Pergamon

0731-7085(94)00083-2

Journal of Pharmaceutical & Biomedical Analysis, Vol. 12, No. 11, pp. 1337–1343, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0731–7085/94 $7.00 + 0.00

Analytical Survey

An analysis of the Washington Conference Report on bioanalytical method validation

C. HARTMANN, D.L. MASSART* and R.D. McDOWALL†

ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium
† Department of Chemistry, University of Surrey, Guildford, Surrey GU2 5HX, UK

Abstract: The Washington Conference Report on bioanalytical method validation is analysed with respect to the requirements for precision and accuracy. It is shown that if the requirements are interpreted too literally, this could lead to disappointment in practice. A better approach is to separate the total measurement error into its constant (bias) and random (precision) components. To ensure that 95% of all methods fall within the acceptance interval of ±15% around the true value, would require, for example, the bias to be ≤8% and the method precision to be ≤8% relative standard deviation (RSD; $n = 5$).

Keywords: *Bioanalytical validation; accuracy; bias; precision; acceptance limits.*

Introduction

Before an analytical method can be used for routine analysis, it must first be demonstrated that the method fulfils certain performance criteria. When this has been documented, the method is said to be validated. The first, and one of the main difficulties for the practising analytical chemist, is to decide exactly which parameters should be measured and to set the performance criteria which have to be fulfilled before a method can be said to be validated. Once the parameters have been fixed, it must be shown that they meet the performance criteria. How to do this is the second difficulty that is faced by the analyst. To provide guidance in solving this problem, specifically in the field of bioanalytical method validation, a conference was organized in Washington in December 1990. The conclusions of this meeting of professionals from industry, academia and regulatory agencies are summarized in a document, known as the Guidelines of the Washington Conference [1].

This conference achieved an important aim in bringing scientists together to discuss essential principles for the validation of an analytical method and to set minimum standards for method performance. However, there was still some controversy at the meeting and a complete consensus could not be reached. A review meeting took place in June 1994 in Munich, Germany.

It is necessary to look critically at certain aspects of the guidelines in this document. The objectives of the present article are to show what the acceptance limits really mean and to formulate recommendations to improve the guidelines. However, the terminology used in the guidelines must be examined first.

Terminology

A glossary in the Guidelines defines most of the analytical terms used in the validation of a method. However, internationally accepted definitions such as those by ISO or IUPAC already exist and have been carefully elaborated over many years. These definitions do not always agree but it is not a good idea to develop yet another set. It would be better to reach a consensus on one single terminology that is broad enough to be used in all fields of analysis.

* Author to whom correspondence should be addressed.

In this text the most recent ISO guidelines [2] will be followed to explain some of the deficiencies of the terminology introduced by the Washington consensus guidelines. The glossary of these guidelines defines *accuracy* as "*Closeness of determined value to the true value. Generally, recovery of added analyte over an appropriate range of concentrations is taken as an indication of accuracy. Whenever possible, the concentration range chosen should bracket the concentration of interest*". The first sentence is close to the definition of ISO (see Appendix) and the authors of the present survey certainly agree that the method should be validated over all expected concentrations. It is correct that recovery can be taken as an indication that a method is accurate but it is no more than an indication. Inclusion of recovery in a definition of accuracy may lead some analysts to conclude that adequate recovery always means that a method is accurate and that, of course, is not true. Suppose that a method is not selective and that some interference is also measured. The result will then be a certain (approximately the same) amount too high in both the unspiked and the spiked sample. However, the difference between the two results, from which the recovery is calculated, will be correct, leading to the false conclusion that the method is accurate.

Still more important is the fact that international organizations such as ISO (see Appendix) make it clear in their definition of accuracy that a result can be affected by a combination of two different kinds of experimental errors; these are systematic and random errors. The systematic error of an analytical method is the difference of the mean value (obtained by the method in the population of the measurements, i.e. with an unlimited number of experiments in each of an unlimited number of laboratories) from the true or an accepted reference value. The measure of this difference is the *bias*. Note that accuracy is the concept and that bias is the measure. Bias, therefore, measures the systematic error.

The definition given above illustrates the essential problem of statistics. If it were possible to carry out an unlimited number of replicate determinations, a mean result would be obtained. This mean result is called the limiting mean (or population mean or true mean) by statisticians and can be compared to the true or accepted reference value. In ISO terms, the *trueness* is determined in this way and the difference between the true mean and the true value is then the bias. However, in practice, only a limited number of experiments can be carried out so that the mean is only an estimate of the true mean and the value obtained for the bias is only an estimate of the true bias. The estimate becomes better when more experiments have been carried out.

Precision and the related terms described below are associated with random errors. In the glossary of the Guidelines, precision is subdivided into within-day (intra-) and between-day (inter-) assay precision. However, in the text of the Guidelines, the terms 'repeatability', 'reproducibility' and even 'imprecision' and 'variability' are employed without definition in the glossary. This reflects the consensus nature of the document; several independent groups wrote the different sections of the guidelines.

ISO defines *repeatability* as the closeness of agreement between independent test results obtained with the same method on identical test material under the same conditions (same laboratory, same operator, same equipment, within short intervals of time). *Reproducibility* is defined as the closeness of agreement between individual test results obtained with the same method on identical test material but in different laboratories with different operators using different equipment and not necessarily in short intervals of time. It is not always relevant or practical to measure reproducibility as affected by the different factors (laboratories, operators, equipment, time) and therefore, ISO defines *intermediate precision measures* for use in one laboratory, where one, two or three of the factors — operator, equipment and time — are changed (see also Appendix). While it would appear that the terms 'repeatability' and 'reproducibility' as used in the Guidelines are meant to follow international terminology and that 'imprecision' in principle describes the converse of 'precision', it is not clear what is meant by 'variability'.

Analysis of Statistical Aspects

Systematic and random errors

Consider in somewhat more detail systematic and random errors and how the latter interfere with the measurement of the former. Random errors are expressed as the standard deviation (SD) or as the relative standard deviation (RSD). Again the true standard deviation (σ) of a method, the so-called population parameter, can only be determined from an unlimited number of experiments. It can be estimated by the standard deviation of the results of n replicate measurements (s). The RSD is the standard deviation of the results of a certain number of measurements divided by the mean of these measurements; this ratio is expressed as a percentage.

Consider a method (I) with a true (i.e. population) bias of −10% and a true (i.e. population) repeatability relative standard deviation (RSD) of 10%. At first sight this method should satisfy the Washington consensus guidelines.

This method is applied in laboratory A and the reference value found is 100. If there was only the systematic error (−10%) of the method and no random error, the result would, therefore, be 90. The repeatability of the method is 10%, i.e. the standard deviation of the normal distribution of the results of replicate single measurements is 9, i.e. 10% of 90. With this repeatability, equation (1) can be used to calculate the concentration interval (*CI*), i.e. the range within which the results of n replicate measurements will be found with a chosen probability $(1-\alpha)$. If $\alpha = 5\%$, then the probability is 95% that a calculated mean value (obtained from n replicate measurements) is included within the calculated interval around the true population mean (μ).

$$CI = \mu \pm z_{(\alpha/2)} \frac{\sigma}{\sqrt{n}} \qquad (1)$$

where σ is the population (i.e. true) standard deviation of replicate single measurements, n is the number of replicates, $\sigma/\sqrt{n}$ is the true standard deviation of the mean and $z_{(\alpha/2)}$ is the tabulated z-value (two-sided) at the significance level α. The intervals $[-z, +z]$ describe the intervals within which a standardized normally distributed variable lies with the probability $(1-\alpha)$.

For $\alpha = 5\%$ and five replicate determinations:

$$CI = 90 \pm 1.96 \frac{9}{\sqrt{5}} = 82.1\text{–}97.9.$$

If $\alpha = 5\%$ but the number of replicates is increased to eight:

$$CI = 90 \pm 1.96 \frac{9}{\sqrt{8}} = 83.8\text{–}96.2,$$

Figure 1 shows the distribution of the results that would have been obtained if an unlimited number of determinations, each with five replicates, had been carried out. The distribution is normal. The most probable value is 90 and from the calculation given above, it is concluded that in 95% of all cases the mean of five determinations will be situated within the limits 82.1–97.9. In many cases, the resulting mean value will actually be, as expected, within the limits of ±15% around the nominal value 100 and there is a high probability that this would have occurred in laboratory A. Laboratory A would then conclude that the method has been validated.

However, the normal distribution around the biased mean exceeds the acceptance limits of the Guidelines at the lower level, as can be seen also in Fig. 1. Therefore, it will also be possible that the mean value of five determinations with this method will not be included in the acceptance interval [85–115]. This means that applying the same method in another laboratory, B or repeating the experiment in laboratory A can easily give a result outside the acceptance interval since results between 82.1 and 85 are not improbable (Fig. 1). This of course will lead to disappointment, for instance, in interlaboratory comparisons. However, such disappointment is due to poor understanding of the very different nature of systematic and random error.

A bioanalyst is faced with the problem that he or she does not *know* the magnitude of the true repeatability and bias and can only *estimate* them. There are tables [3] that provide

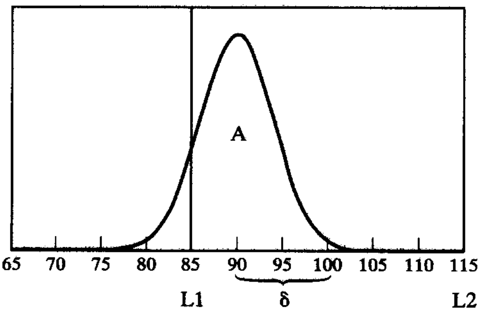

Figure 1
Normal distribution (A) of the mean values of five replicates that will be obtained with a method with a systematic error of −10% (δ) of the true value, a RSD of 10% and tolerance limits at −15% (L1) and +15% (L2) of the true reference value 100.

information on the number of experiments that should be carried out to estimate these parameters sufficiently well. This number depends on three parameters: λ, α and β. λ is defined as the ratio of δ, the minimum bias required to be detected and s, the estimated repeatability.

To understand α and β further, consideration must be given to the statistics of the results. When a method is validated a decision is made about the acceptance of the method. The analyst would like to know the probability that an error has been made in that decision. When two values are compared (for instance an experimental mean value to a reference value) the *null hypothesis* (*H0*) will be to state that the mean result from n replicate measurements is the same as the reference value. The observed deviation from the expected value is then considered to be due to random errors and the method will be considered acceptable. The *alternative hypothesis* (*H1*) will then be to state that the method performance is not acceptable, i.e. that there is a significant and too large difference between the calculated value and the reference value.

Incidentally, it might be preferable to state the hypotheses in a different way, i.e. to rewrite the null and alternative hypotheses in a way similar to that used in bioequivalence testing [4]. This yields two null hypotheses (H0) and one alternative hypothesis (H1):

$$\text{H0: } (\mu_T - \mu)/\mu \leqslant -15\% \text{ or } (\mu_T - \mu)/\mu \geqslant +15\%$$

where μ_T is the true mean for the test sample and μ is the nominal spiked value.

$$\text{H1: } -15\% < (\mu_T - \mu)/\mu < +15\%.$$

To keep the arguments simple, however, this way of stating the hypotheses will not be followed in this text.

It is usual to focus on the acceptance of the null hypothesis. The decision of accepting H0 carries with it two kinds of errors, α-errors and β-errors. The α-error is the error of rejecting the null hypothesis when H0 is in fact true. The often forgotten second error, the so-called β-error, or error of the second kind, is to accept the null hypothesis when H0 is in fact not true. Applied to the problem under consideration, the α-error is the risk that one would conclude that there is a bias, when in fact there is none. The β-error, on the other hand, is the risk that a bias of a certain magnitude will go unnoticed.

It is important to avoid wrong decisions. In order to decrease both kinds of errors simultaneously, data must be generated from a sufficient number of experiments. Often, it will not be practical to perform many experiments to achieve results with very small α- and β-errors. It is, however, necessary to know the risks that are taken in a certain situation.

In method validation it is usual to set $\alpha = 0.05$. In some other fields other significance levels are used. It seems sensible to specify the same value for β but in the analytical literature, there is no guidance about this requirement. The medical literature, however, is often more lenient for β than for α; for instance a value of $\beta = 0.1$ (or even $\beta = 0.2$) may be accepted. For a β value of 0.1, tables [3] show that λ values of 1.75, 1.4 and 1.0 require six, eight and 13 replicates, respectively (two-sided $\alpha = 0.05$). Clearly, λ needs to be known so that a decision can be made on the magnitude of n. As stated above, λ is the ratio of the minimum bias required to be detected and the estimated repeatability. To investigate analytical methods properly, it is imperative to separate the total measurement error into systematic (bias) and random (precision) components. Then the probability can be calculated to meet the requirements or to know how precise and accurate a method should be to comply with the acceptance limits of the Guidelines with only a small chance of a wrong decision.

Separation of the two kinds of errors is also useful in deciding which corrective action should be taken. For instance, random error components (RSD), can be decreased without changing the analysis procedure simply by increasing the number of replicates.

However, when a method is unacceptable owing to a too large systematic error replicate measurements will not be of value. Then the method has to be changed; for example an extraction step could be optimized to obtain a higher recovery.

The significance of the 15% limit

The values of 15 and 20%, which are specified in the document, will not be discussed until later in the paper; clearly, they are values with which the members of the Conference felt comfortable. It is important, however, to know what these values mean in terms of acceptable random error and bias. To be able to understand their meaning, the probability of achieving the ±15% criterion of the Washington Guidelines has been calculated for different combinations of systematic and random errors.

How does the analyst know the α-risk that the requirements of the Guidelines will not be met by the determination of n replicates using a particular method with a certain bias and precision? Knowledge of this risk requires calculation of the areas of those parts of the normal distribution around the biased mean that are not included in the acceptance interval 85–115% (or 80–120% at the limit of quantification). The calculations are illustrated for the earlier examples of method (I) that has a bias of −10% and a repeatability of 10% when five replicate determinations are carried out at the true concentration of 100. To calculate the areas of the parts outside the acceptance limits it is necessary to compute z, the value of the standardized normal distribution:

$$z = \frac{limit - \mu_T}{\sigma_m} \qquad (2)$$

where *limit* is the lower or upper limit of the normal distribution under consideration, μ_T is the true mean of the test samples and σ_m is the standard deviation of the normal distribution of the means of n determinations.

Application of equation (2) leads to the value of the standardized normal distribution for the lower limit:

$$z_1 = \frac{limit - \mu_T}{\sigma/\sqrt{n}} = \frac{85 - 90}{9/\sqrt{5}} = -1.242.$$

Then the percentage of the standardized normal distribution that is lower than this value of z_1 is determined. This percentage corresponds to the probability of obtaining a result that is smaller than the acceptance limit of the Guidelines. The percentage of the area corresponding to −1.242 can be found in standard statistical tables. A second possibility is to calculate the cumulative integral of the frequency function of the standardized normal distribution using equation (3):

$$P(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0.5u^2} \, du \qquad (3)$$

where z is the standardized normal variable and u represents all possible values from $-\infty$ to z.

It is then found that values that are smaller than z_1 correspond to an area of 10.7%. On the other hand, calculation of z for the upper limit leads to:

$$z_1 = \frac{limit - \mu_T}{\sigma/\sqrt{n}} = \frac{115 - 90}{9/\sqrt{5}} = 6.21.$$

The corresponding area of values higher than z_2 is <0.001% and can be neglected. The overall probability to obtain a result within the acceptance limits is, therefore, 100% − 10.7%, i.e. 89.3%.

This probability has been calculated for specific combinations of systematic and random errors. The probability that the mean of five replicate determinations from a method with a positive bias will fall in the interval of ±15% around the true value is given in the graph in Fig. 2(a). When more replicates are analysed, e.g. eight instead of the minimally required five replicates [1], the probability is somewhat larger as can be seen in Fig. 2(b). If calculations are carried out with a negative bias

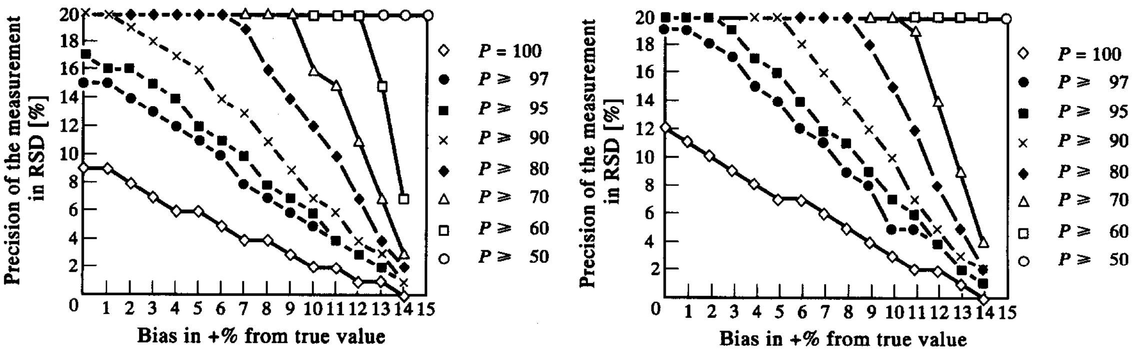

Figure 2
(a) Probability [P in %] that a measurement mean (n = 5) with a given positive bias and a given random measurement error is included in the ±15% interval around the true value. (b) Probability [P in %] that a measurement mean (n = 8) with a given positive bias and a given random measurement error is included in the ±15% interval around the true value.

the results are also slightly better, owing to the relatively smaller random error. In order to obtain mean values within the limits of ±15% around the true value with a probability of ≥95%, giving equal weight to both types of errors, it is concluded from Fig. 2(a) that *the proposed guidelines require the bias and the precision (as RSD) to be ≤8% for five replicates and ≤9% for eight replicates.* When the bias becomes larger, then the precision must increase to meet the acceptance limits with sufficient probability and vice versa. When both precision and bias are >8% ($n = 5$) or >9% ($n = 8$), then the sample size must be increased, i.e. more than five or eight replicates must be analysed so that a reliable decision can be made to accept or reject an analytical method.

From the point of view of the bioanalyst, who does not know the bias and the precision of a method, equal weights for bias and precision mean a value for λ of 1. As shown above this requires a relatively high value for n. If it is preferred that n be not greater than eight, then the estimated precision would be somewhat lower!

Repeatability and reproducibility

It is important to investigate whether the 15 and 20% precision requirements are in line with what can be expected generally. To estimate the reproducibility RSD that is attainable at a certain concentration, the equation proposed by Horwitz *et al.* [5], deduced from interlaboratory studies in many different fields, can be applied:

$$\mathrm{RSD} = 2^{(1-0.5\log x)} \qquad (4)$$

where x is the concentration in μg g^{-1} expressed in negative powers of 10.

With this equation, by calculation for 10 ng ml^{-1} (i.e. a concentration of 0.01 μg g^{-1} or, in negative powers of 10, 10^{-8}), the RSD is $2^{(1-0.5\log 10^{-8})} = 32\%$. More examples are shown in Table 1. The consensus guidelines, on the other hand, only require that the precision, expressed as RSD, must be better than 15%. Equal limits for repeatability and reproducibility are, however, not sensible in practice. Acceptance of a method when the repeatability RSD is close to 15% will lead to disappointment during further validation of the performance. It is unlikely, even with methods that are insensitive to small changes in the experimental conditions, that the reproducibility will be as good as the repeatability. Horwitz [5] deduced from a large number of experiments that, in most cases, the ratio of the RSD of repeatability and the RSD of reproducibility (calculated from interlaboratory studies) will be between 1/2 and 2/3. According to ISO, similar ratios can occur even in one single laboratory when the RSD of repeatability and the RSD of the so-called intermediate precision estimate are compared. Even a ratio of 1/2 or 1/3 is acceptable in chemical analysis [2], when the measure of the intermediate precision is based upon at least three different factors, e.g. time, operators and instruments (see Appendix). In accordance with this experience, it will, therefore, be unlikely in practice to obtain a reproducibility RSD of 15% when the repeatability RSD of the method is 15%. Since the differences between the repeatability RSD and reproducibility RSD can become quite large (Table 1) it is preferable to set separate limits for the repeatability and the reproducibility of the method.

Recommendations

The Washington Guidelines provide a pragmatic approach for the validation of a bioanalytical method. This is a first step towards better quality of bioanalytical data. However, it would be preferable for the terminology to be consistent with existing guidelines in other fields of chemical analysis. More important,

Table 1
Concentration-dependent attainable reproducibility RSD calculated according to the method of Horwitz *et al.* [5]; repeatability is expressed as 1/2 to 2/3 of reproducibility RSD

Concentration (ng ml^{-1})	Reproducibility RSD (%)	Repeatability RSD (%)
10	32	16–21
100	22	11–15
1000	16	8–11

careful attention should be paid to the nature of experimental errors and to statistical considerations. It would be better to make separate recommendations for maximum allowable bias and minimum precision as with the EU guideline for the control of residues in food [6] where separate requirements are given for bias and repeatability for different concentrations. Another example is in clinical chemistry where proposed analytical requirements have to be fulfilled by an analytical quality control procedure in order to comply with the limits of the guidelines of the American National Cholesterol Program; separate limits are given for bias and reproducibility [7].

Acknowledgements — The authors would like to thank the referees, Dr K. Gallimore and Dr I.J. McGilveray, for their helpful suggestions.

References

[1] V.P. Shah, K.K. Midha, S. Dighe, I.J. McGilveray, J.P. Skelly, A. Yacobi, T. Layloff, C.T. Viswanathan, C.E. Cook, R.D. McDowall, K.A. Pittman and S. Spector, *Pharm. Res.* **9**, 588–592 (1992).

[2] International Organization for Standardization, in *Accuracy (Trueness and Precision) of Measurement Methods and Results*, ISO/DIS 5725-1 and 5725-3, (Draft versions 1990/91).

[3] International Organization for Standardization, in *Statistical Methods*, ISO Standards Handbook 3, 3rd Ed. (1989).

[4] V.W. Steinijans and D. Hauschke, *Clin. Research and Reg. Affairs* **10**, 203–220 (1993).

[5] W. Horwitz, L.R. Kamps and K.W. Boyer, *J. Assoc. Off. Anal. Chem.* **63**, 1344–1354 (1980).

[6] The Rules Governing Medicinal Products in the European Community, in *Establishment by the European Community of Maximum Residue Limits (MRLs) for Residues of Veterinary Medicinal Products in Foodstuffs of Animal Origin*, Volume VI, Commission of the European Communities (1991).

[7] J.O. Westgard, P.H. Petersen and D.A. Wiebe, *Clin. Chem.* **37**, 656–661 (1991).

[*Received for review 2 March 1994; revised manuscript received 13 June 1994*]

Appendix: ISO Analytical Definitions [2]

Accuracy. The closeness of agreement between the test result and the accepted reference value. Note — the term accuracy, when applied to a set of observed values, describes a combination of random components and a common systematic error or bias component.

Bias. The difference between the expectation of the test results and an accepted reference value. Note — bias is a systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value.

Trueness. The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value. Note — the measure of trueness is usually expressed in terms of bias.

Precision. The closeness of agreement between independent test results obtained under prescribed conditions. Notes — (1) precision depends only on the distribution of random errors and does not relate to the true value or the specified value. (2) The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Higher imprecision is reflected by a larger standard deviation. (3) 'Independent test results' means results obtained in a manner not influenced by any previous result on the same or similar material.

Repeatability. Precision under repeatability conditions.

Repeatability conditions. Conditions where independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time.

Reproducibility. Precision under reproducibility conditions.

Reproducibility conditions. Conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment.

Intermediate precision conditions. The M-factor different intermediate precision conditions ($M = 1$, 2 or 3) are: (a) $M = 1$ where only one of the three factors (operator, equipment or time) is different, or where the equipment is recalibrated between successive determinations; (b) $M = 2$ where two of the three factors are different; and (c) $M = 3$ where all three factors are different between successive determinations.