OIV-MA-AS1-09 Protocol for the designe, conducts and interpretation of collaborative studies

Introduction

After a number of meetings and workshops, a group of representatives from 27 organizations adopted by consensus a "Protocol for the design, conducts and interpretation of collaborative studies" which was published in Pure & Appl. Chem. 60, 855-864, 1995. A number of organizations have accepted and used this protocol. As a result of their experience and the recommendations of the Codex Committee on Methods of Analysis and Sampling (Joint FAO/WHO Food Standards Programme, Report of the Eighteenth Session, 9-13 November, 1992; FAO, Rome Italy, ALINORM 93/23, Sections 34-39), three minor revisions were recommended for incorporation into the original protocol. These are: (1) Delete the double split level design because the interaction term it generates depends upon the choice of levels and if it is statistically significant, the interaction cannot be physically interpreted. (2) Amplify the definition of "material". (3) Change the outlier removal criterion from 1% to 2.5%.

The revised protocol incorporating the changes is reproduced below. Some minor editorial revisions to improve readability have also been made. The vocabulary and definitions of the document 'Nomenclature of Interlaboratory Studies (Recommendations 1994)' [published in Pure Appl Chem., 66, 1903-1911 (1994)] has been incorporated into this revision, as well as utilizing, as far as possible, the appropriate terms of the International Organization for Standardization (ISO), modified to be applicable to analytical chemistry.

Protocol

1. Preliminary work

Method-performance (collaborative) studies require considerable effort and should be conducted only on methods that have received adequate prior testing. Such within-laboratory testing should include, as applicable, information on the following:

1.1. Preliminary estimates of precision

Estimates of the total within-laboratory standard deviation of the analytical results over the concentration range of interest as a minimum at the upper and lower limits

of the concentration range, with particular emphasis on any standard or specification value.

Note 1: The total within-laboratory standard deviation is a more inclusive measure of imprecision that the ISO repeatability standard deviation, §3.3 below. This standard deviation is the largest of the within-laboratory type precision variables to be

expected from the performance of a method; it includes at least variability from different days and preferably from different calibration curves. It includes between-run (between-batch) as well as within-run (within-batch) variations. In this respect it can be considered as a measure of within-laboratory reproducibility. Unless this value is well within acceptable limits, it cannot be expected that the between-laboratory standard deviation (reproducibility standard deviation) will be any better. This precision term is not estimated from the minimum study described in this protocol.

NOTE 2: The total within-laboratory standard deviation may also be estimated from ruggedness trials that indicate how tightly controlled the experimental factors must be and what their permissible ranges are. These experimentally determined ranges should be incorporated into the description of the method.

1.2. Systematic error (bias)

Estimates of the systematic error of the analytical results over the concentration range and in the substances of interest, as a minimum at the upper and lower limits of the concentration range, with particular emphasis on any standard or specification value.

The results obtained by applying the method to relevant reference materials should be noted.

1.3. Recoveries

The recoveries of "spikes" added to real materials and to extracts, digests, or other treated solutions thereof.

1.4. Applicability

The ability of the method to identify and measure the physical and chemical forms of the analyte likely to be present in the materials, with due regard to matrix effects.

1.5. Interference

The effect of other constituents that are likely to be present at appreciable concentrations in matrices of interest and which may interfere in the determination.

1.6. Method comparison

The results of comparison of the application of the method with existing tested methods intended for similar purposes.

1.7. Calibration Procedures

The procedures specified for calibration and for blank correction must not introduce important bias into the results.

1.8. Method description

The method must be clearly and unambiguously written.

1.9. Significant figures

The initiating laboratory should indicate the number of significant figures to be reported, based on the output of the measuring instrument.

Note: In making statistical calculations from the reported data, the full power of the calculator or computer is to be used with no rounding or truncating until the final reported mean and standard deviations are achieved. At this point the standard deviations are rounded to 2 significant figures and the means and related standard deviations are rounded to accommodate the significant figures of the standard deviation. For example, if S_R = 0.012, c is reported as 0.147, not as 0. 1473 or 0. 15, and RSD_R is reported as 8.2%. (Symbols are defined in Appendix L) If standard deviation calculations must be conducted manually in steps, with the transfer of intermediate results, the number of significant figures to be retained for squared numbers should be at least 2 times the number of figures in the data plus 1.

2. Design of the method-performance study

2.1. Number of materials

For a single type of substance, at least 5 materials (test samples) must be used; only when a single level specification is involved for a single matrix may this minimum required number of materials to be reduced to 3. For this design parameter, the two portions of a split level and the two individual portions of blind replicates per laboratory are considered as a single material.

Note 1: A material is an 'analyte/matrix/concentration' combination to which the method-performance parameters apply. This parameter determines the applicability of a method. For application to a number of different substances, a sufficient number of matrices and levels should be chosen to include potential interferences and the concentration of typical use.

Note 2: The 2 or more test samples of blind or open replicates statistically, are a single material (they are not independent).

NOTE 3: A single split level (Youden pair) statistically analyzed as a pair is a single material; if analyzed statistically and reported as single test samples, they are 2 materials. In addition, the pair can be used to calculate the within-laboratory standard deviation, s_r as

$s_r = \sqrt{(\sum d_1^2)/2n}$ (for duplicates, blind or open

$s_r = \sqrt{(\sum d_1^2)/2 - (n-1)}$ (for duplicates, blind or open

where d_i, the difference between the 2 individual values from the split level for each laboratory and n is the number of laboratories. In this special case, S_R, the among laboratories standard deviation, is merely the average of the two S_R values calculated from the individual components of the split level, and it is used only as a check of the calculations.

Note 4: The blank or negative control may be a material or not depending on the usual purpose of the analysis. For example, in trace analysis, where very low levels (near the limit of quantitation) are often sought, the blanks are considered as materials and are necessary to determine certain 'limits of measurement.' However, if the blank is merely a procedural control in macro analysis (e.g., fat in cheese), it would not be considered a material.

2.2. Number of laboratories

At least 8 laboratories must report results for each material; only when it is impossible to obtain this number (e.g., very expensive instrumentation or specialized laboratories required) may the study be conducted with less, but with an absolute minimum of 5 laboratories. If the study is intended for international use, laboratories from different countries should participate. In the case of methods requiring the use of specialized instruments, the study might include the entire population of available laboratories. In such cases, "n" is used in the denominator for calculating the standard deviation instead of "(n - 1)". Subsequent entrants to the field should demonstrate the ability to perform as well as the original participant.

2.3. Number of Replicates

The repeatability precision parameters must be estimated by using one of the following sets of designs (listed in approximate order of desirability):

2.3.1. Split Level

For each level that is split and which constitutes only a single material for purposes of design and statistical analysis, use 2 nearly identical test samples that differ only slightly in analyte concentration (e.g., <1-5%). Each laboratory must analyse each test sample once and only once.

Note: The statistical criterion that must be met for a pair of test samples to constitute a split level is that the reproducibility standard deviation of the two parts of the single split level must be equal.

2.3.2. Combination blind replicates and split level

Use split levels for some materials and blind replicates for other materials in the same study (single values from each submitted test sample).

2.3.3. Blind replicates

For each material, use blind identical replicates, when data censoring is impossible (e.g., automatic input, calculation, and printout) non-blind identical replicates may be used.

2.3.4. Known replicates

For each material, use known replicates (2 or more analyses of test portions from the same test sample), but only when it is not practical to use one of the preceding designs.

2.3.5. Independent analyses

Use only a single test portion from each material (i.e., do not perform multiple analyses) in the study, but rectify the inability to calculate repeatability parameters by quality control parameters or other within-laboratory data obtained independently of the method-performance study.

3. Statistical analysis (See Flowchart, A.4. 1)

For the statistical analysis of the data, the required statistical procedures listed below must be performed and the results reported. Supplemental, additional procedures are not precluded.

3.1. Valid data

Only valid data should be reported and subjected to statistical treatment. Valid data are those data that would be reported as resulting from the normal performance of laboratory analyses; they are not marred by method deviations, instrument malfunctions, unexpected occurrences during performance, or by clerical, typographical and arithmetical errors.

3.2. One-way analysis of variance

One-way analysis of variance and outlier treatments must be applied separately to each material (test sample) to estimate the components of variance and repeatability and reproducibility parameters.

3.3. Initial estimation

Calculate the mean, c (= the average of laboratory averages), repeatability relative standard deviation, RSD_r and reproducibility relative standard deviation, RSD_R with no outliers removed, but using only valid data.

3.4. Outlier treatment

The estimated precision parameters that must also be reported are based on the initial valid data purged of all outliers flagged by the harmonized 1994 outlier removal procedure. This procedure essentially consists of sequential application of the Cochran and Grubbs tests (at 2.5% probability (P) level, 1-tail for Cochran and 2-tail for

Grubbs) until no further outliers are flagged or until a drop of 22.2% (= 219) in the original number of laboratories providing valid data would occur.

Note: Prompt consultation with a laboratory reporting suspect values may result in correction of mistakes or discovering conditions that lead to invalid data, 3.1.

Recognizing mistakes and invalid data per se is much preferred to relying upon statistical tests to remove deviate values.

3.4.1. Cochran test

First apply Cochran outlier test (1-tail test a P = 2.5%) and remove any laboratory whose critical value exceeds the tabular value given in the tale, Appendix A.3. 1, for the number of laboratories and replicates involved.

3.4.2. Grubbs tests

Apply the single value Grubbs test (2 tail) and remove any outlying laboratory. If no laboratory is flagged, then apply the pair value tests (2 tail) - 2 at the same end and 1 value at each end, P = 2.5% overall. Remove any laboratory(ies) flagged by these tests whose critical value exceeds the tabular value given in the appropriate column of the table Appendix A.3.3. Stop removal when the next application of the test will flag as table, A outliers more that 22.2% (2 of 9) of the laboratories.

Note: The Grubbs tests are to be applied one material at a time to the set of replicate means from all laboratories, and not to the individual values from replicated designs because the distribution of all the values taken together is multimodal, not Caussian, i.e., their differences from the overall mean for that material are not independent.

3.4.3. Final estimation

Recalculate the parameters as in §3.3 after the laboratories flagged by the preceding procedure have been removed. If no outliers were removed by the Cochran-Grubbs sequence, terminate testing. Otherwise, reapply the Cochran-Grubbs sequence to the data purged of the flagged outliers until no further outliers are flagged or until more than a total of 22.2% (2 of 9 laboratories) would be removed in the next cycle. See flowchart A.3.4.

4. Final report

The final report should be published and should include all valid data. Other information and parameters should be reported in a format similar (with respect to the reported items) to the following, as applicable:

[x] Method-performance tests carried out at the international level in [year(s)] by [organisation] in which [y and z] laboratories participated, each performing [k] replicates, gave the following statistical results:

Table of method -Performance parameters

Analyte; Results expressed in [units]

Material [Description and listed in columns across top of table in increasing order of magnitude of means]

Number of laboratories retained after eliminating outliers

Number of outlying laboratories

Code (or designation) of outlying laboratories

Number of accepted results

Mean

True or accepted value, if known

Repeatability standard deviation (S_r)

Repeatability relative standard deviation (RSD_R)

Repeatability limit, r (2.8 x S_r)

Reproducibility standard deviation (S_R)

Reproducibility relative standard deviation (RSD_R)

Reproducibility limit, R (2.8 X S_R)

4.1. Symbols

A set of symbols for use in reports and publications is attached as Appendix 1 (A.1.).

4.2. Definitions

A set of definitions for use in study reports and publications is attached as Appendix 2 (A.2.).

4.3. Miscellaneous

4.3.1. Recovery

Recovery of added analyte as a control on method or laboratory bias should be calculated as follows:

[Marginal] Recovery, %=

(Total analyte found - analyte originally present) x 100/(analyte added)

Although the analyte may be expressed as either concentration or amount, the units must be the same throughout. When the quantity of analyte is determined by analysis, it must be determined in the same way throughout.

Analytical results should be reported uncorrected for recovery. Report recoveries separately.

4.3.2. When S_r, is negative

By definition, S_R is greater than or equal to S_r in method-performance studies; occasionally the estimate of S_r is greater than the estimate of S_R (the average of the replicates is greater than the range of laboratory averages and the calculated S_L^2 is then negative). When this occurs, set S_L = 0 and $S_R = S_r$.

5. References

- Horwitz, W. (1988) Protocol for the design, conduct, and interpretation of method performance studies. Pure & Appl. Chem. 60, 855-864.
- Pocklington, W.D. (1990) Harmonized protocol for the adoption of standardized analytical methods and for the presentation of their performance characteristics. Pure and Appl. Chem. 62, 149-162.
- International Organization for Standardization. International Standard 5725-1986. Under revision in 6 parts; individual parts may be available from National Standards member bodies.

Appendices

Appendix 1. - Symbols

Use the following set of symbols and terms for designating parameters developed by a method-performance study.

Mean (of laboratory averages): x

Standard deviations:s (estimates)

- Repeatability: S_r
- 'Pure' between-laboratory: S_L
- Reproducibility; S_R

Variances: S^2 (with subscripts, r, L, and R)

$$S_R^2 = S_L^2 + S_r^2$$

Relative standard deviations: RSD (with subscripts, r, L, and r)

Maximum tolerable differences

(as defined by ISO 5725-1986);

See A.2.4 and A.2.5)

Repeatability limitr = (2.8 x S_r)

Reproducibility limit R = (2.8 X S_R)

Number of replicates per laboratory :k (general)

Average number of replicates per laboratory i:k (for a balanced design)

Number of laboratories :L

Number of materials (test samples): m

Total number of values in a given assay: n (= kL for a balanced design)

Total number of values in a given study: N (= kLm for an overall balanced design)

————————————————————

If other symbols are used, their relationship to the recommended symbols should be explained fully.

Appendix 2. - Definitions

Use the following definitions. The first three definitions utilize the 1UPAC document "Nomenclature of Interlaboratory Studies" (approved for publication 1994). The next two definitions are assembled from components given in ISO 3534-1:1993. All test results are assumed to be independent, i.e., 'obtained in a manner not influenced by any previous result on the same or similar test object. Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme stipulated conditions.'

1. A.2.1 Method-performance studies

   An interlaboratory study in which all laboratories follow the same written protocol and use the same test method to measure a quantity in sets of identical test items [test samples, materials]. The reported results are used to estimate the performance characteristics of the method. Usually these characteristics are within-laboratory and among-laboratories precision, and when necessary and possible, other pertinent characteristics such as systematic error, recovery, internal quality control parameters, sensitivity, limit of determination, and applicability.

2. A.2.2 Laboratory-performance study

   An interlaboratory study that consists of one or more analyses or measurements by a group of laboratories on one or more homogeneous, stable test items, by the method selected or used by each laboratory. The reported results are compared with those of other laboratories or with the known or assigned reference value, usually with the objective of evaluating or improving laboratory performance.

3. A.2.3 Material certification stud

   An interlaboratory study that assigns a reference value ('true value') to a quantity (concentration or property) in the test item, usually with a stated uncertainty.

4. A.2.4 Repeatability limit (r)

When the mean of the values obtained from two single determinations with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time, lies within the range of the mean values cited in the Final Report, 4.0, the absolute difference between the two test results obtained should be less than or equal to the repeatability limit (r) [= 2.8 x s,) that can generally be inferred by linear interpolation of s_r from the Report.

Note: This definition, and the corresponding definition for reproducibility limit, has been assembled from five cascading terms and expanded to permit application by interpolation to a test item whose mean is not the same as that used to establish the original parameters, which is the usual case in applying these definitions. The term 'repeatability [and reproducibility] limit' is applied specifically to a probability of 95% and is taken as 2.8 x s, [or SRI. The general term for this statistical concept applied to any measure of location (e.g., median) and with other probabilities (e.g., 99%) is "repeatability [and reproducibility] critical difference".

5. A.2.5 Reproducibility limit (R)

When the mean of the values obtained from two single determinations with the same method on identical test items in different laboratories with different operators using different equipment, lies within the range of the mean values cited in the Final Report, 4.0, the absolute difference between the two test results obtained should be less than or equal to the reproducibility limit (R) [= 2.8 x s_R] that can generally be inferred by linear interpolation of S_R from the Report.

Note 1: When the results of the interlaboratory test make it possible, the value of r and R can be indicated as a relative value (e.g., as a percentage of the determined mean value) as an alternative to the absolute value.

Note 2: When the final reported result in the study is an average derived from more than a single value, i.e., k is greater than 1, the value for R must be adjusted according to the following formula before using R to compare the results of a single routine analyses between two laboratories.

$$R' = (R^2 + r^2(1 - [l/k])^{1/2}$$

Similar adjustments must be made for replicate results constituting the final values for S_R and RSD_R, if these will be the reported parameters used for quality control purposes.

Note 3: The repeatability limit, r, may be interpreted as the amount within which two

determinations should agree with each other within a laboratory 95% of the time. The reproducibility limit, R, may be interpreted as the amount within which two separate determinations conducted in different laboratories should agree with each other 95% of the time.

Note 4: Estimates Of S_R can be obtained only from a planned, organized method performance study; estimates of S_r can be obtained from routine work within a laboratory by use of control charts. For occasional analyses, in the absence of control charts, within-laboratory precision may be approximated as one half S_R (Pure and Appl. Chem., 62, 149-162 (1990) , Sec. L3, Note.).

6. A.2.6 One-way analysis of variance

One-way analysis of variance is the statistical procedure for obtaining the estimates of within laboratory and between-laboratory variability on a material-by-material basis. Examples of the calculations for the single level and single-split-level designs can be found in ISO 5725-1986.

Appendix 3. – Critical values

1. A.3.1 Critical values for the Cochran maximum variance ratio at the 2.5% (1 -tail) rejection level, expressed as the percentage the highest variance is of the total variance; r = number of replicates.

No of labs	r=2	r=3	r=4	r=5	=6
4	94.3	81.0	72.5	65.4	62.5
5	88.6	72.6	64.6	58.1	53.9
6	83.2	65.8	58.3	52.2	47.3
7	78.2	60.2	52.2	47.3	42.3
8	73.6	55.6	47.4	43.0	38.5
9	69.3	51.8	43.3	39.3	35.3
10	65.5	48.6	39.9	36.2	32.6
11	62.2	45.8	37.2	33.6	30.3

12	59.2	43.1	35.0	31.3	28.3
13	56.4	40.5	33.2	29.2	26.5
14	53.8	38.3	31.5	27.3	25.0
15	51.5	36.4	29.9	25.7	23.7
16	49.5	34.7	28.4	24.4	22.0
17	47.8	33.2	27.1	23.3	21.2
18	46.0	31.8	25.9	22.4	20.4
19	44.3	30.5	24.8	21.5	19.5
20	42.8	29.3	23.8	20.7	18.7
21	41.5	28.2	22.9	19.9	18.0
22	40.3	27.2	22.0	19.2	17.3
23	39.1	26.3	21.2	18.5	16.6
24	37.9	25.5	20.5	17.8	16.0
25	36.7	24.8	19.9	17.2	15.5
26	35.5	24.1	19.3	16.6	15.0
27	34.5	23.4	18.7	16.1	14.5
28	33.7	22.7	18.1	15.7	14.1
29	33.1	22.1	17.5	15.3	13.7
30	32.5	21.6	16.9	14.9	13.3
35	29.3	19.5	15.3	12.9	11.6
40	26.0	17.0	13.5	11.6	10.2

50	21.6	14.3	11.4	9.7	8.6

Tables A.3.1 and A.3.3 were calculated by R. Albert (October, 1993) by computer simulation involving several runs of approximately 7000 cycles each for each value, and then smoothed. Although Table A.3.1 is strictly applicable only to a balanced design (same number of replicates from all laboratories), it can be applied to an unbalanced design without too much error, if there are only a few deviations.

2. A.3.2 Calculation of Cochran maximum variance outlier ratio

Compute the within-laboratory variance for each laboratory and divide the largest of these variances by the sum of the all of the variances and multiply by 100. The resulting quotient is the Cochran statistic which indicates the presence of a removable outlier if this quotient exceed the critical value listed above in the Cochran table for the number of replicates and laboratories specified.

3. A.3.3 Critical values for the Grubbs extreme deviation outlier tests at the 2.5% (2-tail), 1.25% (1tail) rejection level, expressed as the percent reduction in standard deviations caused by the removal of the suspect value(s).

No. of labs	One highest or lowest	Two highest or two lowest	One highest and one lowest
4	86.1	98.9	99.1
5	73.5	90.9	92.7
6	64.0	81.3	84.0
7	57.0	73.1	76.2
8	51.4	66.5	69.6
9	46.8	61.0	64.1
10	42.8	56.4	59.5
11	39.3	52.5	55.5
12	36.3	49.1	52.1

13	33.8	46.1	49.1
14	31.7	43.5	46.5
15	29.9	41.2	44.1
16	28.3	39.2	42.0
17	26.9	37.4	40.1
18	25.7	35.9	38.4
19	24.6	34.5	36.9
20	23.6	33.2	35.4
21	22.7	31.9	34.0
22	21.9	30.7	32.8
23	21.2	29.7	31.8
24	20.5	28.8	30.8
25	19.8	28.0	29.8
26	19.1	27.1	28.9
27	18.4	26.2	28.1
28	17.8	25.4	27.3
29	17.4	24.7	26.6
30	17.1	24.1	26.0
40	13.3	19.1	20.5
50	11.1	16.2	17.3

4. A.3.4 Calculation of the Grubbs test values

To calculate the single Chubbs test statistic, compute the average for each laboratory and then calculate the standard deviation (M) of these L averages (designate as the original s). Calculate the SD of the set of averages with the highest average removed (SH); calculate the SD of the set of averages with the lowest average removed (SL). The calculate the percentage decrease in SD for both as follows:

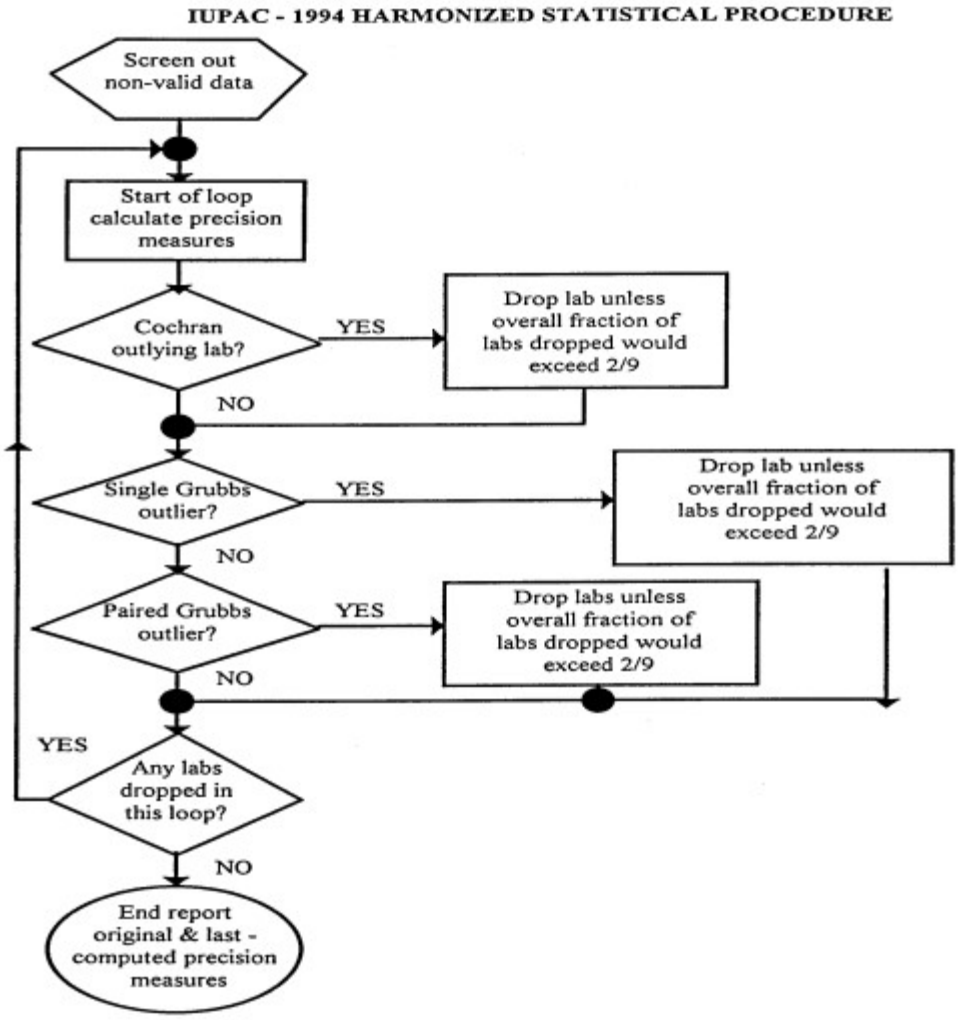
- 100 x [1 - (sL/s] and 100 x [1 - (sH/s)].

The higher of these two percentage decreases is the singe Grubbs test statistic, which signal the presence of an outlier to be omitted at the P = 2.5% level, 2□tail, if it exceeds the critical value listed in the single value column, Column 2, of Table A.3.3 , for the number of laboratory averages used to calculate the original s.

To calculate the paired Grubbs test statistics, calculate the percentage decrease in standard deviation obtained by dropping the two highest averages and also by dropping the two lowest averages, as above. Compare the higher of the percentage changes in standard deviation with the tabular values in column 3 and proceed with (1) or (2): (1) If the tabular value is exceeded, remove the responsible pair. Repeat the cycle again, starting at the beginning with the Cochran extreme variance test again, the Grubbs extreme value test, and the paired Grubbs extreme value test. (2) If no further values are removed, then calculate the percentage change in standard deviation obtained by dropping both the highest extreme value and the lowest extreme value together, and compare with the tabular values in the last column of A.3.3. If the tabular value is exceeded, remove the high-low pair of averages, and start the cycle again with the Cochran test until no further values are removed. In all cases, stop outlier testing when more than 22.2% (2/9) of the averages are removed.

Appendix 4

1. A.4.1. Flowchart for outlier removal


IUPAC - 1994 HARMONIZED STATISTICAL PROCEDURE
Screen out non-valid data
Start of loop calculate precision measures
Cochran outlying lab?
YES
Drop lab unless overall fraction of labs dropped would exceed 2/9
NO
Single Grubbs outlier?
YES
Drop lab unless overall fraction of labs dropped would exceed 2/9
NO
Paired Grubbs outlier?
YES
Drop labs unless overall fraction of labs dropped would exceed 2/9
NO
YES
Any labs dropped in this loop?
NO
End report original & last - computed precision measures