HOKLAS Supplementary Criteria No. 38

'Medical Testing' Test Category – Performance Verification of Automated Analysers

1. INTRODUCTION

1.1 ISO 15189 requires all equipment to be shown, upon installation and in routine use, to be capable of achieving the performance required, and shall comply with specifications relevant to the examinations concerned; and that all methods have to be verified or validated as extensive as are necessary to confirm that they are suitable for the intended use. Automated analysers (autoanalysers) play an important role in the operation of a medical laboratory, especially in the disciplines of chemical pathology and haematology; and to a lesser extent, clinical microbiology and infection. Since the performance of the autoanalysers had been validated by the manufacturer, laboratories are only expected to carry out verification of their autoanalysers before they are put into service if the methodology and reagents used are in strict accordance with the manufacturer's instructions. If the methodology or reagents used in an autoanalyser deviate from the manufacturer's instructions, performance validation shall be conducted as for in-house developed methods.

1.2 The purpose of the verification is to determine whether (i) the total analytical error of the autoanalyser for an analyte is within the total allowable error; (ii) results for an analyte returned from different autoanalysers in the laboratory are comparable; and (iii) the new autoanalyser is suitable for its intended use. Total analytical error consists of three principal contributors: imprecision, method-specific bias, and sample-specific bias, i.e. interference.[1] Performance verification invariably includes studies on imprecision and trueness (method-specific bias) while other factors contributing to the analytical error are studied when there is clinical relevance.

1.3 There are a number of documents published by the Clinical and Laboratory Standards

Institute (CLSI, formerly known as NCCLS) on various studies required for the evaluation of autoanalysers. Laboratories are advised to study these documents for procedural details on how each study is to be conducted and how results are to be analysed. This supplementary criteria document provides an outline of the type of studies usually expected in verifying the performance of an autoanalyser. It is intended to be used by laboratories to verify the performance of an autoanalyser against manufacturer's claims and to verify if the autoanalyser is suitable for its intended use. Validation of autoanalysers by manufacturer is more intensive and is not covered in this document. This supplementary criteria document is applicable only for autoanalysers that report quantitative results on a continuous scale, excluding grading and ranking. For verification of examinations that include a measurement step but do not report a measured quantity value, some studies mentioned in this document may be applicable and can be referred.

1.4 Laboratories should note that reference intervals to be used for interpreting examination results obtained by an autoanalyser also have to be verified before the new autoanalyser is put into service. Please refer to HOKLAS SC-32 for verification of biological reference intervals from other sources.

1.5 Verification of autoanalyser performance is a complicated process, detailed planning before proceeding with the work is recommended. Consideration includes the samples to be used, effects of sample matrix, the timing of test runs, methods for result analyses, acceptance criteria, etc. With careful planning, different studies can be conducted concurrently and different analyses can be done using the same set of data.

2. CIRCUMSTANCES WHERE VERIFICATION OF AUTOANALYSER PERFORMANCE IS REQUIRED

The autoanalyser(s) is(are) to be verified in the following situations:

a. Installation of new autoanalyser(s) irrespective of whether autoanalysers of the same model have been installed;

b. Replacement of an autoanalyser; and

c. Transfer of examination(s) originally done by an autoanalyser to another autoanalyser, regardless of whether the autoanalysers are of the same model.

3. GENERAL APPROACHES

3.1 For laboratories using autoanalysers, a standard operating procedure (SOP) which lays out the general verification protocol for autoanalysers shall be prepared. This SOP shall include the general study protocol, sample types, expected sample size, statistical methods for data analyses and guidelines on selection of acceptance criteria and result interpretation for each type of study. As the studies to be conducted depend on the analyte and its application, the SOP shall also specify the type of verification studies to be performed for each analyte and application. The verification should cover all the analytes and applications that will be used in the new autoanalyser, including the sample types that will be applied in the examination.

3.2 Basic studies required in the verification of autoanalysers are listed below. Similar studies are also applicable for method validation/verification in other testing areas and the list below can be used for reference. This list is not exhaustive and laboratories shall consider including other relevant studies where required. If anyone of the listed studies is considered not applicable for certain analytes, or for certain applications, the laboratory shall document the justifications in the verification report. When using software packages for statistical analyses, the parameters to be obtained using such softwares shall be defined.

3.3 Precision

3.3.1 Precision is not typically represented as a numerical value but is expressed quantitatively in terms of imprecision. Imprecision is defined as the random dispersion of a set of replicate measurements and/or values expressed quantitatively by a statistic, such as standard deviation (SD) or coefficient of variation expressed as a percentage (%CV).[2]

3.3.2 The samples used shall preferably be of the same or similar matrix as the patient samples. A minimum of two analyte levels shall be included in the study. The levels selected shall include the medical decision level(s) or the lower and upper limits of the reference intervals. Because the samples have to be repeatedly tested for many times on different days, they must be stable and have sufficient quantity.

3.3.3 For experimental procedure design, as a minimum, five replicates of each analyte level are to be run per day for five different days (the days need not be consecutive).[2] If only fewer replicates can be tested on each run, more runs should be performed. For example, four replicates can be tested on each run on each of seven days in order to obtain a somewhat more reliable within-laboratory imprecision estimate, without compromising the repeatability estimate. The repeatability and within-laboratory imprecision should be calculated using appropriate statistical tools, e.g. one-way ANOVA.[2]

3.3.4 Alternatively, analyse a minimum of 20 samples of each material within a run or within a day to obtain an estimate of short-term imprecision (repeatability).[3] Performing a single run of each analyte level per day for a total of 20 days to assess between-day imprecision (long-term imprecision).[3] The 20 days need not be consecutive. A study conducted on 20 different days is expected to provide a more realistic estimate of the variation that will be seen in patient samples over time.

3.3.5 Other experimental designs following internationally recognised guidelines can also be used to estimate the repeatability and within-laboratory imprecision / between-day imprecision.

3.3.6 The precision goal shall be stated as the maximum allowable SD and/or %CV at each analyte concentration to be examined. The laboratory shall determine the acceptance criteria for precision taking into consideration the bias of the new autoanalyser so that the total allowable error, i.e. the maximum permissible difference between an individual sample's result and the true value for that sample, is within the limits of performance standard specified by the laboratory based on accepted practice in the field.

3.4 Trueness

3.4.1 Trueness of measurement is defined as the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value.[2] The measure of trueness is usually expressed in terms of bias which is defined as the estimate of a systematic measurement error. If the comparison method is a reference method, then the difference between the two methods measures the bias of the new method. If the comparison method is not a reference method, then the bias of the new method cannot be determined. In this case, the difference between the two methods is considered simply a difference, and not bias.[4]

3.4.2 Trueness may be assessed by analysing materials with known concentration and comparing the results of the new autoanalyser to the expected reference value.[2] Some sources of testing materials with known concentrations, such as reference standards recognised by a professional body, materials prepared by spiking of reference standards, survey materials from PT / EQA programmes and materials used in interlaboratory QC programmes having peer group means (obtained from a minimum of ten participants) may be used for trueness studies.[2]

3.4.3 At least two concentrations that span but do not exceed the analytical measuring interval of the autoanalyser should be used. Selection of an appropriate number of samples with concentrations at or near the medical decision level is recommended.

3.4.4 Trueness may also be assessed by comparing the performance of a new autoanalyser to a reference method but this is usually not practical in a routine laboratory and is not expected.

3.4.5 The bias of the autoanalyser for each analyte shall not be greater than the maximum allowable bias at each concentration to be examined. As stated in clause 3.3.6, the laboratory shall determine the maximum allowable bias, taking into consideration the precision of the new autoanalyser so that the total allowable error is within the limits of performance standards specified by the laboratory. The maximum allowable bias may be expressed as either an absolute concentration or as a percentage of the concentration being studied.[2]

3.5 Inter-instrument comparison

3.5.1 The performance of a new autoanalyser shall be compared with that of the existing autoanalyser whether the existing autoanalyser is to be replaced or be used continually in service. This is to allow monitoring of patient progress when historic and current results are compared.

3.5.2 Patient samples with analyte concentrations that cover the whole analytical measuring interval or entire clinical relevant range shall be used. For comparing the performance of the existing and the new autoanalyser, several different patient samples can be run without replicate on both autoanalysers each day for 3 – 5 days.[4] The days need not be consecutive. Wherever practicable, a minimum of 40 samples shall be tested. Performing measurement on different days allows averaging of any between-day

variability, which may exist for either autoanalysers under comparison.

3.5.3 Having completed the collection of data for analysis, the next step is the visual review of the data. Such a review is useful to get an initial understanding of the difference between the measurement of the existing and the new autoanalysers. A scatter plot of the results obtained from the two autoanalysers using identical scales and ranges for the x and y axes should be constructed and fitted with a linear regression line. The linear regression model (least squares, weighted least squares, Deming or Passing-Bablok) used shall suit the quality of the data collected.

3.5.4 For analysis of data sets obtained from two autoanalysers, use of Deming or Passing-Bablok regression is recommended as both sets of data have random errors. Simple least squares linear regression assumes that only results of new autoanalyser are measured with random error. It should only be used for estimation of difference between two autoanalysers when the correlation coefficient (r) is sufficiently great, i.e. $r \geq 0.975$ (or, equivalently, if $r^2 \geq 0.95$). [Note: The correlation coefficient (r) is often squared to provide the coefficient of determination (r^2)]. In cases of constant difference variability, the estimate of r^2 should be used only as an indicator of the strength of a regression fit and should not be used as the only acceptance criterion of the correlation study.[4] The regression equation obtained from any of the appropriate regression models can be used to estimate the difference of the results between the old and new autoanalysers at any specified value within the measurement interval.[4]

3.5.5 Alternatively, a Bland-Altman plot (difference plot) should be constructed to compare the results and obtain the difference between the results of the two autoanalysers.[4] More samples will increase the confidence of the statistical estimate of the difference.

3.5.6 The laboratory has to specify acceptance criteria on the allowable difference between results from the two autoanalysers and provide justification. Examples of appropriate acceptance criteria including but not limited to: (i) clinical studies; (ii) data from external quality assurance programmes; and (iii) biological variability data.[5] The laboratory shall take necessary action if the results were found to be incomparable according to laboratory's defined criteria.

3.5.7 When more than one autoanalyser is used in the same laboratory to perform an examination, results returned from them have to be comparable so that the autoanalysers can be used interchangeably in routine operation. Therefore, inter-instrument comparison has to be conducted periodically thereafter to ensure that their differences are within acceptable limits on a continual basis.

3.5.7.1 Usually at least 40 patient samples covering the whole analytical measuring interval or the entire clinical relevant range shall be used for comparing performance of different autoanalysers over a period of time. Only concentrations within the reportable range need to be included in the study.

3.5.7.2 Inter-instrument comparisons can be conducted based on frequent monitoring (daily, weekly), or periodic monitoring (quarterly, half-yearly) depending on the stability of the measurement systems.[5] Frequent monitoring generally involves comparing fewer samples while periodic monitoring should be designed to have greater power to detect a difference (i.e. a larger number of samples are tested).[5] The data collected over a period of time are to be analysed statistically by regression and bias determination. When comparing more than two autoanalysers, the results may be analysed by ANOVA or other appropriate statistical techniques

(e.g. range test as specified in CLSI EP31-A-IR). Data collected based on frequent monitoring (a few samples per day or week) shall also be analysed statistically every quarterly or half-yearly. When frequent monitoring is implemented, the laboratory shall also define acceptance criteria for the allowable difference between individual pairs of results.

3.5.7.3 The selection of samples for inter-instrument comparison should take into account the commutability of the materials. The optimal samples used for comparison are native patient samples collected appropriately and processed and stored according to the stability requirements of the analytes. Reference materials, control materials and EQA materials that have demonstrated to be commutable with patient samples for the assay being compared are also suitable for inter-instrument comparison.[5]

3.5.7.4 Results of internal quality control materials may also be used for inter-instrument comparison when comparing results of identical instrument using same lot of reagents and calibrators.[5]

3.5.7.5 If sample stability is a limiting factor such that patient samples cannot be used, materials used for external evaluation of performance may be considered for use as comparison materials.[5]

3.5.8 Inter-instrument comparison shall also be performed among different detection channels of the same autoanalyser if each channel has its own calibration curve. If the same calibration curve is used by different detection channels of the same autoanalyser, correlation study is not necessary.

3.6 Linearity

3.6.1 Linearity is the ability of an autoanalyser to provide results that are directly proportional to the concentration of the analyte in the test sample.[5] The linearity error within the reportable range shall be within the medically allowed tolerance. Samples used for the verification of linearity shall match closely with patient sample matrix. If multiple matrix sample types are analysed by the autoanalyser, a linearity study should be carried out for each sample type.[9] The concentrations of analyte used for the study shall cover the entire clinical relevant range including all medical decision levels and the laboratory's claimed reportable range. Use of a minimum of 5 different concentrations, each measured in duplicate for the verification is recommended. More levels of concentration (7 to 9) and replicates (2 or 3) should be used for validation of linearity of a modified method.[9] Both commercial standard panels and in-house prepared panels consisting of different dilutions of a high level sample can be used for the verification of linearity.

3.6.2 For haematology autoanalysers which function primarily as particle counters, linearity can be studied by making dilutions of blood and calculating the recovery of the expected concentrations. Although commercial 'linearity kits' exist, the materials therein (particles, matrix) often differ significantly from fresh whole blood; thus, their results may not correctly reflect instrument performance when analysing fresh blood. The materials used for linearity study should have the same matrix effect to the analyser as fresh blood. Use of concentrated or diluted fresh blood cells is preferred.[6]

3.7 Limit of Blank (LoB), Limit of Detection (LoD) and Limit of Quantitation (LoQ)

3.7.1 LoB, LoD and LoQ is a set of performance attributes used to characterise measurement accuracy in the low-end region of the measuring interval.

They reflect increasing informational content in the measurement procedure's ability to resolve measurand levels, from an upper boundary on expected blank sample measurements (LoB), to simple detection of measurand presence (LoD), to the minimal measurand amount that can be measured with defined accuracy (LoQ).[7]

3.7.2 Limit of Blank (LoB) is the highest measurement result that is likely to be observed (with a stated probability, α) for a blank sample.[7] The minimal experimental design to verify the manufacturer's LoB claim of an analyser is:

- One reagent lot
- One instrument system
- Three days
- Two blank samples
- Two replicates per sample per day
- 20 total blank replicates (across all samples and days)

3.7.3 The minimal design described in clause 3.7.2 (i.e. one instrument system, three days, two samples, two replicates) does not yield the necessary 20 total replicates per reagent lot. It is necessary for the laboratory to increase one or more design factors to provide a sufficient number of measurement results. The selection of which factors to increase depends on the particular measurement procedure and available resources for examination. The laboratory also may wish to add more factors and/or to increase the number of replicates beyond the minimum in order to increase the power of the verification experiment. The same principle applies to the verification of Limit of Detection (LoD) and Limit of Quantitation (LoQ) described in clauses 3.7.4 and 3.7.7.

3.7.4 Limit of Detection (LoD) is defined as the measured quantity value, obtained by a given measurement procedure, for which the probability of

falsely claiming the absence of a measurand in a material is β, given a probability α of falsely claiming its presence.[7] The minimal experimental design to verify the manufacturer's claim of LoD of an analyser is:

- One reagent lot
- One instrument system
- Three days
- Two samples at the claimed LoD level
- Two replicates per sample per day
- 20 total low level replicates (across all samples and days)

It is necessary to have the associated LoB claim in order to verify the LoD claim. The laboratory shall define the acceptance criteria for verification studies of both LoB and LoD according to recognised international guidelines or manufacturer's instructions.

3.7.5 Limit of Quantitation (LoQ), is the lowest amount of a measurand in a material that can be quantitatively determined with stated accuracy (as independent requirements for precision and bias), under stated experimental conditions.[7] LoQ is defined preferentially in terms of a total error goal or with respect to goals for both bias and precision. There may be situations, however, where bias cannot be determined at the appropriate measurand level and within-laboratory precision is used as the sole acceptance goal. In such cases, the LoQ would be equivalent to the older and now deprecated term "functional sensitivity". Functional sensitivity is a form of the LoQ, in which the threshold for results suitable for quantitative analysis is defined solely in terms of a precision requirement.

3.7.6 Determination of LoQ of an autoanalyser is usually required for immunoassays where detection of low analyte level is critical (e.g. Troponin, PSA, TSH) or if LoQ is defined for the measurement procedure. Whereas in haematology, determination of the LoQ may need to be considered in the

verification of factor assays[8] and under certain clinical situations where accurate quantitation of low WBC concentrations and low platelet concentrations may be important.[6]

3.7.7 The minimal experimental design for verification of the LoQ claimed is

- One reagent lot
- One instrument system
- Three days
- Two samples at the LoQ claim measurand concentration
- Two replicates per sample per day
- 20 total low level replicates (across all samples and days)

3.7.8 To estimate functional sensitivity, samples with several low analyte levels must be studied to determine the precision profile at the low level range. When possible, the matrix of the blanks and samples with low analyte level must be similar to that of natural samples. The level at which a predefined coefficient of variation (%CV) is obtained is taken as the functional sensitivity. The minimum number of replicates used for determining the %CV at each level is usually 20. The predefined CVs to be used by the laboratory shall be supported with quoted reference. To verify a claimed functional sensitivity, the laboratory may determine the %CV of samples with analyte level below or at the functional sensitivity claimed and if the %CV is below the predefined acceptable precision, e.g. 20% for TSH or 10% for Troponin, the claimed functional sensitivity is verified.

3.7.9 In most situations the LoD is lower than the LoQ verified, LoD shall not be used as the lower reporting limit when LoQ is applicable to that analyte.

3.8 Carry-over study

3.8.1 Carry-over is the contamination of a sample by the sample or reagent used in

the analysis immediately before it. This is most often due to contamination by the sampling syringe exposed to the previous sample or reagent.

3.8.2 Estimation on carry-over can be conducted by running a sample containing high level of the target analyte thrice, followed by running a sample containing low level of the target analyte thrice. For the estimation of carry-over in haematology autoanalysers, commercial controls should not be used to substitute fresh whole human blood because of matrix effects and difference in cellular composition.

3.9 Interference

3.9.1 The laboratory can decide whether to include interference study in the verification study of the autoanalyser based on the adequacy of manufacturer's information and the intended use of the analyser. Information on interference provided by the manufacturers can be adopted and used by the laboratory. The rationale for using the manufacturer's criteria and data shall be documented. Appropriate actions need to be in place when interfering substances at concentrations known or likely to cause erroneous results are encountered. Such results should not be reported and/or appropriate comments should be provided in the examination reports.

3.9.2 The three most common types of interference encountered in clinical samples are samples being lipaemic, haemolytic or icteric. When the laboratory conducts interference study to confirm the concentrations of the interferent that will cause significant difference to the examination results, it can be conducted through split samples and spiking.

3.9.3 The sample is split and one of them is spiked with the interfering substance under study and the other is not. The spiking material added to the sample shall not exceed 10% of the sample volume to avoid any significant change

in the sample matrix. An equal volume of diluent should also be added to the unspiked sample to compensate for the dilution effect. The concentrations of the interfering substance in the spiked sample should be at the concentrations commonly encountered in clinical practice. The paired samples should be examined in duplicates.

3.9.4 The increase of the total error caused by the interferent shall not result in altering the decision on diagnosis, treatment, or management of a patient made by a physician. Hence an assay, with the presence of interferent in the patient sample, is clinically acceptable when the total error is less than the total allowable analytical error.

3.10 High dose hook effect

3.10.1 High dose hook effect or antigen excess phenomenon should be studied for assays of analytes for which the normal pathologic concentration range is very wide (e.g. tumour markers) and one-step noncompetitive immunoassay designs are employed. It occurs when the concentrations of analyte in the patient sample exceed the binding capacity of the capture and signal antibodies, making these antibodies unavailable to form antibody-antigen complexes, thus resulting in a severe underestimation of the analyte concentration.

3.10.2 Manufacturer's claim on freedom from high dose hook effect shall be confirmed with patient samples of high analyte concentrations. Such high level samples can be stored samples of the laboratory or from other laboratories. When high level samples are not readily available, this confirmation can be performed after the autoanalyser has been put into routine service.

3.10.3 The laboratory shall make appropriate dilution of the sample under

examination to check if the expected concentration is reported by the autoanalyser. The highest concentration at which the autoanalyser has been demonstrated to exhibit no hook effect shall be documented. This value shall be updated whenever a sample of higher concentration has been correctly reported by the autoanalyser.

3.11 Autodilution

3.11.1 If the built-in autodilution feature of an autoanalyser would be used in routine operation, the precision and accuracy of the autodilution shall be verified. In general, this can be done by comparing the autodilution results with the manual dilution results using the manual dilution result as the target value.

4. VERIFICATION REPORT

4.1 A fully written verification report shall be prepared for the autoanalyser(s) under study. The verification report shall include, but not limited to, the following contents:

a. Name of person(s) carrying out the verification;
b. Objective of the verification;
c. Period of the verification and the analytes and applications covered;
d. Name, model and serial number of the autoanalyser(s);
e. For each analyte and application, details of each study performed including types and number of samples used, number of runs per sample, reagent lot and expiry date, etc and the results obtained;
f. The statistical method used for data analysis for each study and the acceptance criteria (should be defined before the study);
g. Summary data with appropriate graphical presentation and analysis;
h. Results and conclusions;

i. Limitations and precautions, if any;
j. Source of references.

4.2 The laboratory shall be responsible for the contents and information accuracy of the verification report. The verification report shall be reviewed and endorsed by staff with appropriate authority.

REFERENCES

1. Interference Testing in Clinical Chemistry, 3rd Edition. CLSI document EP07 (2018).
2. User Verification of Precision and Estimation of Bias; Approved Guideline – Third Edition. CLSI document EP15-A3 (2014).
3. James O. Westgard. Basic Method Validation, 3rd Edition – The Replication Experiment. 2009 http://www.westgard.com/lesson22.htm (accessed in March 2019).
4. Measurement Procedure Comparison and Bias Estimation Using Patient Samples – Third Edition. CLSI document EP09-Ed3 (2018).
5. Verification of Comparability of Patient Results Within One Health Care System; Approved Guideline (Interim Revision). CLSI document EP31-A-IR (2012).
6. Validation, Verification, and Quality Assurance of Automated Hematology Analyzers; Approved Standard – Second Edition. CLSI document H26-A2 (2010).
7. Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline – Second Edition. CLSI document EP17-A2 (2012).
8. Protocol for the Evaluation, Validation and Implementation of Coagulometers; Proposed Guideline. CLSI document H57-A (2008).
9. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. CLSI document EP06-A (2003).