UK NEQAS

Leucocyte Immunophenotyping

Title:	guide to z scores
Index:	Quality Management 333
Authors:	Liam Whitby
Version:	1.2
Authorised By:	Liam Whitby
Authorisation Date:	24-Jul-2014
Review Date:	25-Jul-2016
Location Of Copy:	unknown

Changes to UK NEQAS Leucocyte Immunophenotyping Performance Monitoring Systems From April 2014

Contents

1. Introduction

UK NEQAS for Leucocyte Immunophenotyping (UKNEQAS LI) currently operates 19 different External Quality Assessment (EQA) programmes covering various aspects of flow cytometry and molecular haemato-oncology. To examine laboratory performance there are 10 different performance monitoring systems in use by this centre, each based on different analytes, statistical methodologies and use different criteria to classify unsatisfactory performance.

An external review of the performance monitoring systems was performed by the Statistical Services Unit of the University of Sheffield in 2012. It was determined that whilst the systems were distinctive from each other they were fit for purpose. It also concluded that that laboratories identified as unsatisfactory were genuinely producing results outside of the expected consensus values.

In order to ensure accreditation to ISO 17043, provide better information on test performance to participants, and provide greater transparency on the derivation of performance monitoring scores, an internal review was held to identify any improvements that could be made to the systems currently in place. The focus for the review was on the quantitative programmes operated by this centre; owing to the availability of International Standards for use in proficiency testing (ISO 13528) that detailed procedures for use in quantitative EQA programmes.

The performance monitoring systems examined under the review were for the following UKNEQAS LI programmes:

Flow Cytometry Programmes

Immune Monitoring; CD34+ Stem Cell Enumeration; Low Level Leucocyte Enumeration; Leukaemia Immunophenotyping (Part 1); Minimal Residual Disease

Molecular Haemato-Oncology Programmes

BCR-ABL Quantitation; Post-SCT Chimerism Monitoring

2. Current Systems

Each of the quantitative programmes listed above used different statistical methods to identify the consensus value for samples and different performance criteria to classify unsatisfactory performance and persistent unsatisfactory performance. A brief outline of the methods used is provided in Table 1.

3. Need for Change

Table 1 shows that the criteria for identifying consensus values and the methods used to identify unsatisfactory performance vary across the programmes. As such it was concluded that this was not transparent to participants, could cause confusion and did not provide information on bias of results or methodologies. Despite the successful external review of the statistics currently in use, it was determined that the scoring systems should be updated to ensure they provide maximum benefit to participants.

4. New Format

An internal exercise was undertaken to study alternative performance monitoring systems that could be implemented but still ensure we were ISO 17043 compliant. As part of this exercise ISO 13528 was used, as it is the sole internationally accepted standard for the interpretation of EQA data. As a result, and taking into consideration the procedures and type of data generated by UK NEQAS LI programmes it was determined to introduce z-scores (as per ISO 13528) and update all quantitative programmes accordingly.

z-Scores

A z-score is calculated using the following formula:

$$z = (x - X)/\hat{\sigma}$$

where x is the result returned by the testing laboratory, X is the assigned value (robust mean) and $\hat{\sigma}$ is the standard deviation for proficiency assessment (robust SD).

The robust mean and robust SD are derived from participant data using Algorithm A (ISO 5725-5) that ensures that all data is included in the generation of the robust mean and robust SD but also minimizes the effect of outliers upon the final values.

Interpretation of z-scores is as follows:

- A result between 2.0 and -2.0 would be classed as satisfactory

- A result between 3.0 and 2.0 or -2.0 and -3.0 is seen as an 'action' result, that highlights a potential issue to the laboratory. Two 'action' results in successive rounds would result in classification as a 'critical'
- A result above 3.0 or below -3.0 is considered to be a 'critical' result requiring immediate investigation by the laboratory

Due to the nature of how z-scores are generated a positive z-score highlights a positive bias in a laboratory's results whereas a negative z-score shows a negative bias. As such, this adds value to the performance monitoring information provided to laboratories because the z-score immediately highlights to the participating centre if their result is above or below the expected consensus value. In addition to the z-score all methodological data featured on reports will be in the format of robust mean and robust SD. This will give participants the option to use the extra provided data to calculate additional "in-house" z-scores based on machine types, methodologies etc and allow them to monitor if there are any "in-house" technical biases. **However, it is important to stress that the z-score issued by UK NEQAS based on all methods will remain the only parameter that is used for performance monitoring.**

Please note that the use of z-scores results in the generation of unique performance scores on a 'per sample' basis, currently performance is monitored on a 'per trial' basis, as such the introduction of z-scores will increase the sensitivity of the performance monitoring systems.

To highlight persistent performance issues and minimise the time to bring these to the participants attention, the definitions of persistent unsatisfactory performance have been updated. In the future persistent unsatisfactory performance will be defined as obtaining a set number (or greater) of critical z-score classifications over a 12 months rolling period, with the number of critical z-score classifications varying from programme to programme and related to the number of samples issued in each EQA send out.

During the course of the review, it was found that the scoring system in place for the leukaemia immunophenotyping programme did not match clinical practice. The current scoring system is quantitative and is based on the percentage expression of individual antigens. However leukaemia diagnosis is based on an overall phenotype. As such, it has been decided that, whilst the leukaemia reports will feature z-scores on individual antigens, robust means and robust SDs on associated tables, the focus of the scoring system will be on the Overall Grade classification. This is based on the overall consensus phenotype of the core antigen and therefore, this programme will be scored in a qualitative manner.

Details of the updated scoring systems can be found in table 2.

To assist participants in the interpretation of the new reports and the data they contain, a series of example reports featuring interpretative notes will be available for download from the UK NEQAS LI website (www.ukneqasli.co.uk). In addition, UK NEQAS LI scientific staff will be available to answer specific questions on reports via email or telephone.

Qualitative programmes

The review focussed, in the first instance, on the quantitative programmes operated by this centre. However it is intended that the performance monitoring systems in place for the qualitative programmes will be reviewed separately in 2014/2105.

Table 1. Outline of statistical methods, scoring systems and performance criteria **currently** used in UK NEQAS LI Quantitative EQA programmes

Programme	Performance Monitored Results	Statistical method used to identify consensus	Scoring system	Definition of unsatisfactory performance	Definition of persistent unsatisfactory performance
Immune Monitoring	Absolute and percentage values for CD3+; CD3+/CD4+ and CD3+/CD8+ lymphocytes	Trimmed mean as defined by Healy[1]	Within +/-1SD of Mean=2 points Within +1/+2SD or -1/-2SD=1 point Outside +2SD or -2SD=0 points	Running score per antigen over 6 samples of <5 points	3 occurrences of unsatisfactory performance in 12 months
CD34+ Stem Cell Enumeration	Absolute values for CD34stem cells	Median Value	≥25th to ≤75th centile=0 points ≥10th to <25th or ≤90th to >75th centile=20 points ≥5th to <10th or ≤95th to >90th centile=35 points <5th or >95th centile=50 points	Running score over 3 samples of ≥100 points	3 occurrences of unsatisfactory performance in 12 months
Low Level Leucocyte Enumeration	Absolute leucocyte count values for red cell and platelet samples	Median Value	≥25th to ≤75th centile=0 points ≥10th to <25th or ≤90th to >75th centile=5 points ≥5th to <10th or ≤95th to >90th centile=15 points <5th or >95th centile=25 points	Running score over 6 samples of ≥100 points on absolute counts	3 occurrences of unsatisfactory performance in 12 months
Minimal Residual Disease	Percentage MRD cells within total population	Median Value	No scoring system in place	No scoring system in place	No scoring system in place
BCR-ABL Quantitation		Median Value of logarithmic reduction between 2 samples	<5th or >95th centile=Amber status	Amber status	2 occurrences of unsatisfactory performance in 12 months
Post-SCT Chimerism Monitoring		Three step process: 1.Results of 100% and 0% removed 2. Upper and lower adjacent points calculated and outliers trimmed 3. Median calculated using remaining data	<5th or >95th centile=Amber status	Amber status	2 occurrences of unsatisfactory performance in 12 months
Leukaemia Immunophenotyping (Part 1)	Percentage expression of core antigens (CD2, CD3, CD5, CD13, CD19 and CD20) on the malignant cell population	Median Value	Incorrect positive/negative classification of antigen based on BCSH guideline[2] cut-off values=50 points. In addition for positive antigens: ≥25th to ≤75th centile = 0 points ≥10th to <25th or ≤90th to >75th centile = 10 points ≥5th to <10th or ≤95th to >90th centile = 20 points <5th or >95th centile=40 points	Running score per antigen over 3 samples of ≥100	3 occurrences of unsatisfactory performance in 12 months

Table 2. Outline of statistical methods, scoring systems and performance criteria to be introduced for UK NEQAS LI Quantitative EQA programmes

Programme	Performance Monitored Results	Statistical method used to identify consensus	Scoring system	Definition of unsatisfactory performance	Definition of persistent unsatisfactory performance
Immune Monitoring	Absolute and percentage values for CD3+; CD3+ CD4+ and CD3+ CD8+ lymphocytes	Robust mean and robust SD derived from participant data using Algorithm A (ISO 5725-5)	z-scores: • 3.0 or < -3.0 = 'critical' • ≤3.0 and >2.0 or ≥-3.0 and <-2.0 = 'action' result. (Two 'action' results in successive rounds would result in classification as a 'critical') • ≤2.0 and ≥-2.0 = satisfactory	One critical z-score Or 2 'action' z-scores in successive rounds	3 occurrences of unsatisfactory performance in 12 months
CD34+ Stem Cell Enumeration	Absolute values for CD34 stem cells				3 occurrences of unsatisfactory performance in 12 months
Low Level Leucocyte Enumeration	Absolute leucocyte count values for red cell and platelet samples				4 occurrences of unsatisfactory performance in 12 months
Minimal Residual Disease	Percentage MRD cells within total population				2 occurrences of unsatisfactory performance in 12 months
BCR-ABL Quantitation					2 occurrences of unsatisfactory performance in 12 months
Post-SCT Chimerism Monitoring					3 occurrences of unsatisfactory performance in 12 months
Leukaemia Immunophenotyping (Part 1)	Overall immunophenotype grade based on core antigen (CD2, CD3, CD5, CD13, CD19 and CD20) expression. Quantitative data on antigens will be transformed and scored qualitatively.	Consensus positive/negative expression on 6 core antigens	Grade A = 6 core antigens in consensus Grade B = 5 core antigens in consensus Grade C = 4 core antigens in consensus Grade D = 3 core antigens in consensus Grade E = 2 core antigens in consensus Grade F = 1 core antigens in consensus Grade G = 0 core antigens in consensus	Grade D or below	2 occurrences of unsatisfactory performance in 12 months