non-UK labs. It was found, however, that 9 of the labs that had given 3 or more 'outwith consensus' assessment results between 2006 and mid-2008 were UK based and 11 were non-UK labs.

It is difficult to identify the precise cause of the submission of 'outwith consensus' Hb A_2 assessment codes as there are a number of factors that could contribute to such errors. An 'outwith consensus' code could be submitted due to the generation of an 'outwith consensus' Hb A_2 result by the laboratory. It could also be due to the participant using a normal range for Hb A_2 which is different from the majority of other participants, since the ranges used participants vary greatly as discussed in the previous section. Finally, it could either be due to a transcription error in which the participant simply ticked the wrong box by mistake or an interpretation error in which the individual who filled out the UK NEQAS return form incorrectly assessed the result. In order to get a better idea of which factor may be responsible for the 3 or more 'outwith consensus' comments submitted by the 20 labs between 2006 and mid-2008, the Hb A_2 results that they submitted were studied in more detail, along with their normal ranges. This allowed us to hypothesise as to the cause of the 'outwith consensus' comment.

The generation of an ‘outwith consensus’ Hb A_2 result and the use of a different normal range to most other participants are thought to account for a large proportion of the repeatedly ‘outwith consensus’ Hb A_2 assessment codes submitted by UK NEQAS participants. A relatively small proportion of the ‘outwith consensus’ codes were thought to have arisen from transcription or interpretation errors in which the incorrect assessment box was ticked for the result generated. A number of the ‘outwith consensus’ assessment codes were likely to be due to a combinations of factors, most of which were due to both generation of an ‘outwith consensus’ Hb A_2 result and the use of a normal range that differed from most other participants.

The methodology used by the participants that submitted 3 or more ‘outwith consensus’ Hb A_2 interpretation results between 2006 and mid-2008 was then investigated. Although the Menarini HA8160 only accounted for roughly 12.5% of the methodology used by UK NEQAS participants at this time, it was used to generate nearly 30% of the Hb A_2 results submitted by participants giving 3 or more ‘outwith consensus’ Hb A_2 assessment codes from 2006 to mid-2008. This is in-keeping with earlier findings that the Menarini HA8160 has a negative bias relative to the other methods used by participants for Hb A_2 measurement.

Interpretive Comment Codes

When the comments that participants attach to their specimen results were studied, only ‘outwith consensus’ comments that were submitted by participants were identified, rather than investigating 'missing comments', e.g. the lack of the 'beta thalassaemia carrier' comment on specimens with a consensus raised Hb A_2 and the lack of 'partner testing recommended' comments on pregnant beta thalassaemia carriers. The reason for this is because participants will often submit free-text comments in addition to using the standard comment codes. They may therefore have opted to submit a free-text comment that is different to the standard comment code but which essentially makes the same point. As it would not have been possible to retrospectively study the free-text comments submitted by each participant for every survey due to the result sheets no longer being available, only the ‘outwith consensus’ comment codes were investigated.

Between 2006 and mid-2008, 30 participants were found to have used comments which state ‘no evidence of thalassaemia’ or ‘no evidence of beta thalassaemia’ for specimens which were from beta thalassaemia carrier individuals. The reason why these labs had missed the fact that the specimen was from a beta thalassaemia carrier could have resulted from a number of errors, as with the submission of the ‘outwith consensus’ Hb A_2 assessment results. It could have been due to the generation of a falsely low Hb A_2 result. It could have been due to the use of a different normal range to most other labs in which the upper limit for the Hb A_2 reference range was higher than that used by most other participants. Finally, it could either be due to a transcription error in which the participant simply wrote down the wrong code by mistake or an interpretation error in which the individual who filled out the UK NEQAS return form incorrectly interpreted the results. It could also have been due to a combination of two or more of these reasons.

When the causes of these 'outwith consensus' comment codes were hypothesised, the generation of falsely low Hb A_2 results appeared to be responsible for the majority of the participants that submitting comments stating that there was 'no evidence of (beta) thalassaemia' for specimens from beta thalassaemia carrier individuals. The finding that the majority of these participants are Menarini users is consistent with the finding discussed earlier that the Menarini HA8160 has a negative bias for Hb A_2 relative to other methods and submethods.

Of the 30 participants that used codes which suggested that they missed the fact that the specimen was from a beta thalassaemia carrier, 37% were UK labs and 63% were based outside of the UK despite the fact that UK outnumbered non-UK participants by 2:1 during this period.

In addition to investigating comments which implied that beta thalassaemia had been missed by participants, cases where comments inappropriately suggesting that beta thalassaemia was present were also studied further. It was found that between 2006 and mid-2008, 42 participants had used comments incorrectly suggesting beta thalassaemia and 21 (50%) of these specimens were from sickle cell carriers. Of these 21 sickle cell carrier specimens, 11 were actually at risk of coexisting alpha thalassaemia rather than beta thalassaemia. It is difficult to establish why these participants concluded that the specimen was from a patient with sickle plus beta thalassaemia as the Hb A_2 result is not requested for these specimens. For those specimens with microcytic red cell indices and a risk of alpha thalassaemia, participants could have got confused between the indications for coexisting alpha and coexisting beta thalassaemia. For the specimens containing Hb S but with normal red cell indices, it could be that a falsely high Hb S quantitation result lead participants to conclude that there was a risk of coexisting beta thalassaemia. Alternatively, the falsely high Hb A_2 result which is obtained by most methods for Hb S carrier specimens could have lead some participants to incorrectly assume that the specimens were from individuals with both Hb S and beta thalassaemia. More work would need to be done in this area in order to determine the most likely causes of the incorrect use of beta thalassaemia comments for specimens containing Hb S. They could be due to analytical errors or interpretive mistakes by the laboratory but our initial findings imply that there may be an education issue regarding the risk of coexisting alpha and beta thalassaemia in patients with Hb S.

For the 21 participants that incorrectly used the beta thalassaemia comment code in the absence of Hb S, the generation of falsely high Hb A_2 results again appeared to be responsible for why beta thalassaemia was suggested by the majority of these participants. Transcription or result interpretation errors also appeared to be the likely cause of a number of the incorrect beta thalassaemia carrier comments and in one case, it seemed likely that the error was due to a transposition of UK NEQAS specimen results.

Of the 42 labs that incorrectly concluded that specimens were from beta thalassaemia carriers, 43% were UK-based and 57% were non-UK labs which again does not reflect the relative proportions of UK and non-UK participants. One explanation for the submission of relatively fewer 'outwith consensus'

interpretive comment codes by UK compared to non-UK participants is the potential influence of the NHS Sickle Cell and Thalassaemia Screening Programme. Through it's Handbook for Laboratories, the Programme has helped to standardise the interpretation of haemoglobinopathy screen results in antenatal screening laboratories in England.

Initial Conclusions

Since the implementation of the NHS Sickle Cell and Thalassaemia Screening Programme, there has been a reduction in the SD and CV for Hb A_2 measurement by UK NEQAS participants which demonstrates an improvement in the correlation of the Hb A_2 results submitted. This has coincided with a change in methodology from the majority of participants using column chromatography for Hb A_2 measurement to using HPLC, which is likely to be one of the major factors responsible for the improvement seen in the SD and CV values.

A number of different methods for analysing the historic Hb A_2 data gathered by UK NEQAS have shown that the overall mean Hb A_2 value does not reflect the distribution of the data, particularly when participant groups of varying sizes show a positive or negative bias relative to the overall mean. The trends found by analysing the historic data were also evident within the results of the borderline Hb A_2 specimen sent out in 2009 in that the Menarini HA8160 demonstrated a consistent negative bias for Hb A_2 and the Tosoh G7 has a positive bias but only when the Hb A_2 level is borderline or raised.

The Menarini HA8160 is therefore more likely to produce a result below the cut-off for beta thalassaemia when most other analysers would obtain a result above the cut-off and class the patient as a beta thalassaemia carrier. In the latest version of the Handbook for Laboratories issued by the NHS Sickle Cell and Thalassaemia Screening Programme, however, the Hb A_2 cut-off for beta thalassaemia has changed. In the original testing algorithm, a patient had to have a Hb A_2 value of greater than 3.5% and microcytic red cell indices in order to be at risk of beta thalassaemia (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). In the version released in late 2009, the beta thalassaemia risk group had changed to having an Hb A_2 value of greater than **or equal to** 3.5% with microcytic red cell indices (NHS Sickle Cell and Thalassaemia Screening Programme, 2009). This change means that antenatal screening laboratories that use analysers that have a negative bias, such as the Menarini HA8160, should be less likely to miss pregnancies in which the unborm child is at risk of beta thalassaemia major. There is, however, still an increased potential for Menarini HA8160 users to miss the diagnosis of beta thalassaemia carrier individuals when this instrument is used in conjunction with the universal Hb A_2 cut-off set by the NHS Sickle Cell and Thalassaemia Screening Programme.

The Tosoh G7, with it's positive bias for Hb A_2, is more likely to produce a result above the beta thalassaemia cut-off and therefore class a patient as a beta thalassaemia carrier when most other analysers would obtain a result below the cut-off. Although this perhaps has potentially less serious consequences than missing a beta thalassaemia carrier individual, it has the potential to cause unnecessary worry and inappropriate further testing. If a pregnant lady with a

borderline Hb A_2 level was classed as a beta thalassaemia carrier using this analyser, the couple would have to be informed of the potential risk to the unborn child and partner specimens would be requested. If partner specimens were unobtainable, the mother may have to worry throughout the pregnancy about whether or not the baby will be affected until the outcome is revealed upon newborn screening. The NHS Sickle Cell and Thalassaemia Screening Programmes Handbook for Laboratories states that 'If the partner is unavailable for testing or his haemoglobinopathy status is unknown, a risk assessment should be done. The programme supports the woman being offered prenatal diagnosis' (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). A pregnant lady may therefore be offered prenatal diagnosis on the basis of her result alone, although in view of the positive bias of this particular type of analyser, it seems advisable that repeat testing should constitute part of the risk assessment. If the lady's partner was tested and also classed as a beta thalassaemia carrier (which is more likely using this analyser), the couple may again be offered prenatal diagnosis. The Handbook for Laboratories does state, however, that 'Ideally, identification of the β-thalassaemia mutations should be carried out before fetal tissue sampling', which may act as a safety net to catch any individuals incorrectly diagnosed as beta thalassaemia carriers and prevent inappropriate fetal sampling. It will not, however, prevent the unnecessary worry and requesting of partner specimens which is more likely to occur for couples tested in areas using Tosoh G7 analysers for antenatal screening.

When studying the effects of comparing participants results to the all methods median, the method specific median and the submethod specific median, little difference was found between comparing participants results to the all methods median and the method specific median. This is likely to be due to the fact that the HPLC method group constitutes the majority of participants. When participant's results were compared to the submethod specific median, however, very different numbers and proportions of outliers were identified.

There are advantages and disadvantages with moving from a method specific comparison to either an all methods or submethod specific approach. A potential advantage with moving to comparing participants results to the submethod specific mean or median is that participants using an analyser with a bias for Hb A_2 are less likely to be persistently identified as outliers. UK NEQAS have always aimed to try to improve the performance of laboratories which are generating results that are 'outwith consensus', however it does not aim to penalise participants for factors that are out of their control. It may be, for example, that a Menarini HA8160 user's Hb A_2 results are significantly different to either the overall or HPLC trimmed mean or median, however their results may be fairly consistent with the results of other participants using the same analyser. A submethod specific approach would mean that participants would only be identified as outliers if they submitted a result that was 'outwith consensus' compared to participants using the same analyser and therefore participants would no longer be penalised for using an analyser that has a bias which is out of their control. Similarly, using either an all methods or method specific approach, a participant may not be identified as 'outwith consensus' if they are generating results that are vastly different to their analyser group mean but that are close to the overall or method group mean, whereas they would be identified as 'outwith

consensus' using a submethod specific approach. A disadvantage with using a submethod specific comparison, however, is that some submethod groups would have tighter SDs and CVs whereas others would have much wider SDs and CVs but this would not be reflected by the proportions of outliers identified within each analyser group. Participants within analyser groups with higher SDs and CVs would therefore be able to submit much more variable Hb A_2 results without being identified as outliers than better performing groups with lower SDs and CVs.

Since the NHS Sickle Cell and Thalassaemia Screening Programme has implemented a single Hb A_2 cut-off for beta thalassaemia and are reluctant to allow the use of different cut-offs for different method and analyser groups, UK NEQAS may be more likely to change to using the all methods trimmed mean if they are to change their Hb A_2 scoring system at all. Potential disadvantages with this approach include persistently identifying participants as outliers when their particular method or analyser has a bias for Hb A_2, as mentioned above. Although this may seem unfair for those participants as the bias is out of their control, it may in fact have the advantage of encouraging the instrument manufacturers to address the bias problem. Another potential disadvantage with this approach is that the market leader will have a large effect on where the all methods mean lies, so if the number of participants using an analyser with a bias increased dramatically, the all methods mean would shift and participants using other methods and analysers would be more likely to be identified as outliers.

The UK NEQAS Abnormal Haemoglobins Scheme organisers continuously review the performance scoring system and this project is an extension of their review process. Any changes to the scoring system are made in collaboration with UK NEQAS Steering Committee advisors and the National Quality Assurance Advisory Panel (NQAAP) for Haematology. More work will need to be carried out to determine whether greater improvement in the measurement of Hb A_2 is likely to arise from all method or submethod performance scoring systems.

Since it had already been established that certain methods/submethods appear to have a bias for Hb A_2, it was expected that different participant groups may be using different normal reference ranges for Hb A_2. What was not expected, however, was the difference in the reference ranges being used by laboratories that are using the same method or analyser for Hb A_2 measurement. It would be interesting to find out how and when different participants established their reference ranges for Hb A_2.

It was suspected that differences in the reference ranges for Hb A_2 used by participants may account for some of the differences seen in the assessment of Hb A_2 results. When participants are asked to assess the Hb A_2 results that they generate in terms of being low, normal, high or uncertain, the primary reasons for the submission of an 'outwith consensus' Hb A_2 assessment code appear to be due to generation of an 'outwith consensus' Hb A_2 value, application of varying normal ranges between participants or a combination of these two reasons.

Some of the differences in the maximum levels of the normal reference ranges used by participants and difference in the interpretation of Hb A_2 results in the UK

could be explained by the fact that they are either laboratories that do not perform antenatal screening or they fall outside the remit of the English NHS (i.e. are located within Wales, Scotland or Northern Ireland). These participants therefore do not necessarily need to apply the 3.5% cut-off for a normal Hb A_2 level set within the testing algorithm in the 'Standards for the linked Antenatal and Newborn Screening Programme' (NHS Sickle Cell and Thalassaemia Screening Programme, 2006). With the new BCSH guideline released in early 2009, however, stating that a 'national recommended cut-off Hb A_2 of 3.5% or above has been set as the action point for the diagnosis of carriers of β thalassaemia', perhaps more UK laboratories will adopt this Hb A_2 level as the top of their normal range (British Committee for Standards in Haematology, 2009). The introduction of these guidelines may not effect the ranges used or the interpretation of results by non-UK participants, however. It would be very interesting to determine where participants have derived their reference ranges from and to establish why the reference ranges of some UK participants do not reflect the cut-offs recommended within current guidelines. Pathology Harmony is an initiative funded by the Department of Health which is working towards harmonisation in UK pathology laboratories (www.pathologyharmony.co.uk). It was established in January 2007 and may ultimately assist with the standardisation of the Hb A_2 reference ranges used by UK laboratories.

When the 'outwith consensus' comment codes submitted by participants between 2006 and mid-2008 were studied, the primary cause of an incorrect assessment of beta thalassaemia carrier status appeared to be the generation of an 'outwith consensus' Hb A_2 value. There are therefore clear differences in the Hb A_2 values generated by different HPLC analysers and these differences are likely to result in differences in the interpretation of Hb A_2 results which may lead to a clinically incorrect assessment of beta thalassaemia carrier status.

The NHS Sickle Cell and Thalassaemia Screening Programme established the 3.5% cut-off for beta thalassaemia, however the programme is aware that this will not allow the detection of all beta thalassaemia carrier individuals. The original Handbook for Laboratories that was released in 2006 stated that 'a raised Hb A_2 can sometimes just be borderline-raised in the range of 3.3-3.8%. In combination with a β^0 thalassaemia mutation, these mutations result in a disorder ranging in severity from thalassaemia intermedia to thalassaemia major. A cut-off value of 3.5% for the Hb A_2 level will miss some cases of normal Hb A_2 β-thalassaemia' (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). In the revised handbook released in 2009, the Hb A_2 cut-off for a beta thalassaemia carrier had changed from greater than 3.5% to greater than **and including** 3.5% (with microcytic red cell indices) in an effort to positively identify more carriers of beta thalassaemia (NHS Sickle Cell and Thalassaemia Screening Programme, 2009).

There is no 'gold standard' for Hb A_2 measurement therefore it is very difficult to determine the 'true' Hb A_2 level within a specimen. Further research may identify the use of tandem mass spectrometry as useful tool for providing an accurate Hb A_2 measurement. Also, the establishment of a new international standard for Hb A_2 may prove valuable for the calibration and standardisation of different methods of measurement. In the meantime, it is important that all laboratories

are measuring comparable levels of Hb A_2 within specimens so that the results of thalassaemia screening are consistent nationwide. UK NEQAS supports the standardisation of haemoglobinopathy screening through projects such as this and aims to minimise misdiagnosis from avoidable causes.

Further Work

There are a number of different aspects that could be investigated further to benefit this project to evaluate Hb A_2 and related performance data. Firstly, it would be interesting to further investigate the results of the borderline Hb A_2 specimen sent out in 2009 (0902AH1) to look at differences in both the assessment codes and comment codes submitted by different participant groups. It would also be useful to confirm our findings with this specimen by dispatching further very borderline Hb A_2 specimens in later surveys and examining the results of these specimens in similar detail.

UK NEQAS has also been investigating the possibility of having the Hb A_2 level measured on a number of subsequent EQA specimens by an expert laboratory using a variety of methods, including mass spectrometry, column chromatography and HPLC. This may assist with establishing a reference Hb A_2 level for a series of specimens.

It would also be interesting to compare the overall results and findings of participants that perform antenatal screening to those that are not antenatal screening laboratories in the UK. This would allow us to identify the proportion of participants that are antenatal screening laboratories that do not appear to be following different elements of the NHS Sickle Cell and Thalassaemia Screening Programme guidelines (NHS Sickle Cell and Thalassaemia Screening Programme, 2006a). It would also be interesting to look at the responses to the high prevalence and low prevalence questionnaires carried out by the NHS Sickle Cell and Thalassaemia Screening Programme in conjunction with the findings of this project. These questionnaires were conducted for all high and low prevalence antenatal screening laboratories in the UK in 2008. A further questionnaire asking participants to give the source of their Hb A_2 reference range would also be extremely informative.

Finally, UK NEQAS will re-run data analyses to validate any proposals to alter the performance scoring system used for monitoring Hb A_2 measurement to examine the effects that the changes would have on the scores of its participants. These changes could include altering the multiplier used in the performance scoring equation and/or moving from comparing participants results to the method specific trimmed mean to either the all methods mean or submethod specific mean.

Further work of a similar but slightly different nature could be to investigate trends and differences in the measurement and interpretation of other haemoglobin fractions using similar approaches to those used to investigate Hb A_2 measurement within this project. It may be interesting and informative to investigate, for example, Hb F and Hb S quantitation, and perhaps quantitation of other relatively common haemoglobin variants, since the NHS Sickle Cell and

Thalassamia Screening Programme Handbook for Laboratories also specifies ranges and cut-offs for these other haemoglobin fractions (NHS Sickle Cell and Thalassaemia Screening Programme, 2009).

Acknowledgements

The following individuals should be acknowledged for their support and involvement in the project:
Anne Mahon, Mary West and all of the staff at UK NEQAS (H), Watford.
Joan Henthorn, Central Middlesex Hospital, London.
Dr. S. Mitchell Lewis, Hammersmith Hospital, London.
The NHS Sickle Cell and Thalassaemia Screening Programme with special thanks to Allison Streetly, David Worthington and Elizabeth Dormandy.
The UK NEQAS (H) Special Scientific Advisory Group (SSAG).
Neil Porter and the staff in Haematology at the Royal Hallamshire Hospital, Sheffield.
The participants of the UK NEQAS (H) Abnormal Haemoglobins scheme.

References

British Committee for Standards in Haematology (2009) Significant haemoglobinopathies: guidelines for screening and diagnosis.
Hoffbrand, A. V., Pettit, J. E. and Moss, P. A. H. (2001) Genetic disorders of haemoglobin. In: *Essential Haematology*, 4th edn., pp 71-90. Blackwell Science Ltd., Oxford.
NHS Sickle Cell and Thalassaemia Screening Programme (2006) Standards for the linked Antenatal and Newborn Screening Programme.
NHS Sickle Cell and Thalassaemia Screening Programme (2006a) Handbook for Laboratories.
NHS Sickle Cell and Thalassaemia Screening Programme (2009) Handbook for Laboratories.
UK NEQAS (H) (2008) UK NEQAS for General Haematology Participants' Handbook.

Websites used:
http://www.investopedia.com/terms/c/coefficientofvariation.asp

http://sct.screening.nhs.uk/aboutus

http://www.pathologyharmony.co.uk/

Appendices

Comments from the manufacturers

The four major manufacturers noted in this report were sent a pre-release copy of the report in July 2010 for comment. Their responses are reproduced in full, in the order that they were received.

Appendix 1: Tosoh Bioscience
Appendix 2: Helena Biosciences
Appendix 3: Bio-Rad Laboratories
Appendix 4: Menarini Diagnostics

Appendix 1: Response from Tosoh Bioscience

Received by email 20 August 2010

Dear Mrs De la Salle,

Thank you for providing us the opportunity to comment on this interesting report.
We believe this is a very useful and necessary study to allow us to get a better understanding of the situation today regarding the screening for Thalassaemias and Hb-pathies in the UK.

Our comments:

- We believe this study clearly demonstrates the need for an International standard. Both in terms of method as well as in terms of standard material. Today the only available material is the WHO material that was developed in the eighties and with an assigned HbA2 value of 5,3%. We believe standards should be available at different levels: normal level (between 2%HbA2 and 3,5% HbA2), a standard in the borderline area (between 3,6% and 3,9%) and a standard in the "abnormal" level (HbA2 greater than 4%). If this type of material would be available it would be much more evident for manufacturers to correctly align analysers. Today there is no means of validating which analyser is measuring the correct value.
- Since 2008 Tosoh has released the G8 β-Thalassaemia Analysis Mode. This system will give less positive bias than the G7 analyser. We are confident that this will become evident in the near future when more and more data is collected for the G8 in this scheme.
- We at Tosoh believe that it is critical to make a judgement only based on the combination of different data sets. The HbA2 value alone can never be used to diagnose if a patient is a thalassaemia carrier or not. It is criticial to review the HbA2 value, the Red Cell indices as well as the iron levels before a "diagnosis" is made. If all factors are taken into consideration the amount of unnecessary screenings, and patient worry, originating from the positive bias should be very limited.
- We believe it would be best that values are assigned to the UKNEQAS samples using a reference method and not the all method mean. As rightly mentioned in the report, the use of the all method mean could bias the "assigned value" due to the market share of the different analysers.
- We are very supportive of the proposed further studies regarding the reference ranges. We believe it is critical that all users are working with the same reference ranges, but this will only be possible if a reference method and reference material becomes available.

Our Question:

- Do you by any chance know how the cut-off of less than or equal to 3,5% for HbA2 was determined? Which method was used? Which study population?

We at Tosoh are very supportive of these kind of studies and we are willing to improve our tests. However, I believe we first need standardization before we can improve, today we can only align to a single standard in the abnormal range.

If you have any further remarks or questions for Tosoh, please do not hesitate to contact me.

Kindest regards,

Kirsten

Kirsten Van Garsse
Product Manager - HPLC Solutions - EMEA

TOSOH BIOSCIENCE Dedicated to earn your trust.

DIAGNOSTICS BUSINESS UNIT - EUROPE
Direct: +32 (0)13 61 84 44 - Fax: +32 (0)13 66 47 49
www.tosohbioscience.eu

Appendix 2: Response from Helena Biosciences

Received by email 23 August 2010

Dear Barbara,

Further to your email below, I would like to thank you for providing the report to preview.

Helena has no specific response in relation to the report, however I would be grateful if you could notify us when it is formally released.

Helena is currently in final development of a Haemoglobin IEF method for our newly launched automated capillary electrophoresis analyser (V8) and would welcome input from NEQAS with this method. We hope to offer a combined variant and Thalassaemia method for both the UK and international markets and input and assistance from a respected scheme such as NEQAS Haematology would be appreciated.

Best Regards

Stephen

Stephen Bell
Senior Sales Support Scientist
Helena Biosciences

Main Office:+44 (0) 191 4828440

www.helena-biosciences.com

Registered in England: 1796207

Appendix 3: Response from Bio-Rad Laboratories

Received by email 23 September 2010

Dear Barbara,

I will preface these by saying that many, if not all, of my comments here are personal opinion and may not be entirely supported by Bio-Rad Laboratories.
I also rather suspect that I'm not coming up with anything you have not already taken into account.

That accepted, here they are :

In my rather simplistic view the purpose of external QA schemes is to assess the accuracy, precision and reproducibility of methodologies for particular analytes - I tend to describe this to my colleagues as degree of "trueness".
The goal being to, in the first instance observe and report on the degree of trueness across platforms and methodologies and, in the second instance discover the underlying causes for differences and seek to resolve them.
Ultimately meaning that recoveries and clinical interpretations are the same throughout the region (local or global).

In respect of HbA2, in my opinion, fundamental to the accuracy of results is the accuracy of the calibration materials used by any given device.
In the case of HbA2, I do not believe that there is an internationally recognised calibration material - in consequence, not surprisingly, we can and do see different values assigned for same samples.

Looking specifically at the Bio-Rad platforms, as a general consideration, the values assignment is accomplished by means of (I believe) NIBSC standards and normal patient pools.
In view of the lack of an international calibration material it is perhaps as good as one may hope for. The process used by other companies will, no doubt, vary - contributing to the variability in recoveries.

Importantly, if we look at the calibration value assignment by platform (D-10, Variant Classic, Variant II) and by assay (b-Thal and Dual) we see differences.
So, when running calibrators on the different platforms we get different "response factors" which are subsequently applied to quality controls and samples.
These differences are part of the process by which same samples are assigned different values, as so clearly demonstrated by your data.
This matter is recognised by the manufacturer and is being addressed. I would anticipate that in the near future each of the Bio-Rad platforms and assays will produce similar results (allowing for normal acceptable variations). This process would then give rise to a harmonisation of the collective Bio-Rad peer group.

What it will not address, of course, is:

a) The differences in the frequency with which customers calibrate their instruments (we are aware that not all customers follow manufacturers guidelines) and the impact on correction factors.
b) The regularity and "completeness" of operator maintenance - which also has some impact on calibrator response factors.

Both of which play a part in instrument to instrument variability, albeit within a hopefully acceptable tolerance range.

Looking at manufacturer to manufacturer differences in recovery, no surprises that in the absence of a single calibration reference point, we can and do see different recoveries for same

samples - once again clearly demonstrated by your data - leading to the positive and negative bias' when compared to an all method mean.

In respect of using an all method mean I readily accept that, in the absence of an international reference material it is probably as good as it gets. However the all method mean does not really tell us what the true HbA2 value is for any given sample - it just says what it is - an all method mean. Accepting that different methods give positive and negative bias' to the all method mean and each other we still cannot say which is closest to trueness, so , as you may guess I am a little uncomfortable with it. This is further compounded by the fact that the NHS screening guidelines normal range is fixed - my question is how was this number arrived at (what method) and is it truly applicable in the current environment when we know that different manufacturer methods / platforms give different recoveries for the same sample. Could it be agreed that a particular instrument / method becomes the "Gold standard" as a temporary measure.

I absolutely agree that manufacturers should be under pressure to harmonise calibration procedures, which I think is essential, but it needs to be undertaken in cooperation with guidelines / directives from an agreed "steering committee" - I cannot envisage laboratory workers accepting calibration materials derived by the manufacturers in isolation. Rather like the recently implemented IFCC standardisation for HbA1c, this process would take time.

In the interim, and in view of the fact that different instruments clearly have differing recoveries, perhaps labs should be able to use different normal ranges by manufacturer to harmonise clinical interpretations as a temporary bridging strategy until a common calibrant is available and in use. Alternatively, what about instrument specific correction factors to bring about a temporary harmonisation ?

My final comment is in respect of borderline results - I feel that whilst there is no "universal" calibrator, instruments will continue to have positive and negative bias' to each other and the all method mean.
This in turn, I believe, means that a borderline may be low / normal by one method and elevated by another - meaning some methods will give false negative and another false positive - if a suitable instrument "correction factor" was applied this could to some extent be resolved.

I hope the above is of some use and I would be very happy to discuss further if it would be helpful.

Best Regards

Jon

Jonathan. D. Strotton,
Northern European Product Manager,
Diabetes & Haemoglobinopathies
Bio-Rad Laboratories,
Bio-Rad House,
Maxted Road,
Hemel Hempstead,
Herts, HP2 7DX
UK
Office: +44 208 328 2264

Appendix 4: Response from Menarini Diagnostics
Received by email 5 October 2010

5th October 2010

Mrs B. De la Salle
UK NEQAS (H)
PO Box 14
Watford
Herts
WD18 0FJ

Dear Barbara

Many thanks for your report on the evaluation of HbA2 and related performance data. We wholeheartedly agree with the stated aims of the study; most notably "… to establish more evidence to evaluate the 3.5% HbA2 cut-off for beta thalassaemia carrier status."

Bearing in mind this aim we were quite disappointed at the wholly partial statement about the Menarini HA-8160 found in paragraph 4 of the executive summary. It is true that the Menarini HA-8160 has a negative bias **against the all method mean** but this bias is **relative to the group**, **not absolute**. This is not made at all clear in the Executive Summary and the wording used will clearly lead the reader to believe that any differences are absolute and thus indicate deficiencies on the part of the HA-8160. We should be clear that this is absolutely not because the HA-8160 is performing incorrectly or giving falsely low results. It is because a falsely high 'universal' cut-off has been set using a group mean which includes a large number of uncalibrated systems which have an absolute positive bias and have, therefore skewed the cut-off figure.

When we started to promote the HA-8160 in the UK, the predominant HPLC system at that time was the Bio-Rad Variant and we found very early on that our results for HbA2 were lower than that system by around 0.2-0.3% at the cut-off. For this reason, we had many communications with customers who made the same assumption that you have; that the HA-8160 was giving incorrectly low results. On many occasions the laboratories concerned purchased the WHO International Reference Reagent for Haemoglobin A2 (NIBSC Code 89/666) and tested it on the HA-8160. Without exception, they found that the HA-8160 was always giving results in line with the expected WHO reference values. As you are aware, a number of those customers have gone on to use the WHO International Reference Reagent for Haemoglobin A2 to calibrate the HA-8160. You are further aware that results from the laboratories using this material show no difference with the laboratories using the Menarini calibrator. I should probably add that this is hardly a surprise in that the Menarini-supplied calibration material, manufactured by Eurotrol is traceable back to the WHO standard because the system used to assign the values in the Euotrol calibrator is actually calibrated with the primary WHO standard. For your information, I have enclosed a Eurotrol A2 calibrator pack insert where you will see that it clearly states that the calibration is traceable to the internationally recognised HbA2 reference material from the WHO.

In summary, we fully accept that the HA-8160 is giving results which are lower than the Bio-Rad and Tosoh systems but **we are sure that the HA-8160 is giving the correct results.** Past experience with the same competitor systems when used for measuring HbA1c, and many years of EQA data for that analyte have shown us that the analytical separation and integration performance of the HA-8160 is unsurpassed by anything on the market and we are very confident that the system is giving values closer to the absolute HbA2 values than any other system. The HA-8160 is directly traceable to the WHO standard for HbA2 and we are confident that our calibration and traceability mechanisms are sound. If there are differences between the HA-8160 and other systems, which in some cases are not even calibrated, or calibrated with an internal standard with no external reference system, **the problem clearly lies with the other systems.** In the absence of any evidence of WHO traceability in other systems, we believe that it is clear that it is the results from those systems which should be questioned, not those from the HA-8160. We therefore believe that the thrust of the executive summary is entirely wrong. The HA-8160 can be proved to be giving correct results, so it is clear that the other systems mentioned in the report all have a positive bias against the HA-8160 which is traceable back to the WHO standard. We find it strange that you have not mentioned calibration, traceability or absolute accuracy instead have concentrated on performance relative to the mean which we know has results from a large number of uncalibrated systems and is, therefore, falsely high. This approach is bound to 'lock in' poor accuracy to the system which is clearly not your aim so we really don't understand why you have chosen this relative, rather than absolute approach in your methodology.

It seems that some of the problems which you have documented are related to the desire to use a fixed 'universal' cut-off of 3.5%, no matter which system is used. Although the HA-8160 is traceable back to the WHO standard, we are aware that many other systems are not. In this situation we would assert that trying to use a fixed cut-off is bound to fail due to the well-known calibration differences between systems and your report has simply documented a fact which has been known by manufacturers for many years. For example, within the body of the report you refer to the possibility that blood from a beta thalassaemia carrier with a HbA2 level close to the cut-off of 3.5% is more likely to be 'mis-reported' as negative on the HA-8160. We believe that this is a fundamentally incorrect statement because the HA-8160 is not performing incorrectly or giving falsely low results. The reason for this apparent 'problem' is the attempt to set a 'universal' **relative** cut-off based on a group mean which has been distorted upwards by a large number of uncalibrated systems which have an **absolute** positive bias. The possibility of mis-classification is not because of the performance of the HA-8160, it exists because the NHS Sickle Cell and Thalassaemia Screening programme is recommending a 'universal' cut-off which has been distorted upwards by uncalibrated systems and is thus, unfortunately, unsuitable for the purpose it is intended for.

It is clear from your report that you wish to improve the commutability of HbA2 results across all UK laboratories and thereby improve consistency of reporting patient samples as either positive or negative for Thalassaemia. In order to achieve this we have a simple suggestion. I probably do not need to remind you that it is a legal requirement of manufacturers of CE-IVD marked devices to ensure that systems should be calibrated to the highest order calibration material available. In this case that material is the WHO International Reference Reagent for Haemoglobin A2 and it is clear that some manufacturers are not meeting their obligations under CE-IVD with regard to A2

calibration traceability. We would suggest that you remind manufacturers that they have an obligation in this regard and you could also suggest to UK laboratories that they enquire of their suppliers whether their system is calibrated against, and traceable to the WHO International Reference Reagent for Haemoglobin A2. This would pressurise manufacturers who do not currently have a traceable calibration system to meet their legal requirements under CE-IVD.

We note and welcome your suggestion to send future EQA samples to an 'expert laboratory' in order to derive absolute values for HbA2. We have always believed that having absolute target values, traceable back to the highest order calibration possible is essential for any EQA system and we look forward to seeing the performance of all of the systems displayed in absolute rather than relative terms. If you decide to move in this direction, we would suggest that the analytical method(s) and calibration traceability path(s) for the 'expert' laboratory should be made clear to all participants and manufacturers.

Thank you again for all of your efforts and the opportunity to respond to the report. We would be very happy to work with UK NEQAS and the NHS Sickle Cell and Thalassaemia Screening programme to help resolve the issues around the cut-off value for HbA2 and thus improve diagnosis of Thalassaemia in the UK.

Yours sincerely

Paul Tolan
Managing Director