Journal of Pharmaceutical and Biomedical Analysis
17 (1998) 193–218

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Review article

Validation of bioanalytical chromatographic methods

C. Hartmann [1,a], J. Smeyers-Verbeke [a], D.L. Massart [a,*], R.D. McDowall [b]

[a] *Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium*
[b] *Department of Chemistry, University of Surrey, Guildford, Surrey GU2 5XH, UK*

Received 5 May 1997; received in revised form 6 August 1997

Abstract

A strategy is discussed for the validation of chromatographic methods that are developed to quantify drugs in biological matrices. Both the validation terminology and the hypothesis testing are briefly reviewed. The emphasis is on the design of the experiments required to allow a reliable conclusion about acceptance or rejection of the bioanalytical method. In particular, it is explained how to evaluate the calibration line, devise experiments to estimate precision and bias and how to determine the stability of the analyte between the time of the sample collection and the analysis of the processed sample. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Bioanalytical method validation; Experimental design; Hypothesis testing; Bias; Precision

1. Introduction

Owing to increasing interdependence among the countries during the last few years it has become necessary for the results of many analytical methods to be acceptable internationally. Consequently, to assure a common (minimum) level of quality, the need for and use of validated methods has increased. Bioanalysis is one of the branches of analytical science that requires method validation: for example without proving that the results of a bioequivalence study are based on a validated method, such a study is nowadays meaningless and will not be recognised by health official authorities. For many bioanalytical chemists it is, however, not obvious how to perform a reliable method validation and how the requirements for a regulatory submission can be reached in an economical way which is still statistically correct.

In 1990, a conference was held in Washington to discuss what a validation of bioanalytical methods should consist of, i.e. which analytical parameters (bias, precision, etc.) need to be documented to validate a method employed in bioavailability, bioequivalence and pharmacokinetic studies in man and animals. Moreover for some of these parameters minimum acceptance

[*] Corresponding author.

[1] Current address. Novartis Pharma AG, Technical R&D, Chemical and Analytical Development, Analytical Research and Development, 4002 Basle, Switzerland.

PII S0731-7085(97)00198-2

requirements were given. Although this was a good start to standardise validation in bioanalysis, several of the recommendations given during this meeting and gathered in a conference report [1] are subject to critique. The critique concerns both the usefulness of some of the recommendations and the lack of advice for the practical execution of a validation study.

Hartmann et al. [2] showed that with acceptance criteria for systematic and random error (bias and precision) and an experimental design as specified in the Washington conference report [1], there is a relatively large probability to conclude that the bias is not acceptable although the true bias fulfils the minimum requirements. Even worse, the probability is also large to adopt methods for the routine analysis that are in fact not suitable for their intended purposes. During the up-date meeting of this Washington conference in Munich in 1994, it was moreover pointed out that the indicated quality control requirements are not satisfactory [3], which is a confirmation of the work performed by Kringle [4]. Furthermore, Hooper [5] illustrated by means of the specificity requirements of the Washington conference report [1], that there is in general a need for acceptance requirements which are based on statistical considerations. Otherwise it is not possible to deduce from the validation results a reliable conclusion on the acceptability (or non-conformance) of a bioanalytical method and to guarantee a certain level of quality for the methods that fulfil the requirements in the Conference report.

An important shortcoming of the Washington conference report and most other available guidelines is, that not enough attention is paid to the experimental designs to be used. It is therefore the aim of this paper to provide information to assist an analyst in setting-up the validation experiments and, at the same time, to provide guidance for a sound statistical data evaluation. Since method validation is too broad a field to be covered accurately by a single paper, we will mainly focus on the validation of quantitative chromatographic methods with conventional detection systems used in bioanalysis. (Additional validation activities can, for example, be required when liquid chromatographic systems are coupled with mass spectrometers, LC-MS(/MS).) Special interest will be given to the experimental designs for the determination of bias and precision, the evaluation of the calibration model and stability studies. It is assumed that pure analytical reference standards for the analyte to be determined are available, an assumption which influences the bias evaluation, as discussed below in more detail.

2. Terminology

Before discussing how to carry out the validation experiments, it seems important to stress that the validation in bioanalysis should not be considered as an isolated field. Work on method validation is going on in different analytical fields. Each of them has of course its own characteristics and problems. However, as it has already been pointed out in the analysis of the Washington conference report [2], it does not seem promising to develop, for example, a validation terminology to be used in bioanalysis alone, without reference to other fields.

A harmonised validation terminology is the minimum basis required for a discussion between scientists of the same or different analytical fields. A consensus on a common terminology for all analytical fields is therefore required. For the moment it is not yet possible to propose a validation terminology which is also in agreement with the recommendations of important international organisations such as ISO (International Organization for Standardization), IUPAC (International Union of Pure and Applied Chemistry) and AOAC (Association of Official Analytical Chemists), since differences exist between their documents or draft documents.

For the validation of pharmaceutical drug formulations the discussion on a consensus terminology is relatively advanced. It is suggested to follow in general the proposal elaborated for the validation of drug formulations by the joint initiative of the pharmaceutical industry and the regulatory agencies of the three major regulatory authorities (the European Union, the USA and Japan), the International Conference on Harmonisation (ICH). In this way, at least the pharma-

Table 1
Performance characteristics to be considered during the validation of a quantitative method in bioanalysis

Method parameters (adapted from [6])	Short description (based on [6])
Bias[d]	Systematic difference between the true method mean and the true (reference) value
Precision	Random error of the method
Repeatability	Precision measured under the best condition possible (short period, one analyst...)
Intermediate Precision[a]	Precision measure of the within-laboratory variation due to different days, analysts, equipments, etc.
(Robustness)	Capacity of a method to remain unaffected by small variations in the method parameters as could possibly occur during the normal use of the method (pH, mobile phase composition,...)
(Reproducibility)	Precision measure determined by interlaboratory studies
Specificity[b]	Ability to determine the analyte in presence of other compounds
Limit of detection[c]	Lowest sample concentration that can be detected
Limit of quantitation	Lowest sample concentration that can be quantified with suitable bias and precision
Linearity	Ability of the method to obtain test results which are proportional to the concentration in the sample
Range[e]	Concentration interval within the method has a documented suitable performance
Stability	Absence of an influence of time on the concentration of the analyte in a sample

Remarks to Table 1 [6]:

[a] In cases where the reproducibility has been performed, intermediate precision is not needed.

[b] Lack of specificity of an analytical procedure could be compensated by other supporting analytical procedure(s).

[c] May be needed in some cases.

Our remarks to Table 1:

[d] Accuracy: difference of individual values from the 'true' or 'assigned' or 'accepted' value; Trueness: difference of an average for a group of individual values from the 'true' or 'assigned' or 'accepted' value; Bias: 'long term' or expected difference from an average of many groups of individual values from the 'true' or 'assigned' or 'accepted' value [7].

[e] Is defined as the concentration range in the measurement samples and not as the concentration range of the calibration standards.

ceutical industry would apply the same definitions and speak the same validation language. An adapted version of this terminology is given in Table 1.

In Table 1 we have adopted the AOAC terminology instead of that used by ICH for the description of systematic error. Although AOAC and ISO do not agree on the exact definition of 'accuracy', they both agree that accuracy is the combination of systematic and random error components, whereas the estimate of the pure systematic error should be indicated with 'bias' [7,8]. It seems preferable not to use at all the term 'accuracy' [9] or to consider, as AOAC does, the terms 'accuracy' and also the less frequently used term 'trueness' as estimates of the bias [7] obtained under defined conditions, in the same way that 'repeatability' and 'intermediate precision' are measures of 'precision' obtained under specific conditions [7,8].

Discussions are going on in the field of pharmaceutical analysis to additionally evaluate the 'robustness' of a method. This becomes important when one laboratory develops methods, which are then transferred to other laboratories, as described, for example, by Brooks and Weinfeld [10], but also when a method will be used over a longer period of time and small changes in the application of the method are likely to occur. A systematic investigation of parameters which need to be carefully controlled (= robustness testing) reduces the probability of having problems during the routine application of the method. It is therefore, also in bioanalysis, useful to evaluate the robustness in such cases.

The term 'reproducibility' is reserved [6–9] for interlaboratory studies. For completeness, 'reproducibility' is still mentioned in Table 1 as in a former draft of the ICH [11]. However, in bioanalysis it does not seem necessary to perform systematic interlaboratory studies when an intermediate precision measure has been evaluated. According to ISO [8], intermediate precision measures can further be classified. Depending on the

Table 2
Errors related to the hypothesis that the analytical method is not biased

Real situation	Decision	
	Test method accepted, since considered not biased	Test method rejected, since considered biased
There is no bias	Correct decision Probability = $1-\alpha$ (Confidence level)	**Type I error** Probability = α (Significance level)
There is a bias	**Type II error** Probability = β	Correct decision Probability = $1-\beta$ (Power)

factors investigated, specific M-factor different intermediate precision measures are evaluated. A one-factor different intermediate precision measure (i.e. $M = 1$) is, for example, determined when the precision is measured over a longer time period by one operator with only one equipment. This measure is called 'time-different intermediate precision'.

In bioanalysis, it is suggested to add 'stability' to the ICH list of the parameters which are to be evaluated. The stability of an analyte in biological matrices is a much more critical factor than it is for pharmaceutical drug formulations and this fact is in practice more and more taken into account. Notice that the term stability is also specifically considered in the validation strategy for bioanalytical methods which is currently prepared by the French group SFSTP (Société Française des Sciences et Techniques Pharmaceutiques).

To avoid confusion it should be noted that ICH makes no difference between the terms 'selectivity' and 'specificity' [6], an approach which has, however, been subject to the critique of Vessman [12].

3. Hypothesis tests

The validation experiments are usually analysed statistically and this eventually requires hypothesis tests. It is not as evident as it might seem, which hypothesis should be tested. There are two fundamentally different approaches that are possible. Let us consider as an example the evaluation of the bias. (The reason for including this section in the paper is that many bioanalysts come from a biological background with little theoretical knowledge of statistics who are often not aware of the reliability of a conclusion taken based on the significance of a given statistical test.)

One can follow the classical approach of point hypothesis testing, as, for example, applied in the well documented validation approach for the analysis of pharmaceutical drug formulations by the French group SFSTP [13]. The null hypothesis tested is that there is no bias, i.e. that the bias is zero. A method is declared free of bias and accepted when there is not enough statistical evidence that the method bias is significantly different from zero. This means that one only pays attention to the probability to wrongly decide that there is a bias when in fact there is none (α-error, Table 2). However, it is more and more frequently recognised that a small bias should not lead to rejection of the method. This is certainly true for bioanalysis.

According to the Washington conference report [1], the maximum bias for a method should be $\leq \pm 15\%$ (and at the limit of quantitation $\leq \pm 20\%$). How such acceptance limits should be taken into account during the data evaluation is, however, not specified. Frequently, one tests the hypothesis that there is no bias and when the test indicates a significant bias one simply compares the estimated bias with the limit value, e.g. $\pm 15\%$ ($\pm 20\%$). The probability of obtaining a bias estimate $\leq \pm 15\%$ has been calculated for several combinations of true bias and precision [2]. This study indicated that an evaluation based on this approach is not reliable. The latter therefore cannot be recommended. Hypothesis testing as described in the next paragraph should be preferred.

The other approach uses statistics as they are applied for the evaluation of bioequivalence studies. As one of the major purposes of bioanalysis is to carry out pharmacokinetics, one might as well apply the same statistical approach. This was recently studied for analytical method validation [14]. The hypothesis testing is based on the philosophy that the β-error needs to be fixed. In terms of the evaluation of the bias, the β-error is the probability of concluding that there is no bias, when in reality the method is biased (Table 2). The β-error clearly is at least as important as the α-error in a method validation context. To fix the β-error, the bias estimate must be tested against the acceptance limits. This can be done by performing interval hypothesis testing and reformulating in an appropriate way the (classical) null and alternative hypotheses [14]. A method then is accepted when the probability is satisfactory that the true method bias lies within specified acceptance limits (e.g. the ± 15% (± 20%) limits of the Washington conference report [1]). This second approach is now also preferred by the SFSTP group in their draft document on the validation of bioanalytical methods.

The decision on what hypothesis testing approach is preferred needs to be taken by consensus. For the moment point hypothesis testing is much more popular. The recommendations that follow will therefore mainly consider this classical way of hypothesis testing, although we consider this approach fundamentally less suitable. Some possibilities and consequences of applying interval hypothesis testing will, nevertheless, be mentioned in the following sections.

4. Strategy for validation

Validation can be defined as the process of documenting that the method under consideration is suitable for its intended purpose. In this definition two aspects should be stressed. The first is that any decision must be written down, i.e. one must be able to provide documented evidence that the conclusion is correct. In general one should have statistical reasons for the acceptance (or rejection) of a method. The second is, that the required extent of a validation must only go as far as it is needed for the goal of the application of the method, for example, depending on the stage of drug development [15–17]. This means that only the analytical parameters need to be validated that are of importance for the routine application. Only a minimal number of validation experiments are, for example, performed for the first toxicological trials, whereas for bioequivalence studies it is crucial both to exactly know the method performance and to reach tighter requirements than in the beginning of the method validation in order to limit the number of subjects required for a reliable conclusion. This documents that the extent of a validation is not only influenced by safety considerations but also by reasons of benefits and costs. As mentioned above, it makes sense, therefore, for example, to add stability to the list of Table 1, whereas it will not be necessary to document the reproducibility of a candidate method in many cases. Analogously, the imposed acceptance requirements should be based on what is needed for the intended application. Specifying minimum requirements either on general or on in-house consensus [16] can provide a guideline. Care has, however, to be taken by a bioanalyst not only to consider the imposed minimum requirements, but to make sure that the aims of the future study can be reached with a feasible workload.

A full method validation requires a rather high workload and should therefore only start when promising results are obtained for the explorative validation performed during the early drug development phase, i.e. when the preliminary experiments indicate that the required quality (for precision, range, etc., see below) will be reached [15]. The latter experiments are very important to give insight in the possibilities and limitations of the analytical method. It is, for example, recommended to verify that a sufficient resolution can be reached between closely eluting substances. The main focus during this early stage of the validation should then be on the study of the calibration range in order to provide the basis to select an appropriate calibration model and to define the concentration range within which it is likely that the method will satisfy the acceptance requirements. It is suggested to consider rather many calibration levels with only a limited number of replicates and to mainly visually evaluate the experimental data (see Section 5.5.). A first

Table 3
Simplified experimental set-up for an ideal validation situation of quantitative chromatographic methods to be used in one laboratory (for details see text)

Parameters to be validated	Experimental approach for an ideal situation
Calibration model/linearity	One sequence of all experiments Four evenly spread levels × nine independent replicates [26]
Precision[a]: repeatability and (time different) intermediate precision	Three concentrations Eight days × two independent replicates [8]
Bias[a]	Evaluation from experiments performed to document the method precision
Specificity	Combined with the bias evaluation
Limit of quantitation (LOQ)	First guess from linearity experiments, confirmation by precision and bias experiments
(Assay) Range	Deduced from the precision and bias experiments
Stability	Two concentrations (top and bottom of bias and precision) Approximately four test conditions About six replicates for ‘fresh’ and ‘stability’ samples
Robustness	Recommended for methods for long-term use and/or interlaboratory use. Depending on the number of factors to be evaluated, fractional factorial, or Plackett-Burman design [59,60].

[a] Forms basis for on-going quality assessment of precision and bias as the method is applied in routine.

(coarse) estimate of precision and the quantitation limit(s) of the analytical method (see also below) can be deduced from replicate measurements of the calibration standards and the results of these experiments can be used to define the calibration range to be considered further (see, for example, the SFSTP approach for bioanalytical methods).

Based on the results obtained during the exploratory validation, it then is recommended to follow a certain validation strategy to take all the performance criteria appropriately into account and to document during the full validation that the expectation (e.g. about the calibration model) and the specific acceptance requirements on the method (e.g. on the precision) are fulfilled. The evaluation of all these experiments allows then to exactly formulate the analytical procedure as it is to be applied in routine, i.e. not only to indicate the preparation of the reference standards and reagents, the design and the formulae of the calibration function but also to specify how long and at which conditions a sample may be stored prior to analysis. For the validation, the following sequence of experiments is suggested (Fig. 1). This approach tries to make maximum use of resources in the laboratory. If a method fails to meet the acceptance criteria, using this scheme the analytical procedure will be rejected quickly. It should, however, be noted that a specific experimental situation can require a deviation from the proposed sequence of the validation experiments.

A simplified summary of the experiments of such a full method validation is presented in Table 3. Details of the proposed statistically sound designs as well as the evaluation of the experiments performed are discussed below. At first sight the number of experiments recommended, might seem to be unusually high. It should, however, be realised that efforts made during the validation phase will be recompensed while running the method routinely. On the one hand, more knowledge is available about the performance of the method and critical steps of the procedure can be improved before applying the method in routine. On the other hand, bottlenecks of the analytical method that are realised during the thorough validation can efficiently be solved, since the person having the broadest knowledge of the procedure, the developer of the method, is confronted with the problems and not the ana-

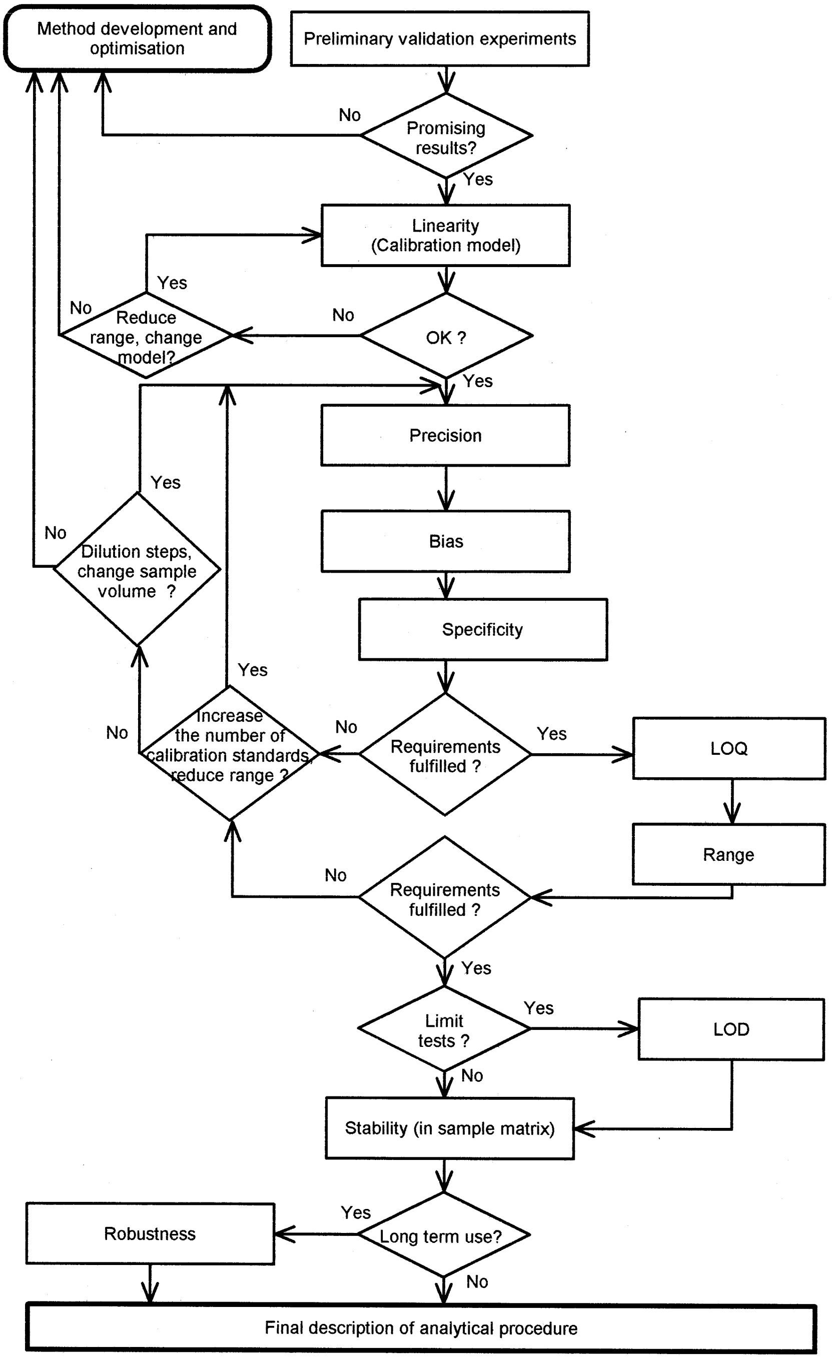
Method development and optimisation
Preliminary validation experiments
No
Promising results?
Yes
Linearity (Calibration model)
Yes
No
Reduce range, change model?
No
OK ?
Yes
Precision
Yes
No
Dilution steps, change sample volume ?
Bias
Specificity
Yes
No
Increase the number of calibration standards, reduce range ?
No
Requirements fulfilled ?
Yes
LOQ
Range
No
Requirements fulfilled ?
Yes
Yes
Limit tests ?
LOD
No
Stability (in sample matrix)
Yes
Robustness
Long term use?
No
Final description of analytical procedure

Fig. 1. Bioanalytical method validation. Strategy for the full validation of quantitative chromatographic methods to be used in one laboratory (for details see text).

lysts performing the routine analysis who might be unfamiliar with possible alternatives to the different steps of the analytical procedure.

5. Linearity and selection of the calibration model

5.1. Definition

The term linearity is generally accepted, but not very clear. According to the ICH-definition ‘the linearity of an analytical procedure is its ability (within a given range) to obtain test results which are directly proportional to the concentration (amount) of analyte in the sample’ [6].

This definition can be interpreted in two ways. The linearity study can be considered as being part of the bias investigation, if it is understood as being the linear relationship between found and known concentrations, e.g. [15]. The linearity can, however, also be considered as being the study of the calibration line, as discussed in the validation methodology of the ICH [18]. The former approach should be included in the bias study as such. If the bias is acceptable, the linearity in that sense is then also acceptable. It is more useful to limit the evaluation of the linearity to the verification of an assumed calibration model. We will follow this second interpretation.

5.2. Model selection

As mentioned above, the selection of the calibration model should be done during the early method validation. It must be pointed out that the selection of the calibration model is crucial for the quality of bias and precision that can be reached with a given method during the routine application [19–22]. According to the Washington conference report [1] the response function should be represented either by a graphical technique or by an algorithm. Since the latter can be evaluated more extensively, namely visually and statistically, the evaluation of a mathematical function will be discussed further.

5.3. Problems in bioanalysis

For many bioanalytical methods the concentration ranges are usually rather broad, e.g. 1–100, 1–1000 [23] or even wider. In broad calibration ranges even relatively small deviations from an assumed model, e.g. a straight line, can lead to substantial errors in the predicted concentrations at the extremes of the calibration range. The analysis is also complicated, because there are often interactions with matrix components. The calibration lines are, therefore, usually prepared in the sample matrix [23]. It must be mentioned that the precision of the calibration line then in general is worse than the precision of a calibration line based on aqueous standards and that this affects the time-different intermediate precision of the method [24]. Nevertheless, accepted practice in bioanalysis is to prepare standard curves in biological matrix and to compare like with like.

5.4. Sequence of the linearity evaluation

The principles of evaluating a calibration function in bioanalysis are similar to other fields and are for all models alike. Both the behaviour of the variance and the goodness-of-fit of the model must be documented. The following evaluation steps are recommended as soon as one has an indication that the precision at the intended concentration levels will be acceptable [3]: (1) selection of the design during the exploratory validation or the early phase of the full method validation; (2) performing the experiments; (3) visual evaluation of the selected model; (4) purging the data from outliers; (5) defining the behaviour of the variance over the calibration range; (6) statistically testing the model fit.

Since the selection of the design is influenced by the needs for a reliable evaluation, we will start with the recommendations for the data evaluation (see also Fig. 2) and discuss the experimental designs in the context of the statistical testing.

5.5. Visual evaluation

Frequently the visual evaluation is merely based on the classical regression plot alone. However, to

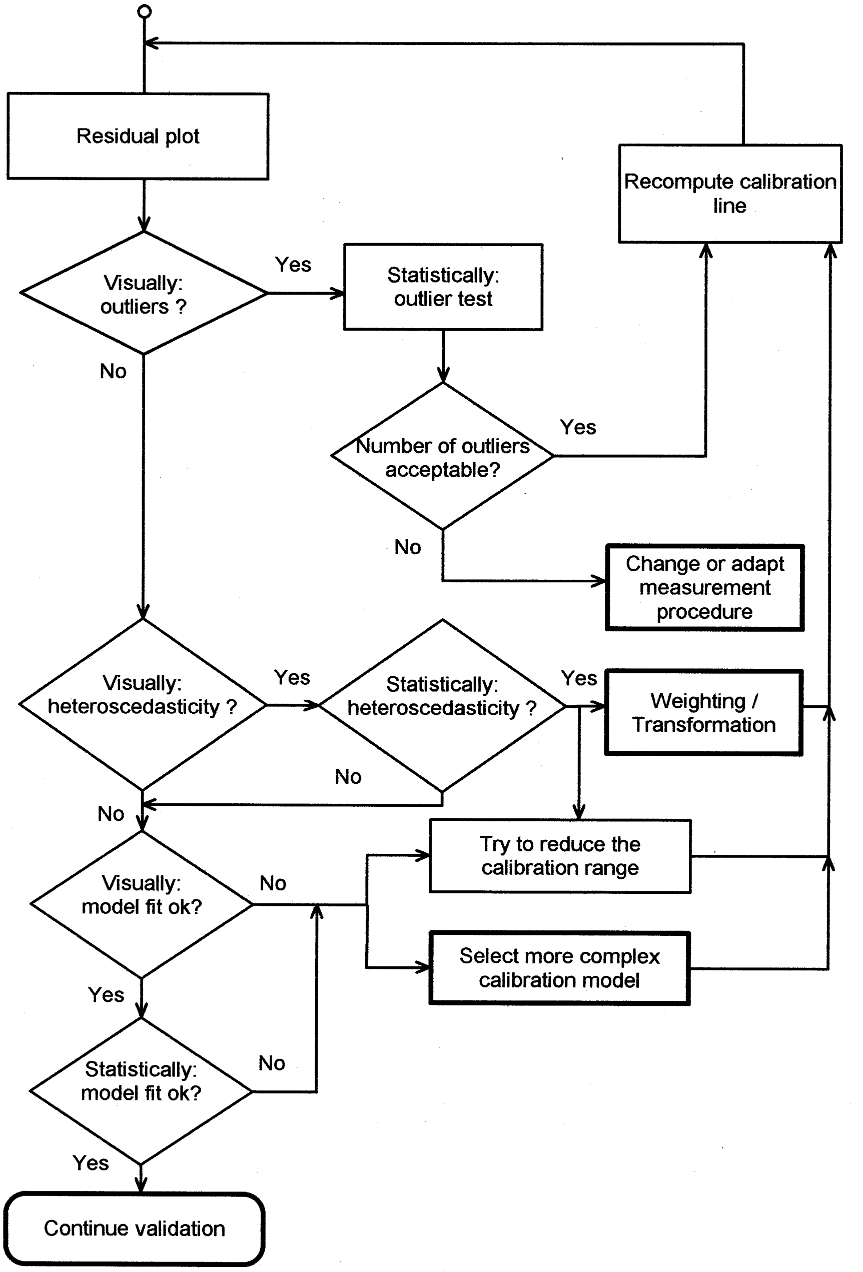
Residual plot
Visually: outliers ?
Yes
No
Statistically: outlier test
Number of outliers acceptable?
Yes
No
Recompute calibration line
Change or adapt measurement procedure
Visually: heteroscedasticity ?
Yes
Statistically: heteroscedasticity ?
Yes
No
No
Weighting / Transformation
Try to reduce the calibration range
Visually: model fit ok?
No
Yes
Select more complex calibration model
Statistically: model fit ok?
No
Yes
Continue validation

Fig. 2. Flowchart for the evaluation of the calibration function (adapted from reference [3]).

detect more easily deviations from the assumed model one should preferably also evaluate the residual plot. (A residual is the difference between the measured and the predicted response for a given calibration standard.) Deviations from the expected homogeneous pattern of the residuals around zero can give hints of a lack-of-fit of the model or of outliers in the replicates at one level.

The residual plot also provides an indication how the measurement variance changes with concentration (= variance function) (see also references [3,25]).

5.6. Outliers

Most statistical tests to examine the behaviour of the variance and the goodness-of-fit of the assumed calibration model require that the data are normally distributed. This can hardly be tested, but is usually true when the data are not affected by outliers. For the regression analysis one must make a clear distinction between 'regression outliers', i.e. outliers with respect to the model, and outliers in the replicates at one concentration level.

It is expected that the regression outliers will be detected while testing the goodness-of-fit of the model. Only outliers in the replicates will therefore be discussed here, i.e. aberrant values in the group of measurements performed at one concentration level. A possible approach to detect outliers, recommended in a recent ISO document [8], is to apply the single and paired Grubbs' outlier tests. The removal of outliers is a two-step process. First one screens for values that are significant at $\alpha = 5\%$, so-called stragglers [8]. If a reason is found (e.g. a transcription or a spiking error), that does not affect the remaining measurements [25], the values may be corrected or discarded. These values should otherwise be retained for the calculations unless they are also significant at the 1% significance level, i.e. when they are considered as 'outliers'. More than four replicates are required to reliably detect stragglers by the single Grubbs outlier test and even more than six replicates are needed for the detection of pairs of stragglers [26].

Not only the elimination of possible outliers should carefully be considered, also the total number of outliers must be controlled. The frequency of outliers is informative to show if the method works well. Penninckx et al. suggested to accept no more than two stragglers during the evaluation of the calibration model [26]. One could also reason that when outliers occur in the calibration line, large errors are also likely to occur in the analysis of the samples and that therefore, one should not allow outliers at all. A general consensus on how to handle outliers would be useful. Anyway, if many values are outlying the reason should be looked for and the experiments should be repeated.

5.7. Behaviour of the variance

When the assumed calibration model has been re-calculated without the identified outliers, the variance over the calibration range should be evaluated, since the latter can influence the calculations required for the verification of the goodness-of-fit of the model. The simplest regression model is the usually applied ordinary least squares model, which is not forced through the origin. This (default) model is based on the assumption that the variance is constant. When this assumption is not fulfilled, the regression calculations become more complex (see below). It is therefore important to check that the measurement variance is constant (= homoscedasticity), i.e. that it does not change with concentration (= heteroscedasticity).

Many bioanalytical procedures are heteroscedastic but often with a constant relative standard deviation (R.S.D.), which means that the standard deviation proportionally increases with concentration. To detect this type of heteroscedasticity it is suggested to compare the variances at the highest and the lowest (non-zero) concentration level by means of a one-sided F-test ($\alpha = 5\%$) [3,26,27]. Heteroscedasticity then is detected more reliably by the F-test than by alternative tests such as the Cochran, Hartley or Bartlett's tests [26]. Another possibility for detecting heteroscedasticity would be to apply the randomisation test proposed by Penninckx et al. [26].

When heteroscedasticity is confirmed and the calibration range cannot be reduced, weighting factors must be determined which are inversely proportional to the variance at the given concentration level, e.g. weighting factors equal to $1/C$ or $1/C^2$, where C is the concentration [3,28]. Alternatively, the variance function must be evaluated [29] or the data can be transformed [3]. To evaluate if the weighting factors (or the transformation) are satisfactory it is recommended to evaluate visually the residual plot of the weighted (or the transformed) calibration line, before applying again the *F*-test on the weighted residuals (or on the residuals of the transformed data).

5.8. Goodness-of-fit

For the evaluation of the goodness-of-fit of a calibration model, one must first decide on the number of concentration levels and the number of replicates per level to be studied. For this evaluation it is from a statistical point of view best to concentrate the experiments at few levels and to have a high number of replicates per level as it has already been described for the detection of outliers, to document homoscedasticity of the measurements and to determine weighting factors [26]. It depends on the complexity of the calibration model how many levels and replicates are actually needed and which statistical tests are optimal. However, independent of the results of the exploratory method validation, several replicates should be performed. It should also be noticed that a design as described in the draft document on bioanalytical method validation by the SFSTP, with several calibration standards at the lower concentration levels but only a single level at a high concentration cannot be recommended, since this highest calibration standard has too large an influence on the least squares line and the statistical analysis. Once a model has been established during the early validation phase, the number of calibration standards can be reduced for the further validation experiments and the routine analysis (see Section 5.10.).

For most chromatographic methods it is expected that the data can be explained by a straight line relationship. From the point of view of the statistical evaluation (β-error considerations) it is best to consider only three (evenly spaced) concentration levels, even for calibration ranges of 1–1000. To provide a possibility for a preliminary evaluation of a second degree model when the simple straight line is not acceptable, it has been recommended to study four concentration levels [26]. Moreover it has been suggested to analyse nine replicates at each of these levels, which means that a total of 36 samples should be analysed [26] (see also Table 3). This is, of course, rather much and it would be good to have a consensus on how thorough one has to be. However, analysing only duplicates, e.g. as suggested by Lang and Bolton [30], cannot be recommended from practical and statistical points of view, since there are not enough data to detect reliably outliers and it might not be possible either to determine reliably weighting factors. Furthermore, it should also be stressed that the runs of the validation experiments should reflect the way the laboratory intends to use the method and then the analysis of many samples is not unusual.

After performing the experiments and the above described evaluation steps it has been proposed to verify that a second degree model does not fit the data better [26,27]. The latter can be done by demonstrating that the second order term is not significantly different from zero (two-sided *t*-test, $\alpha = 5\%$). A complementary or alternative evaluation of the goodness-of-fit of the linear first order model is based on the ANOVA (analysis of the variance) lack-of-fit test [25,26]. Although it is often done [24], it should be stressed that a calibration model cannot be validated by the correlation coefficient [32]. Notice also that it is not necessary that the straight line passes through zero, unless a single point calibration is intended for the routine application. The straight line model should therefore not be forced through zero.

For heteroscedastic measurement methods, alternative calibration models should be evaluated using the same original data but weighted or transformed (see Section 5.7.). The lack-of-fit of the model can again be studied with the ANOVA lack-of-fit test [3,25,26].

If the assumed straight line model is not acceptable over the whole intended range, one should evaluate whether it might be possible to reduce the calibration range (so that the assumed model might be applicable), since it is best to opt for the simplest model possible consistent with the Conference report [1]. Otherwise one can preliminary evaluate a second degree model when at least four concentration levels are represented. More complex models than second order should, however, be avoided because the parameters of such models are less precisely estimated (or the workload increases).

When from the exploratory validation or from the experiments described above, there is a fair indication that a second degree model is required, it is recommended to evaluate this as follows. Calibration standards should be analysed at least at four concentration levels. However, it has been recommended to consider then seven levels evenly distributed over the calibration range [26]. To keep the workload still feasibly low it is suggested to unevenly distribute the total number of observations over the levels and to perform nine observations at the extremes of the calibration range and at each of the five intermediate levels only six observations (i.e. a total of 48 separate samples) [26]. ANOVA can be performed to study lack-of-fit.

5.9. Linearity experiments

To reliably test a calibration model all calibration standards must be prepared and analysed independently (Table 3), since a representative estimate of the measurement variance is required. Care must also be taken to analyse all standards randomly and to perform the chromatographic analysis in the shortest time possible (one sequence, Table 3). Therefore, the often followed practice of using the calibration experiments to obtain an estimate of an intermediate precision (see reference [23]), unfortunately, cannot be recommended. Moreover the experiments should be performed by a single analyst, who is familiar with the analytical procedure. In practice this is usually the method developer or a member of the team who has developed the method.

Additionally to the above indicated experiments one should also analyse at least one matrix blank to confirm that the matrix blanks are pure and that there is no interfering impurity in the mobile phase nor in the reagents used. The results of these blanks should, however, only visually be evaluated and not be considered for the calculations (see Section 5.8.). Notice that blank samples should not only be run during the method validation but also during the routine analysis, e.g. a blank for each subject of a bioequivalence study (see also Section 7.).

5.10. Design for routine application

5.10.1. Model

Once the model is validated, the number and the distribution of the calibration standards often need to be adapted to reduce the workload for the routine analysis but to still reach the acceptance requirements for bias and precision [3,22,24].

According to the Washington conference report, a minimum of five to eight calibration standards should be considered. Hill et al. [22] suggested the following default distribution of eight calibration standards ([LOQ (Limit of Quantitation, see Section 8.) = 1, top standard = 100]). The second calibration standard close to the lower extreme of the range provides a large probability to quantify almost all samples measured [3].

0, LOQ, concentration twice the LOQ, 5, 10, 20, 40, 80, 100

Notice that the concentration zero (a blank) is only used to visually verify the purity of the reagents but that its result is not considered for the calculations. From a statistical point of view there is no reason to have so many levels, once the calibration line has been shown to be linear. The design with duplicate analyses of four concentrations (and one blank):

$0, 2 \times \text{LOQ}, 2 \times 10, 2 \times 50\ 2 \times 100$ would also allow a good estimation of the calibration line, be more practical and the replicates can be used to detect a gross outlier if this were to be present (Section 5.10.2.).

It should be noted that the calibration design selected for the routine application must then be applied for all further validation experiments. Otherwise, the estimates for bias and precision (see Section 6.) are not realistic, since the calibration line can influence the intermediate precision estimates and the bias estimates [24].

5.10.2. Outliers

During the routine application usually no or a few replicates of the calibration standards are analysed. Consequently it is more difficult to trace outliers and to decide on the importance of a measurement response lying far from the calculated line. Any test to be used must also be simple and quick to be practicable.

When replicates of the calibration standards are analysed it is possible to evaluate the maximum difference between the replicates against the so-called repeatability limit [33]. For duplicate analysis, this limit is defined as 2.8 times the estimated repeatability standard deviation. If the difference is larger, the detailed recommendations of ISO can be followed for the further evaluation [33].

It does not appear useful to simply evaluate the percentage difference of each calibration point to the regression line separately [34]. Depending on the precision of the measurements, the deviation of the measured from the predicted response can be rather high even for a perfect method performance. Besides, a relatively high measurement response for one standard may correct for a relatively low one of another so that the calibration line itself is not necessarily affected.

A possible approach could be to compare the regression coefficients of the usually applied (weighted) least squares calculations with the ones of a robust regression line, e.g. Least Median of Squares [35]. The term ‘robust’ means that the regression line is less affected by outliers. Criteria for acceptance have not been developed for chromatographic methods at this time. A possibility is, however, to follow a strategy similar to the one worked out for AAS (atomic absorption spectrometric) methods [36]. The quality of a calibration line is judged by the quality coefficient (QC) and when this quality is acceptable (in AAS, e.g. a QC equal to 5%), the line is accepted. Otherwise the Least Median of Squares approach is applied to identify possible outliers and after removal of these aberrant values ordinary least squares is applied on the remaining data (= reweighted least squares).

Another possible approach, which seems both practicable and acceptable, is to evaluate the absence of outliers and the appropriateness of a daily calibration line indirectly by the results obtained with quality control samples. Care must, however, be taken to use a control procedure based on the quality of the analytical procedure and not simply to follow the approach outlined in the Washington conference report [1]. The latter is not sensitive enough to detect a performance change of the analytical procedure because it is not related to the quality of a given method [3,4,37]. This shortcoming is remedied by the approach that was recently discussed by Selinger [38]. The number of independently prepared quality control samples needed and their acceptance criteria are determined depending on the sample size of the batch and on the percentage defectives considered acceptable.

Again, outliers should not be rejected blindly without examining whether an assignable cause for the outlier(s) can be found and the total number of rejected outliers needs to be controlled, too.

5.11. Null hypotheses tested

All recommendations given in this section followed the usual approach of point hypothesis testing. The null hypothesis of the outlier tests is that the suspected value is part of the normal variation of the data, i.e. that there is no (non-random) effect. It makes sense to focus on the risk not to reject too many ‘good’ values, since otherwise the measurement variance is underestimated. For outlier tests this null hypothesis seems therefore to be appropriate.

Usually one would like to demonstrate that the data are homoscedastic, although from a study performed by Horwitz et al. [39] an effect of the concentration on the precision is to be expected. Moreover by testing the homoscedasticity of the data as described, the risk is not controlled to conclude that the variance is constant when in fact there is heteroscedasticity. Despite this critique, this approach is acceptable, since a small deviation from homogeneity of the variances will hardly affect the ANOVA lack-of-fit test when the sample sizes of the groups to be pooled are nearly equal [40].

The use of the point hypothesis *F*-test (in the ANOVA lack-of-fit) in the evaluation of the linearity should, however, be questioned. Here, it would probably be more appropriate to base one’s conclusion not only on the statistical significance but to consider also the practical relevance of a detected deviation. On the one hand, for many determinations a sufficient approximation to the true sample concentration may be reached (i.e. an acceptable bias) even when there is, for example, a slight curvature in an assumed linear first order model. On the other hand, a poor measurement precision can mask an important lack-of-fit of the assumed calibration model. Consequently one should rather take into account acceptance limits for a departure from the calibration model. This can be evaluated in several ways. One can, for example, estimate the bias introduced by the regression model by comparing the concentration of the calibration standards with those predicted from the model. If the deviation still falls within acceptable limits, one would decide to adopt the model anyway. For the evaluation of the calibration model, it seems therefore more useful to apply the above mentioned approach of interval hypothesis testing.

6. Precision and bias

When the results of the linearity study are acceptable one can start to evaluate the precision and the bias. Both are important performance characteristics used to decide on the acceptance of a method. Consequently, the minimal requirements imposed by regulatory agencies concern mainly these two parameters. The Washington conference report [1] also only specifies limits for precision and bias.

Sometimes the recovery is considered as an additional parameter as recovery can be measured in the same experiments as precision and bias. In the draft of the SFSTP on bioanalytical methods it is, for example, stated that recovery experiments are important during the exploratory validation and might indicate that further method development is required. It is, however, not obvious which acceptance requirements should be imposed. In general, it is most important to reach a reproducible recovery, which is high enough to satisfy the requirements of detecting and/or quantifying low sample concentrations, even when the recovery itself is low [5]. There seems therefore no need to require a minimum recovery of, e.g. 70% as used by Braggio et al. [15]. Eventually the recovery is reflected by the bias (and the specificity) of the method and, therefore, during the full validation, it is considered sufficient to evaluate the bias.

6.1. Sequence of the evaluation

The steps for the evaluation of precision and bias data are similar to the ones described in the section on linearity. After setting up the experimental design and carrying out the experiments, it is recommended to visually evaluate the data scatter. Depending on the design selected for the bias evaluation (details see Section 6.4.) a more sophisticated visual evaluation can, for example, be based on box plots [41], when several replicates are available at a given concentration level (see Fig. 3), or on a plot proposed by Bland and Altman [42] when several samples are measured with each of two methods (see Section 6.4.3.). After the statistical evaluation, the results for precision and bias must be compared with the imposed acceptance requirements. The validation is either continued or the test method is rejected since it requires further development.

In this section again the usually applied point null hypothesis will be considered. However, it is recommended to consider also the β-error when setting up the experimental designs to reduce the risk to accept methods that do in fact not fulfil the imposed quality requirements.

6.2. Precision estimates

As the concentration ranges are usually rather wide in bioanalysis, it is necessary to document the precision, at least, at three concentration levels, namely near the lowest level where quantification is required, the median level (or simply the middle of the range) and near the highest concentration expected in the measurement samples. The extent of the evaluation of the different precision estimates (repeatability, M-factor—different intermediate precision measures and reproducibility) depends on the purpose of the method. For many bioequivalence studies it is sufficient to

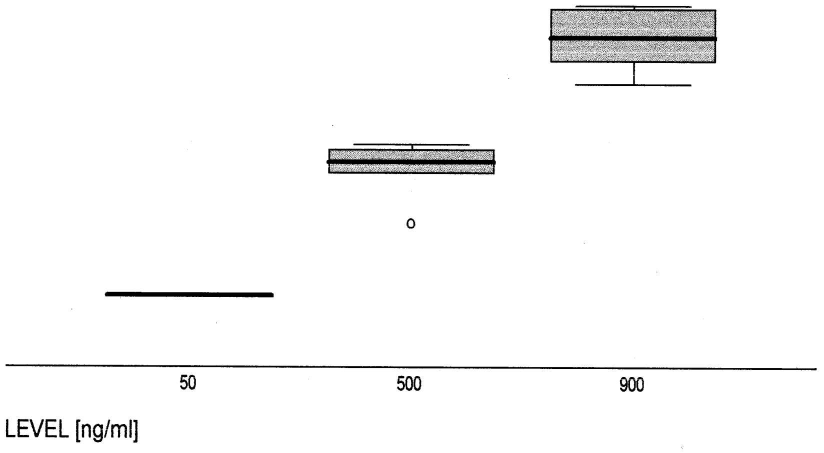

Fig. 3. Examples of box plots. Several replicates ($n = 6$) are performed at each of the studied three concentration levels. There is a tendency of an increase of the spread of the measurement responses (heteroscedasticity) and an outlier in the data (○) at the concentration level 500 ng ml^{-1}.

consider the repeatability and the time-different intermediate precision. In these situations it is suggested to analyse during eight days at least two replicates from the same sample pool (Table 3). The data of each concentration level are evaluated by a one-way ANOVA [24,31]. The suggested 8 days × two replicates are in line with the ISO recommendations [8]. This design is more balanced than the more classical designs with a comparable total number of observations (e.g. six replicates × 3 days), i.e. the degrees of freedom are similar for both the repeatability and the time-different intermediate precision (seven for days, eight for repeatability in the eight × two design and two for days, 15 for repeatability in the six × three design). The 8 days × two replicates design requires more experiments and work than the five determinations per concentration level suggested in the Washington conference report [1]. However, increasing the quality of the validation experiments helps to avoid problems later during the routine application of the method. Furthermore, for methods that do not reach the imposed quality requirements the eight × two design allows to identify more reliably the precision component (e.g. repeatability and/or time-different component) that is responsible for the unacceptable overall precision [24].

If one wishes, more factors can be evaluated, such as the effect of the sample preparation step, different operators, etc. It is then recommended to perform the experiments according to a fully nested design [8]. Such a design is, for example, performed when each of three operators analyses over 4 days two replicates each. A practical application of a fully nested design has, for example, been described by Yang et al. [43].

The reproducibility evaluation of the method should (when required) only start after completion of the full validation in one laboratory and preferably robustness tests should be performed first (see Section 12.). Reproducibility is determined in interlaboratory studies. These studies should be performed according to available guidelines, such as those of ISO [8].

6.3. Precision experiments

Which samples should be measured depends mainly on their availability. To obtain realistic estimates of the precision (and also of the bias, see Section 6.4.) it is, however, always necessary to analyse independently prepared samples that were taken through the whole procedure and to perform the measurements in a random order.

Table 4
Overall α*-error of an experimental design with several statistically independent tests each applied at $\alpha = 5\%$

Overall α*-error in %					
Three levels = 3 tests	Four levels = 4 tests	Five levels = 5 tests	Six levels = 6 tests	Seven levels = 7 tests	Eight levels = 8 tests
14.3	18.5	22.6	26.5	30.2	33.6

6.4. Approaches for the bias evaluation

Several situations are possible which influence the experimental set-up for the evaluation of the bias. Notice that a bias estimate calculated from the mean of observations analysed during only one day can be strongly influenced by the daily calibration line. Consequently, the latter should be avoided and the bias should be estimated from the average concentration of samples that have been analysed with several calibration lines.

6.4.1. Evaluation of spiked samples

When blank matrix is available which can be homogeneously mixed with the reference standard, one can prepare relatively large pools of spiked samples at the three above mentioned concentration levels and combine the bias estimation with the experiments performed for precision (Table 3).

The bias can statistically be evaluated by means of independent t-tests or by regression analysis. Both approaches have their advantages and limitations for a given experimental design.

6.4.1.1. t-Tests. The most important drawback of a data evaluation by multiple independent t-tests is that the α-error which is related to the conclusion about acceptance of the test method increases. For independent tests this increase of the effective α-level, α*, can be approximated by:

$$\alpha * \approx [1-(1-\alpha)^k] \quad (1)$$

with k, the number of levels studied and α, the significance level of the individual tests.

Suppose, for example, that one intends to evaluate the bias at $k = 8$ levels with independent t-tests each at $\alpha = 5\%$. One then runs a risk of concluding in more than one third of all cases that the bias is not acceptable at one of the levels studied when in fact there is no real bias (overall $\alpha^* = 33.6\%$; Table 4). This increase of the effective α-error makes the decision about acceptability of a method more complex. Suppose that for a method satisfactory results are obtained at all but one concentration level. When this level is one of the boundaries of the investigated range, one will probably conclude that the method is not acceptable at this lowest or highest concentration level studied and that the (assay) range (see Section 9.) must be reduced. However, when tests at intermediate concentration levels indicate a significant bias, it will often not be concluded that the method performance is only satisfactory at certain concentration levels and the significant test result will be ignored. Without increasing the number of samples to be analysed, two approaches are possible to avoid these problems. The significance levels at which each of the individual tests are performed can be reduced by the so-called Bonferroni correction [44] or the number of levels and therefore the number of tests performed must be limited. Since not only the α-error increases when several tests are applied but also the probability to accept a method with an in fact not acceptable performance (= β-error; Table 2), the second approach should be preferred.

The increase of the error-levels is substantial when more than three to four tests are performed to come to an overall conclusion [45]. Consequently, the application of more than four independent t-tests should be avoided. When only one (or a few) of the tests performed indicate(s) a significant bias, it is suggested to repeat the independent t-tests at the adapted α-levels (Bonferroni correction). This allows to verify whether the results still are significant and the test method therefore probably biased or whether the significant test result might be due to the increased overall α*-error.

The sample size needed at each test level depends on the precision, on the magnitude of bias that needs to be detected (i.e. on the acceptance limit for the bias) and on the selected acceptable α- and β-errors. When one individual test is performed at $\alpha = 5\%$, the sample sizes needed can be read from the ISO graphs [46]. To detect, for example, a bias twice as large as the known standard deviation by means of one two-sided independent t-test ($\alpha = 5\%$, $\beta = 10\%$), three replicates need to be analysed. This number increases to ten when the measurement standard deviation and the bias to be detected are of the same magnitude.

6.4.1.2. Regression analysis. For the bias evaluation by regression analysis it is also recommended to limit the number of concentration levels studied and to have many replicates per level, although this approach is less affected by the number of levels studied [47]. When the regression line between the predicted concentrations and the known concentrations used to spike blank matrix is calculated, a straight line relationship is expected with a slope not significantly different from 1 and an intercept not significantly different from 0. Visually, one should evaluate the regression plot as well as the residual plot. The statistical evaluation of the coefficients of the ordinary least squares line can give valuable information, since a bias in the slope (= proportional bias) is frequently due to matrix effects and a bias in the intercept (= constant bias) can usually be ascribed to blank problems. For the detection of a proportional bias the number of data points required to detect a given true bias with given α- and β-errors, depends on the ratio of bias and precision and increases in an inversely proportional way with the variance of the sample concentrations in the measurement range [47]. The relatively large ranges usually considered in bioanalysis [23] are therefore an advantage in this case. This means also that for comparable ranges and precision the workload to detect a given bias is similar (Table 5a). A drawback of the regression analysis is, however, that the probability of detecting a constant bias is also affected by the mean concentration of the samples analysed (Table 5b) [47]. As shown in Table 5, the sample sizes required for the detection of a constant bias is larger than that for a similar proportional bias. The differences become important when the mean of the range is shifted from zero. It has, therefore, been suggested to perform additionally a t-test at the extremes of the measurement range to verify the absence of bias [3].

Owing to the correlation of slope and intercept, the individual confidence intervals for slope and intercept do not overlap completely. It is therefore recommended to evaluate the deviations from the expected slope 1 and intercept 0 not only by individual tests, but also by the joint test proposed by Mandel and Linnig [48]. With this joint test a more reliable conclusion (= smaller β-error) is possible when a specified bias must be detected with a given total number of measurements or alternatively less experiments need to be performed to reach a given β-error.

6.4.2. Standard addition

For some compounds (e.g. for certain endogenous compounds) no blank matrix is available. The bias can then be evaluated by the method of standard additions. A sample pool is divided in sub-samples of equal volumes, which are spiked by adding a constant, small volume with an increasing amount of analyte. A necessary condition is that homogeneous mixtures can be prepared. The standard addition line then is the line calculated between the measurement responses and the known added concentrations. When the calibration standards are not matrix-matched it is then possible to evaluate the bias by comparing the slopes of the standard addition line and the calibration line.

It is suggested to consider three concentration levels for both lines. The number of measurement points required to detect a given true bias by comparing the slopes of both lines can be calculated for given α- and β-errors as proposed by Penninckx et al. [49].

Table 5
Sample size required for the bias detection by ordinary least squares regression analysis[a]

Concentration Range	Levels	$\beta \leq 5\%$	$\beta \leq 10\%$	$\beta \leq 20\%$	$\beta \leq 40\%$
(a) Total number of observations required for the detection of a proportional bias of 10%, $\alpha = 5\%$ and β as indicated					
1–10	3	30	24	18	12
	6	42	36	30	18
1–100	3	21	18	15	9
	6	30	24	18	12
1–1000	3	21	18	12	9
	6	30	24	18	12
100–1000	3	30	24	18	12
	6	42	36	30	18
100–10 000	3	21	18	15	9
	6	30	24	18	12
(b) Total number of observations required for the detection of a constant bias of 10% of the mean of the range, $\alpha = 5\%$ and β as indicated					
1–10	3	45	36	27	18
	6	60	48	36	24
1–100	3	36	30	21	15
	6	48	36	30	18
1–1000	3	33	27	21	15
	6	42	36	30	18
100–1000	3	45	36	27	18
	6	60	48	36	24
100–10 000	3	45	36	27	18
	6	60	48	36	24

[a] Known measurement standard deviation (σ) of 10% of the mean of the range and three or six levels, evenly distributed over the range. Equal number of replicates at each level. Calculations based on ref. [47].

For a second order model it is more difficult to compare two lines. As for the linearity evaluation, more concentration levels need then to be considered than for a straight line relationship. For the data evaluation it has been suggested to apply a linearization procedure so that eventually straight lines can be compared [49].

With the standard addition approach it is not possible to detect a constant bias. The latter is, however, not a limitation for certain drug interaction studies. When a constant bias has to be evaluated one can use reference materials with known concentrations. Certified reference samples are, however, not always available. The constant bias must then be evaluated by a method comparison study (see Section 6.4.3.).

6.4.3. Method comparison

Method comparison is the bias evaluation procedure of choice when it is not possible to spike blank matrix homogeneously or when no blank matrix is available. It is then the only way to determine a possible constant bias component. Other situations where a method comparison should be performed are the evaluation of the method performance after a method transfer between two laboratories or when one wants to introduce an alternative to a well-established method.

It is recommended to analyse several replicates at each of the above defined three concentration levels. In any case, one must consider samples that are representative for the whole measurement range of the future study. The data are evaluated by *t*-tests or by regression.

6.4.3.1. t-Tests. For the detection of a given true bias between two measurement means by independent *t*-tests, the number of replicates required can be estimated from the ISO graphs [46]. Assuming that test and reference method have the same precision, at least six replicates have to be

Table 6
Number of data points required for the detection of a proportional bias of 10% by orthogonal least squares analysis when the slope is tested individually[a]

Concentration Range	Levels	$\beta \leq 5\%$	$\beta \leq 10\%$	$\beta \leq 20\%$	$\beta \leq 40\%$
1–10	3	18	15	12	9
	6	24	24	18	12
1–100	3	12	12	9	6
	6	18	18	12	12
1–1000	3	12	9	9	6
	6	18	18	12	6
100–1000	3	18	15	12	9
	6	24	24	18	12
100–10 000	3	12	12	9	6
	6	18	18	12	12

[a] Known measurement standard deviation (σ) of 5% of the mean of the range for both methods. Equal number of replicates at each of three or six levels, evenly distributed over the range, $\alpha = 5\%$ and β as indicated. Calculations based on ref. [47].

analysed with both methods to detect a bias twice as large as the known measurement standard deviation by means of one independent t-test ($\alpha = 5\%$, $\beta = 10\%$). When more than two and certainly when more than four concentration levels are considered for the bias evaluation it is suggested to perform a regression analysis (see Section 6.4.3.2.).

When samples are analysed each by a test and a reference method one could perform a paired t-test to evaluate the bias. Such a paired design has the advantage that a large diversity of samples (e.g. due to inter-individual differences, different concentrations,...) can be evaluated in an economical way.

The assumption underlying the paired t-test is, that the differences between the data pairs follow a normal distribution. This means that there may only be a constant but not a proportional bias and that the data must be homoscedastic. The former is in general not known before performing the comparison and for bioanalytical methods the latter is often not fulfilled. In the relatively large ranges considered in bioanalysis, a violation of these assumptions increases the number of data pairs required to detect a given bias in comparison with a situation with fulfilled test assumptions [45]. Therefore it is suggested to evaluate such a paired design by regression analysis, when it is not known whether the assumptions of the paired t-test are, at least approximately, fulfilled.

6.4.3.2. Regression analysis. For method comparisons both regression variables are the results of measurements. Both are therefore subject to error. Since one of the key assumptions of the ordinary applied Least Squares (LS) regression calculations is that the x-variable is error-free, LS should not be applied and the orthogonal least squares regression line must be calculated instead [50]. Care must be taken to consider sample sizes that make it probable to detect a relevant bias (i.e., that the β-error is low). Formulae have been proposed to estimate the sample sizes required to detect a specified proportional or constant bias with given α- and β-errors by testing slope and intercept individually [47]. Some examples for sample sizes required are shown in Table 6. The data evaluation of the orthogonal regression analysis is analogous to the one described above for LS. The data should first be evaluated visually [51] and the regression coefficients should not only be evaluated individually, but the joint test on slope and intercept should be performed to increase the probability to detect a real bias [47].

6.5. Advantage of interval hypothesis testing

It should be reminded that the recommendations made to limit the β-error to a certain level are not required when the bias is evaluated against specified acceptance limits by interval hypothesis testing. With the latter the safeguard against too large β-errors is already built in, because the tests are performed at a fixed β-error. This explains also why the interval hypothesis approach is considered in the SFSTP draft for bioanalytical methods.

When acceptance limits are not reached, the bioanalyst must decide if the requirements were not met due to too large a bias or due to too small the sample size and, therefore, a too low precision of the bias estimate.

7. Specificity

7.1. State of the art

In bioanalysis often many substances (endogenous substances, metabolites, degradation products, co-administered drugs, etc.) can potentially interfere in the determination of the analyte of interest. The extent of the specificity experiments is mainly determined by the application of the method, but also by the instrumental technique used.

Several validation documents (e.g. [1,8]) only require six different sources of blank matrices to be analysed. One must demonstrate that there is no interference in the chromatographic region of the analyte and, if used, of the internal standard. Sometimes, however, a certain percentage of contaminated samples or a certain level of interference expressed as percentage of the expected measurement response, e.g. of the expected maximum concentration $C_{\max}$, is considered acceptable [5].

7.2. Critique on the current practice

Suppose that it is accepted that during a proposed study of 20 samples up to 10% of the samples may be contaminated. To reach a 94% probability of detecting in at least two samples (i.e. in 10%) an interfering substance during the validation, a true contamination incidence of 21% is required [5]. With a lower contamination incidence, that, however, still is larger than the allowed 10% there is a large risk that less than two samples will be detected to be contaminated. Owing to these large sample sizes required to reach a high probability of detecting an interferent, it can be deduced that the probability of detecting an interference is rather low when only six blanks are analysed. Accepting an interference level in blank samples which is related to the results of future analyses gives problems, too, since these values (e.g. the $C_{\max}$) are often not yet known during the validation phase.

7.3. Recommended approach

For the evaluation of the specificity against known possible interferents, the following approach is recommended. A method that lacks specificity results in a systematic error, a bias. Recalling that usually a certain (limited) bias can still lead to acceptance of the method, the specificity of the method is sufficient as long as the bias is not affected to a relevant extent (see also reference [52]). It seems therefore more sensible to judge the specificity of a method on the resulting bias and to include the experiments to document the method specificity in the bias evaluation (Table 3).

In general it can be expected that the extent of interference increases with the concentration of the interferent, although there are exceptions [53]. Therefore, it is assumed that the largest influence of an interferent will be noticed at the lowest analyte concentration. Consequently, to evaluate the specificity against substances which are expected to be present and could possibly interfere, it is suggested to spike the lowest concentration level of the samples for the bias evaluation with the highest expected concentration of the possibly interfering substance (e.g. co-administered drugs) and to document whether the bias still reaches the acceptance limits.

A difficult situation occurs when one needs to investigate the specificity against substances which are usually present in the sample, so that no matrix free from interferent can be used. The results obtained with a matrix pool spiked with a high concentration of the suspected interferent then are compared with the ones of a pool which contains an average level of the suspected interferent, i.e. some kind of method comparison study needs to be performed [53].

However, one does not only need to investigate known likely interferents, one should also exclude that any other substance would interfere. Documenting the specificity against substances of indeterminate origin, which are always present or present only in a limited number of samples (e.g. due to circadian rhythms), is more complex. The latter is additionally complicated, when the calibration standards are prepared in the matrix, as it is usually done in bioanalysis [23]. Since it is impossible to test against all possible interferences, a compromise needs to be found between the workload and the attempt to unequivocally establish the specificity of the analytical procedure. It is suggested to spike different sources of blank matrices (say about 20) with the lowest concentration of the analyte that needs to be quantified. Each of these samples should then be analysed by the method under consideration and additionally also with an already validated method or at least with a different analytical procedure. Based on the comparison of the two series of results the specificity (the bias) of the newly developed method can then be evaluated (see also Section 6.4.3.).

8. Quantitation limit (LOQ)

The information obtained during the evaluation of the bias and the precision (see Section 6.) gives the basis to specify, the (lower) limit of quantitation ((L)LOQ). Some organisations, e.g. IUPAC [54], require only an evaluation of the precision but not of the bias at the LOQ. Since at all other levels where quantification is intended one must document precision and bias, it seems more sensible that at the LOQ also both parameters are evaluated. The level of quality which is needed at the LOQ depends on the goal of the analysis. If the requirements are not fulfilled at the tested concentration one can evaluate a higher concentration level which still complies with the concentration range requirements for the method. Otherwise one must go back to the method development. To avoid this latter inconvenience, in the draft on bioanalytical method validation of SFSTP, it is recommended to evaluate the LOQ during the exploratory validation phase (even when at this time usually no bias experiments are performed). In the SFSTP approach, the LOQ then is, together with the calibration model, the main point investigated.

9. (Assay) Range

Since any extrapolation should be avoided, not only the lowest but also the highest concentration in a sample that can be quantified with suitable precision and bias (sometimes called the upper limit of quantitation, ULOQ) needs to be indicated. The interval between these two extreme concentrations defines the (assay) range. It should be stressed that the assay range must not coincide with the concentration range of the calibration standards. In bioanalysis, it is quite usual to dilute or up-concentrate, since the concentration differences in the measurement samples can be extremely large. Such a dilution (or up-concentration) must of course be carefully validated as well. To clarify, the assay range then corresponds to the interval between the lowest sample concentration (before up-concentration) and the upper limit (before dilution) for which it has been documented that the quality of the method precision and bias is acceptable (Table 3).

10. Detection limit (LOD)

If needed, one can investigate at which concentration the analytical procedure allows to detect (but not necessarily to quantify) the analyte, i.e. the limit of detection (LOD) can be evaluated. According to the ICH [6] it is in general not necessary to evaluate the detection limit for quantitative methods, because most interest is given to determine the lowest level at which quantification is possible (LOQ, see Section 8.). Since it can also become useful to know the LOD for quantitative methods (e.g. for a pharmacokinetic study) it will be discussed briefly.

10.1. Theory

The detection limit expressed in response units, L_D, is defined as [31]:

$$L_D = \mu_{bl} + k\sigma_{bl} \quad (2)$$

where μ_{bl} represents the response of the blank which is estimated as $\bar{y}_{bl}$, the mean blank signal ($\bar{y}_{bl} = 0$ for blank-corrected signals) and σ_{bl} the true standard deviation of the blank measurements estimated by s_{bl}.

The corresponding detection limit expressed in concentration units, x_D, then is obtained for blank corrected signals as:

$$x_D = ks_{bl}/b_1 \quad (3)$$

with b_1, the slope of the calibration line.

A multiplication factor $k = 3$ is generally considered minimal. This implies for a normal distribution and a known standard deviation of the blank (σ_{bl}) a risk α of 7% to conclude that the analyte is present when it is absent (false positive decision) and a risk β of 7% to conclude that the analyte is absent when it is present (false negative decision). This can be understood from the fact that with $\alpha = 0.07$ and $\beta = 0.07$ (both one-sided):

$$k = z_\alpha + z_\beta = 1.5 + 1.5 = 3 \quad (4)$$

10.2. Limitations

10.2.1. Representative samples

It should be stressed that it is important that the LOD is evaluated with representative blank samples. When possible the precision of the blank measurements should be determined from a blank matrix. When no analyte-free matrix is available and when it cannot be prepared by chemical degradation or enzymatic conversion [16], the LOD is sometimes estimated from reagent blanks. It is, however, likely that the LOD then is underestimated.

10.2.2. Number of samples

Generally, the standard deviation of the blank measurements, σ_{bl}, is not known and only an estimate, s_{bl}, is available. The uncertainty in the estimation of the standard deviation can be taken into account using t-values instead of the z-values of the standardised normal distributions. This practice, which is recommended in a recent document by IUPAC [54], is, however, usually not applied. Moreover there is no consensus on the number of replicate measurements to include in the estimation of the variability of the blank, s_{bl}. Sometimes the standard deviation of the blank is estimated from the residual standard deviation determined for the calibration line or from the standard deviation of the y-intercept [6].

An additional complication which is almost always neglected is, that for the transformation of L_D into x_D by Eq. (3), the random error in the slope b_1 should be taken into account [54]. This means that the LOD estimate cannot be considered as a constant for a given method, but that the LOD will differ from day to day within the instrument and that it will differ as well from instrument to instrument.

10.2.3. Chromatographic methods

For chromatographic methods, there are additional problems to estimate the LOD. Currently the LOD is estimated from a certain ratio (in general three) of the analyte response, measured in peak height, to the maximum fluctuation of the background noise measured over a certain distance (20 times the peak width at half height) [13,55,56]. However, for bioanalytical methods often many peaks occur even in blank matrices. These peaks can interfere with the background noise and lead to an overestimation of the noise amplitude and consequently also of the LOD. This approach is therefore not appropriate for bioanalytical methods. Moreover, it cannot be taken for granted either that the peak height is the optimal decision basis. The determination of the sample concentration is frequently based on the response measured as peak area. Since blank samples do usually not give measurable peak areas, one cannot apply the procedure recommended by IUPAC [54]. For a CE (capillary electrophoresis) method, used for the determination of mineral elements in food and botanical materials, it has therefore been suggested to spike reagent blanks at (or to consider real samples of) a low concentration level, e.g. the LOD as estimated from the approach based on the peak height. Replicate analyses are performed and the concentration corresponding to three times the standard deviation of the peak areas is then considered as the LOD [43].

All these limitations and problems indicate that there is a need for a consensus on how to determine the LOD for bioanalytical methods to make possible a comparison of the LOD results of different laboratories.

11. Analyte stability in the matrix

When the above described experiments fulfil the acceptance requirements and when it is known what is actually feasible with a given method, the analyte stability in the matrix can be evaluated. Of course, one must avoid that the above described experiments are affected by not stable reagent solutions and/or a possible instability of the solutions ready to be injected (case (ii), see below). The latter stability experiments should therefore be performed as early as possible during a validation or appropriate precautions must be taken into account to avoid that the validation experiments are affected.

The stability of the analyte is often critical in biological matrices even during relative short time periods. Degradation is not unusual even when all precautions are taken to avoid specifically known stability problems of the analyte (e.g. light protection). It is, therefore, important to verify that there is no relevant degradation between the time of the collection of the samples and their analysis that would compromise the results of the study. It should also be stressed that important differences of the stability of a substance can exist in corresponding matrices of even very close species. Unfortunately no guidance is provided how the stability should be investigated.

11.1. Test conditions

In the Washington conference report, it is recommended to establish the stability of the analyte in the biological matrix at the intended storage temperature and to study the influence of thaw-and-freeze cycles [1]. A possible degradation of the analyte should, however, also be controlled in other situations. It seems advisable to investigate the following stability conditions (Table 3):

(i) The benchtop stability of the analyte in the matrix at ambient temperature (i.e. the stability of the analyte before processing a sample) to decide if a preservative regime has to be included in the sampling procedure,

(ii) The stability of the analyte in the final extract during the anticipated maximum analysis time and the time needed to repeat the analysis (if enough volume is available), which, for automatic injections, can often be up to 24–48 h,

(iii) The stability of the analyte during three (or more) thaw-and-freeze cycles (a cycle consisting, for example, of thawing the sample, letting it stand during 1 h at ambient temperature and freezing it for at least 24 h), and

(iv) The (long-term) stability of frozen samples prior to the analysis.

11.2. Experimental design

Timm et al. have proposed to compare the results of the samples that were processed through the stability conditions with reference samples [57] and to evaluate the data by interval hypothesis testing. The practicability of this approach has been confirmed, e.g. by Hooper [5]. This means that the stability study is the only part of method validation where the need for interval hypothesis testing has been recognised and where it is actually applied in practice in a few instances.

The number of levels to be evaluated depends on the concentration range of the method. For all cases mentioned above, the stability should at least be documented at the extremes of this range (Table 3). To allow a reliable conclusion with a still feasible workload, it is recommended to evaluate the stability only for the largest time the samples would possibly be stored. This means also that one should only evaluate the stability after the last thaw-and-freeze cycle and that there is no need to evaluate the stability after each cycle. Only for a rather long storage period prior to the analysis, e.g. a long-term storage of 6 months, it may be preferable to include an evaluation at an intermediate time point in order to assure that the stability can be guaranteed at least during a certain period. Another possibility is to start additionally a short stability study of, e.g. 2 months at the same time as the full long-term study.

The number of replicates required must be determined according to the imposed requirements, which should depend on the goal of the study. A minimum of six replicates seems, however, in any case sensible (Table 3). Recently it has, for example, been proposed that the concentration ratio for stored and reference samples should be between 90 and 110%, whereas the 90%-confidence interval of this concentration ratio should lie within the acceptance interval 80–120% [5]. Timm et al. [57] suggested to test a possible degradation only against a lower acceptance limit situated at − 10% from the mean concentration calculated for the reference samples.

In principle two sample preparation approaches are possible. Either 'fresh' reference samples are prepared the day of the analysis, or all samples are prepared the same day, but the reference samples are stored in liquid nitrogen or in a freezer at temperatures < -130°C. It has been documented by Dadgar et al. [58] that any degradation is (almost) stopped under these conditions. Despite the high costs, the latter approach is therefore always recommended when real samples are analysed with a critical stability. Care should also be taken to store the stability samples in the same type of container as intended for the routine application to take into account, e.g. possible interaction effects with the container walls. Furthermore, reference and 'stored' samples should be analysed in a random order and in the shortest time possible.

The data are again visually evaluated, at least by a scatter (or a box) plot and, if indicated, outlier tests are performed. After comparing the precision of the 'stored' and the reference samples by an F-test, one statistically evaluates the stability by interval hypothesis testing [14,57].

12. Robustness

The robustness evaluation is not mentioned in the Conference report [1] and it is not considered in most validation guidelines. However, considering the amount of time wasted for problem-solving during a routine application, robustness testing has certainly an impact for bioanalytical methods that are used over longer periods of time and/or in different laboratories (Table 3). In contrast to the method development where one looks for the optimal method conditions and therefore needs to screen a larger experimental region, during the validation phase only the effects of small changes in the experimental conditions need to be studied by robustness tests. The latter experiments are aimed at defining the steps of the established and validated analytical procedure that need to be controlled carefully during the routine application in order to generate good quality data. For extraction steps one should, e.g. evaluate the influence of slight changes in the pH, in the ionic strength or in the volume mixture of the organic and the aqueous phase, an important factor for evaporation steps is the temperature whereas for chromatographic methods the effect of slight changes in the mobile phase composition, in the pH of the buffer, in the ambient temperature and in the detection wavelength could be investigated. In the field of pharmaceutical applications, robustness studies have, e.g. been described by Vander Heyden et al. [59,60].

13. Linking validation and routine application

It should be reminded that the effort of a method validation is undertaken to guarantee during the routine analysis a quality of the measurement data as needed for the given application. For the moment, however, a validation is too frequently considered as an additional burden imposed by regulatory agencies and far too rarely one tries to include the knowledge gained during a validation in (the quality control of) the routine application. Only one point shall be mentioned here. Considering the same concentrations for the quality control samples as studied during the validation experiments for precision and bias, facilitates the on-going assessment of precision and bias as the method is applied. It helps, for example, to specify acceptance limits for a run that really reflect the quality of the method used. The estimates of the bias and the precision measures can, for example, be used to determine the initial limits of the control charts, such as already mentioned by Lang and Bolton [30], so that a reliable control is possible from the first run of the routine analysis. At this moment the time spent on defining appropriate requirements with respect to the specification limits is re-paid as well, since only then it is possible to routinely reach a required quality level with an acceptable proportion of falsely accepted and falsely rejected runs (see β- and α-error in Section 3. and the proposal by Selinger [38]).

14. Conclusion

It is necessary to plan the validation study and its evaluation in detail before starting the experiments. The goals of the bioanalytical method and minimum requirements imposed by official organisations must be taken into account and balanced with the statistical needs on experimental designs that allow a reliable conclusion about the quality of the data produced. Despite the progress made during the last years, the knowledge of the best practices possible is still missing for several steps of the validation. In many areas, it is necessary to decide on a compromise between the quality of the statistical decisions (particularly the β-error to be expected) and the workload required. It would be useful to achieve a consensus on such compromises.

The terminology should be harmonised and a consensus is also needed on how to take into account minimum acceptance requirements, such as, for example, those for bias and precision specified during the Washington conference in 1990 [1]. The use of interval hypothesis tests seems to be recommended in this context.

It is realised that proper method validation requires a lot of work. However, this effort is re-paid by the time saved when running the method routinely.

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, S. Spector, Pharm. Res. 9 (1992) 588–592.

[2] C. Hartmann, D.L. Massart, R.D. McDowall, J. Pharm. Biomed. Anal. 12 (1994) 1337–1343.

[3] C. Hartmann, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen, D.L. Massart, R.D. McDowall, in: H.H. Blume, K.K. Midha (Eds.), Bio-International 2, Medpharm Scientific Publishers, Stuttgart, 1995, pp. 331–346.

[4] R.O. Kringle, Pharm. Res. 11 (1994) 556–560.

[5] J.W. Hooper, in: H.H. Blume, K.K. Midha (Eds.), Bio-International 2, Medpharm Scientific Publishers, Stuttgart, 1995, pp. 347–355.

[6] Commission of the European Communities, Committee for Proprietary Medicinal Products, Note for Guidance, Validation of analytical procedures, III/5626/94 Final, 1994.

[7] Association of Official Analytical Chemists, Referee 6 (1995).

[8] International Organization for Standardization, Accuracy (Trueness and Precision) of Measurement Methods and Results, ISO/DIS 5725-1 to 5725-3, Geneva, 1994.

[9] F.J. van de Vaart, Pharmaceutisch Weekblad 127 (1992) 1229–1235.

[10] M.A. Brooks, R.E. Weinfeld, Drug Dev. Ind. Pharm. 11 (1985) 1703–1728.

[11] A. Artiges, Pharmaeuropa 5 (1993) 343–348.

[12] J. Vessman, J. Pharm. Biomed. Anal. 14 (1996) 867–869.

[13] J. Caporal-Gautier, J.M. Nivet, P. Algranti, M. Guilloteau, M. Histe, L. Lallier, J.J. N'Guyen-Huu, R. Russotto, STP Pharma Pratiques 2 (1992) 205–240.

[14] C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen, D.L. Massart, Anal. Chem. 67 (1995) 4491–4499.

[15] S. Braggio, R.J. Barnaby, P. Grossi, M. Cugola, J. Pharm. Biomed. Anal. 14 (1996) 375–388.

[16] A.R. Buick, M.V. Doig, S.C. Jeal, G.S. Land, R.D. McDowall, J. Pharm. Biomed. Anal. 8 (1990) 629–637.

[17] J.R. Lang, S. Bolton, J. Pharm. Biomed. Anal. 9 (1991) 357–361.

[18] International Conference on Harmonisation, Final Draft Guideline for Validation of Analytical Procedures: Methodology, 1996.

[19] J.L. Burrows, J. Pharm. Biomed. Anal. 11 (1993) 523–531.

[20] H.T. Karnes, C. March, J. Pharm. Biomed. Anal. 9 (1991) 911–918.

[21] L. Aarons, Analyst 106 (1981) 1249–1254.

[22] H.M. Hill, A.G. Causey, D. Lessard, K. Selinger, J. Herman, in: E. Reid, I.D. Wilson (Eds.), Methodological Surveys in Biochemistry and Analysis, vol. 22, Royal Society of Chemistry, 1992, pp. 111–118.

[23] P. Arnoux, R. Morrison, Xenobiotica 22 (1992) 757–764.

[24] C. Hartmann, Analusis 22 (1994) M19–M21.

[25] International Organization for Standardization, Linear Calibration using Reference Material, ISO 11095, Geneva, 1996.

[26] W. Penninckx, C. Hartmann, D.L. Massart, J. Smeyers-Verbeke, J. Anal. At. Spectrom. 11 (1996) 237–246.

[27] International Organization for Standardization, Water Quality—Calibration and Evaluation of Analytical Methods and Estimation of Performance Characteristics, ISO 8466-1 and 8466-2 Geneva, 1990.

[28] G.S. Land, W.J. Leavens, B.C. Weatherley, in: E. Reid, I.D. Wilson (Eds.), Methodological Surveys in Biochemistry and Analysis, vol. 22, Royal Society of Chemistry, 1992, pp. 103–110.

[29] M. Davidian, P.D. Haaland, J. Chem. Intell. Lab. Syst. 9 (1990) 231–248.

[30] J.R. Lang, S. Bolton, J. Pharm. Biomed. Anal. 9 (1991) 435–442.

[31] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.

[32] Analytical Methods Committee, Analyst 113 (1988) 1469–1471.

[33] International Organization for Standardization, Accuracy (Trueness and Precision) of Measurement Methods and Results, ISO/DIS 5725-6, Geneva, 1994.

[34] P. Hubert, P. Chiap, J. Crommen, Stage de Validation, Université de Liège, Liège, 28 June–2 July 1993.

[35] P. J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.

[36] P. Vankeerberghen, J. Smeyers-Verbeke, D.L. Massart, J. Anal. At. Spectrom. 11 (1996) 149–158.

[37] H.T. Karnes, C. March, Pharm. Res. 10 (1993) 1420–1425.

[38] K.A. Selinger, J. Pharm. Biomed. Anal. 13 (1995) 1427–1436.

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

[40] S. Wallenstein, C.L. Zucker, J.L. Fleiss, Circ. Res. 47 (1980) 1–9.

[41] S.J. Haswell (Ed.), Practical Guide to Chemometrics, Marcel Dekker, New York, 1992.

[42] J.M. Bland, D.G. Altman, Lancet 8 (1986) 307–310.

[43] Q. Yang, C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, J. Chromatogr. A 717 (1995) 415–425.

[44] G.W. Snedecor, W.G. Cochran, Statistical Methods, 7th ed., Iowa State University Press, Ames, 1980.

[45] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, Comparison of the results of two analytical procedures, AOAC Lage Landen Symp., poster presentation, Maastricht, The Netherlands, 14 November 1995.

[46] International Organization for Standardization, Statistical Methods, ISO Standards Handbook 3, 3rd ed., Geneva, 1989.

[47] C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, D.L. Massart, Anal. Chim. Acta 338 (1997) 19–40.

[48] J. Mandel, F.J. Linnig, Anal. Chem. 29 (1957) 743–749.

[49] W. Penninckx, P. Vankeerberghen, D.L. Massart, J. Smeyers-Verbeke, J. Anal. At. Spectrom. 10 (1995) 207–214.

[50] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, Analusis 21 (1993) 125–132.

[51] C. Hartmann, D.L. Massart, J. Assoc. Off. Anal. Chem. 77 (1994) 1318–1325.

[52] D. Dadgar, P.E. Burnett, J. Pharm. Biomed. Anal. 14 (1995) 23–31.

[53] National Committee for Clinical Laboratory Standards, Interference Testing in Clinical Chemistry, Proposed Guideline, NCCLS publication EP7-P, Villanova, PA, 1986.

[54] International Union of Pure and Applied Chemistry, Pure Appl. Chem. 67 (1995) 1699–1723.

[55] J.E. Knoll, J. Chromatogr. Sci. 23 (1985) 422–425.

[56] G.P. Carr, J.C. Wahlich, J. Pharm. Biomed. Anal. 8 (1990) 613–618.

[57] U. Timm, M. Wall, D. Dell, J. Pharm. Sci. 74 (1985) 972–977.

[58] D. Dadgar, P.E. Burnett, M.G. Choc, K. Gallicano, J.W. Hooper, J. Pharm. Biomed. Anal. 13 (1995) 89–97.

[59] Y. Vander Heyden, K. Luypaert, C. Hartmann, D.L. Massart, J. Hoogmartens, J. De Beer, Anal. Chim. Acta 312 (1995) 245–262.

[60] Y. Vander Heyden, C. Hartmann, D.L. Massart, L. Michel, P. Kiechle, F. Erni, Anal. Chim. Acta 316 (1995) 15–26.