Review: Bioanalytical method validation – How, how much and why ?

Frank T. Peters and Hans H. Maurer

Department of Experimental and Clinical Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, University of Saarland, D-66421 Homburg (Saar)
E-mail: frank.peters@uniklinik-saarland.de

1. Introduction

The reliability of analytical findings is a matter of great importance in forensic and clinical toxicology, as it is of course a prerequisite for correct interpretation of toxicological findings. Unreliable results might not only be contested in court, but could also lead to unjustified legal consequences for the defendant or to wrong treatment of the patient. The importance of validation, at least of routine analytical methods, can therefore hardly be overestimated. This is especially true in the context of quality management and accreditation, which have become matters of increasing importance in analytical toxicology in recent years. This is also reflected in the increasing requirements of peer reviewed scientific journals concerning method validation. Therefore, this topic should extensively be discussed on an international level to reach a consensus on the extent of validation experiments and on acceptance criteria for validation parameters of bioanalytical methods in forensic (and clinical) toxicology.

In the last decade, similar discussions have been going on in the closely related field of pharmacokinetic studies for registration of pharmaceuticals. This is reflected by a number of publications on this topic in the last decade, of which the most important are discussed here.

2. Important publications on validation (1991 to present)

A review on validation of bioanalytical methods was published by Karnes et al. in 1991 which was intended to provide guidance for bioanalytical chemists [10]. One year later, Shah et al. published their report on the conference on "Analytical Methods Validation: Bioavailability, Bioequivalence and Pharmacokinetic Studies" held in Washington in 1990 (Conference Report) [13]. During this conference, consensus was reached on which parameters of bioanalytical methods should be evaluated, and some acceptance criteria were established. In the following years, this report was actually used as guidance by bioanalysts. Despite the fact, however, that some principle questions had been answered during this conference, no specific recommendations on practical issues like experimental designs or statistical evaluation had been made. In 1994, Hartmann et al. analyzed the Conference Report performing statistical experiments on the established acceptance criteria for accuracy and precision [5]. Based on their results they questioned the suitability of these criteria for practical application. From 1995 to 1997, application issues like experimental designs and statistical methods for bioanalytical method validation were discussed in a number of publications of Dadgar et al. [3, 4], Wieling et al. [17], Bressolle et al. [1] and Causon [2]. An excellent review on validation of bioanalytical chromatographic methods has been published by Hartmann et al. in 1998, in which theoretical and practical issues were discussed in detail [6]. Finally, in an update conference of the Washington conference, experiences and progress since the first conference have been discussed. The results were again published by Shah et al. in a report (Conference Report II) [14], which has also been used as a template for their own guidelines by the U.S. Food and Drug Administration (FDA) [15]. Besides, it should be mentioned that some journals like Journal of Chromatography B [11] or Clinical Chemistry have established their own criteria for validation. Two other documents that seem to be important in this context have been developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) and approved by the regulatory agencies of the European Union, the United States of America and Japan. Despite the fact, that these were focussed on analytical methods for pharmaceutical products rather than bioanalysis, they still contain helpful guidance on some principal questions and definitions in the field of analytical method validation. The first document, approved in 1994, concentrated on the theoretical background and definitions [7], the second, approved in 1996, on methodology and practical issues [8]. Both can be downloaded from the ICH homepage free of charge (www.ich.org).

The aim of our review is to present and compare the contents of the above mentioned publications on (bio)analytical method validation, and to discuss possible implications for forensic and clinical toxicology.

3. Terminology

The first problem encountered when studying literature on method validation is the different sets of terminology employed by different authors. A detailed discussion of this problem can be found in the two papers of Hartmann et al. [5, 6]. In their review [6], it was proposed to adhere in principle to the terminology established by the ICH [7], except for accuracy, for which the use of a more differentiated definition was recommended (cf. 4.3.). However, the ICH terminology lacked a definition for stability, which is an important parameter in bioanalytical method validation. Furthermore, the ICH definition of selectivity did not take into account interferences, that might occur in bioanalysis (e.g. from metabolites). For both parameters, however, reasonable definitions were provided by Conference Report II [14].

4. Validation parameters

There is a general agreement, that at least the following validation parameters should be evaluated for quantitative procedures: selectivity, calibration model, stability, accuracy (bias, precision) and limit of quantification. Additional parameters which might have to be evaluated include limit of detection, recovery, reproducibility and ruggedness (robustness) [1-4, 6, 11, 13, 14, 17].

4.1 Selectivity (Specificity)

In the Conference Report II, selectivity was defined as follows: Selectivity is the ability of the bioanalytical method to measure unequivocally and to differentiate the analyte(s) in the presence of components, which may be expected to be present. Typically, these might include metabolites, impurities, degradants, matrix components, etc. [14]. This definition is very similar to the one established by the ICH [7], but takes into account the possible presence of metabolites, and is therefore more applicable for bioanalytical methods.

There are two points of view on when a method should be regarded to be selective. One way to establish method selectivity is to prove the lack of response in blank matrix [1-4, 6, 8, 10, 11, 13, 14, 17]. The requirement established by the Conference Report [13] to analyze at least six different sources of blank matrix has become state of the art. However, this approach has been subject to criticism in the review of Hartmann et al., who stated from statistical considerations, that relatively rare interferences will remain undetected with a rather high probability [6]. For the same reason, Dadgar et al. proposed to evaluate at least 10-20 sources of blank samples [3]. However, in the Conference Report II [14], even analysis of only one source of blank matrix was deemed acceptable, if hyphenated mass spectrometric methods are used for detection.

The second approach is based on the assumption that small interferences can be accepted as long as precision and bias remain within certain acceptance limits. This approach was preferred by Dadgar et al. and Hartmann et al. [3, 6]. Both authors proposed analysis of up to 20 blank samples spiked with analyte at the lower limit of quantification (LLOQ) and, if possible, with interferents at their highest likely concentrations. In this approach, the method can be considered sufficiently selective if precision and accuracy data for these LLOQ samples are acceptable. For a detailed account of experimental designs and statistical methods to establish selectivity see ref. [3].

Whereas the selectivity experiments for the first approach can be performed during a prevalidation phase (no need for quantification), those for the second approach are usually performed together with the precision and accuracy experiments during the main validation phase.

At this point it must be mentioned, that the term specificity is used interchangeably with selectivity, although in a strict sense specificity refers to methods, which produce a response for a single analyte, whereas selectivity refers to methods that produce responses for a number of chemical entities, which may or may not be distinguished [10]. Selective multianalyte methods (e.g. for different drugs of abuse in blood) should of course be able to differentiate all interesting analytes from each other and from the matrix.

4.2 Calibration model

The choice of an appropriate calibration model is necessary for reliable quantification. Therefore, the relationship between the concentration of analyte in the sample and the corresponding detector response must be investigated. This can be done by analyzing spiked calibration samples and plotting the resulting responses versus the corresponding concentrations. The resulting standard curves can then be further evaluated by graphical or mathematical methods, the latter also allowing statistical evaluation of the response functions.

Whereas there is general agreement that calibration samples should be prepared in blank matrix and that their concentrations must cover the whole calibration range, recommendations on how many concentration levels should be studied with how many replicates per concentration level differ significantly [1, 2, 4, 6, 11, 14, 17]. In the Conference Report II, "a sufficient number of standards to define adequately the relationship between concentration and response" was demanded. Furthermore, it was stated that at least five to eight concentration levels should be studied for linear and maybe more for non-linear relationships [14]. However, no information was given on how many replicates should be analyzed at each level. The guidelines established by the ICH and those of the Journal of Chromatography B also required at least five concentration levels, but again no specific requirements for the number of replicates at each level were given [8, 11]. Causon recommended six replicates at each of six concentration levels, whereas Wieling et al. used eight concentration levels in triplicate [2, 17]. Based on studies by Penninckx et al. [16], Hartmann et al. proposed in their review to rather use fewer concentration levels with a greater number of replicates (e.g. four evenly spread levels with nine replicates) [6]. This approach not only allows the reliable detection of outliers, but also a better evaluation of the behaviour of variance across the calibration range. The latter is important for choosing the right statistical model for the evaluation of the calibration curve. The often used ordinary least squares model for linear regression is only applicable for homoscedastic data sets (constant variance over the whole range), whereas in case of heteroscedasticity (significant difference between variances at lowest and highest concentration levels) the data should mathematically be transformed or a weighted least squares model should be applied [1, 2, 6, 14, 17]. Usually, linear models are preferable, but, if necessary, the use of non-linear models is not only acceptable but even recommended. However, more concentration levels are needed for the evaluation of non-linear models than for linear models [6, 13, 14].

After outliers have been purged from the data and a model has been evaluated visually and/or by e.g. residual plots, the model fit should also be tested by appropriate statistical methods [6, 8, 13, 14, 17]. The fit of unweighted regression models (homoscedastic data) can be tested by the ANOVA lack-of-fit test [6, 17]. A detailed discussion of alternative statistical tests for both unweighted and weighted calibration models can be found in ref. [16]. The widespread practice to evaluate a calibration model via its coefficients of correlation or determination is not acceptable from a statistical point of view [6].

However, one important point should be kept in mind when statistically testing the model fit: The higher the precision of a method, the higher the probability to detect a statistically significant deviation from the assumed calibration model [6, 10, 17]. Therefore, the relevance of the deviation from the assumed model must also be taken into account. If the accuracy data (bias and precision) are within the required acceptance limits and an alternative calibration model is not applicable, slight deviations from the assumed model may be neglected [6, 17].
Once a calibration model has been established, the calibration curves for other validation experiments (precision, bias, stability etc.) and for routine analysis can be prepared with fewer concentration levels and fewer or no replicates [6, 17].

4.3. Accuracy

The accuracy of a method is affected by systematic (bias) as well as random (precision) error components. [5, 6] This fact has been taken into account in the definition of accuracy as established by the International Organization for Standardization (ISO) [9]. However, it must be mentioned, that accuracy is often used to describe only the systematic error component, i.e. in the sense of bias [1, 2, 7, 10, 11, 13, 14, 17]. In the following, the term accuracy will be used in the sense of bias, which will be indicated in brackets.

4.3.1. Bias

According to ISO, bias is the difference between the expectation of test results and an accepted reference value [9]. It may consist of more than one systematic error component. Bias can be measured as a percent deviation from the accepted reference value. The term trueness expresses the deviation of the mean value of a large series of measurements from the accepted reference value. It can be expressed in terms of bias.
Due to the high workload of analyzing such large series, trueness is usually not determined during method validation, but rather from the results of a great number of quality control samples (QC samples) during routine application.

4.3.2. Precision

According to ICH, precision is the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogenous sample under the prescribed conditions and may be considered at three levels: repeatability, intermediate precision and reproducibility [7]. Precision is usually measured in terms of imprecision expressed as an absolute or relative standard deviation and does not relate to reference values.

4.3.2.1. Repeatability

Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability is also termed intra-assay precision [7]. Repeatability is sometimes also termed within-run or within-day precision.

4.3.2.2. Intermediate precision

Intermediate precision expresses within-laboratories variations: different days, different analysts, different equipment etc. [7]. The ISO definition used the term "M-factor different intermediate precision", where the M-factor expresses the number of factors (operator, equipment or time) that differ between successive determinations [9]. Intermediate precision is sometimes also called between-run, between-day or inter-assay precision.

4.3.2.3. Reproducibility

Reproducibility expresses the precision between laboratories (collaborative studies, usually applied to standardization of methodology) [7]. Reproducibility only has to be studied, if a method is supposed to be used in different laboratories.
Unfortunately, some authors also used the term reproducibility for within-laboratory studies at the level of intermediate precision [2, 11]. This should, however, be avoided in order to prevent confusion.

As already mentioned above precision and bias can be estimated from the analysis of QC samples under specified conditions. As both precision and bias can vary substantially over the calibration range, it is necessary to evaluate these parameters at least at three concentration levels (low, medium, high) [6, 8, 10, 13, 14]. In the Conference Report II, it was further defined that the low QC sample must be within three times LLOQ [14]. The Journal of Chromatography B requirement is to study precision and bias at two concentration levels (low and high), whereas in the experimental design proposed by Wieling et al. four concentration levels (LLOQ, low, medium, high) were studied [11, 17]. Causon also suggested to estimate precision at four concentration levels [2]. Several authors have specified acceptance limits for precision and/or accuracy (bias) [1, 2, 11, 13, 14]. The Conference Reports required precision to be within 15% relative standard deviation (RSD) except at the LLOQ where 20% RSD is accepted. Bias is required to be within ±15% of the accepted true value, except at the LLOQ where ±20% are accepted [13, 14]. These requirements have been subject to criticism in the analysis of the Conference Report by Hartmann et al. [5]. They concluded from statistical considerations that it is not realistic to apply the same acceptance criteria at different levels of precision (repeatability, reproducibility) as RSD under reproducibility conditions is usually considerably greater than under repeatability conditions. Furthermore, if precision and bias estimates are close to the acceptance limits, the probability to reject an actually acceptable method (β-error) is quite high. Causon proposed the same acceptance limits of 15% RSD for precision and ±15% for accuracy (bias) for all concentration levels [2].
The guidelines established by the Journal of Chromatography B required precision to be within 10% RSD for the high QC samples and within 20% RSD for the low QC sample. Acceptance criteria for accuracy (bias) were not specified there [11].

Again the proposals on how many replicates at each concentration levels should be analyzed vary considerably. The Conference Reports and Journal of Chromatography B guidelines required at least five replicates at each concentration level [11, 13, 14]. However, one would assume that these requirements apply to repeatability studies; at least no specific recommendations are given for studies of intermediate precision or reproducibility. Some more practical approaches to this problem have been described by Wieling et al. [17], Causon [2] and Hartmann et al. [6]. In their experimental design, Wieling et al. analyzed three replicates at each of four concentration levels on each of five days. Similar approaches were suggested by Causon (six replicates at each of four concentrations on each of four occasions) and Hartmann et al. (two replicates at each concentration level on each of eight days). All three used one-way ANOVA to estimate within-run precision (repeatability) and between-run precision (intermediate precision). In the design proposed by Hartmann et al. the degrees of freedom for both estimations are most balanced, namely eight for within-run precision and seven for between-run precision. In the information for authors of the Clinical Chemistry journal, an experimental design with two replicates per run, two runs per day over 20 days for each concentration level is recommended, which has been established by the NCCLS [12]. This not only allows estimation of within-run and between-run standard deviations, but also of within-day, between-day and total standard deviations, which are in fact all estimation of precision at different levels. However, it seems questionable if the additional information provided by this approach can justify the high workload and costs compared to the other experimental designs.

Daily variations of the calibration curve can influence bias estimation. Therefore, bias estimation should be based on data calculated from several calibration curves [6]. In the experimental design of Wieling et al., the results for QC samples were calculated via daily calibration curves. Therefore, the overall means from these results at the different concentration levels reliably reflect the average bias of the method at the corresponding concentration level. Alternatively, as described in the same paper, the bias can be estimated using confidence limits around the calculated mean values at each concentration [17]. If the calculated confidence interval includes the accepted true value, one can assume the method to be free of bias at a given level of statistical significance. Another way to test the significance of the calculated bias is to perform a t-test against the accepted true value.

However, even methods exhibiting a statistically significant bias can still be acceptable, if the calculated bias lies within previously established acceptance limits. Other methods for bias evaluation can be found in ref. [6].

4.4. Limits

4.4.1. Lower limit of quantification (LLOQ)

The LLOQ is the lowest amount of an analyte in a sample that can be quantitatively determined with suitable precision and accuracy (bias) [7, 14]. There are different approaches to the determination of LLOQ.

4.4.1.1. LLOQ based on precision and accuracy (bias) data [1, 2, 6-8, 13, 14]

This is probably the most practical approach and defines the LLOQ as the lowest concentration of a sample that can still be quantified with acceptable precision and accuracy (bias). In the Conference reports, the acceptance criteria for these two parameters at LLOQ are 20% RSD for precision and ±20% for bias. Only Causon suggested 15% RSD and ±15% respectively [2]. It should be pointed out, however, that these parameters must be determined using an LLOQ sample independent from the calibration curve. The advantage of this approach is the fact, that the estimation of LLOQ is based on the same quantification procedure used for real samples.

4.4.1.2. LLOQ based on signal to noise ratio (S/N) [8, 11]

This approach can only be applied if there is baseline noise, e.g. to chromatographic methods. Signal and noise can then be defined as the height of the analyte peak (signal) and the amplitude between the highest and lowest point of the baseline (noise) in a certain area around the analyte peak. For LLOQ, S/N is usually required to be equal to or greater than 10.

The estimation of baseline noise can be quite difficult for bioanalytical methods, if matrix peaks elute close to the analyte peak.

4.4.1.3. LLOQ based on standard deviation of the response from blank samples [8]

Another definition of LLOQ is the concentration that corresponds to a detector response that is k-times greater than the estimated standard deviation of blank samples (sbl). From the detector signal, the LLOQ can be calculated using the slope of the calibration curve (S) with following formula: LLOQ = k·sbl/S (for blank corrected signals).

This approach is only applicable for methods where sbl can be estimated from replicate analysis of blank samples. It is therefore not applicable for most quantitative chromatographic methods, as here the response is usually measured in terms of peak area units, which can of course not be measured in a blank sample analyzed with a selective method.

4.4.1.4. LLOQ based on a specific calibration curve in the range of LLOQ [8]

In this approach, a specific calibration curve is established from samples containing the analyte in the range of LLOQ. One must not use the calibration curve over the whole range of quantification for this determination. The standard deviation of the blank can then be estimated from the residual standard deviation of the regression line or the standard deviation of the y-intercept. The calculations of LLOQ are basically the same as described under 4.4.1.3. This approach is also applicable for chromatographic methods.

4.4.2. Upper limit of quantification (ULOQ)

The upper limit of quantification is the maximum analyte concentration of a sample, that can be quantified with acceptable precision and accuracy (bias). In general the ULOQ is identical with the concentration of the highest calibration standard [14].

4.4.3. Limit of detection (LOD)

Quantification below LLOQ is by definition not acceptable [4, 6-8, 13, 14]. Therefore, below this value a method can only produce semiquantitative or qualitative data. However, it can still be important to know the LOD of the method. According to ICH, it is the lowest concentration of analyte in a sample which can be detected but not necessarily quantified as an exact value. According to Conference Report II, it is the lowest concentration of an analyte in a sample, that the bioanalytical procedure can reliably differentiate from background noise [7, 14].

The approaches for estimation of the LOD are basically the same as those described for LLOQ under 4.4.1.2 - 4.4.1.4. However, for LOD a S/N or k-factor equal to or greater than three is usually chosen [6, 8, 10, 17]. If the calibration curve approach is used for determination of the LOD, only calibrators containing the analyte in the range of LOD must be used.

4.5. Stability

The definition according to Conference Report II was as follows: The chemical stability of an analyte in a given matrix under specific conditions for given time intervals [14]. Stability of the analyte during the whole analytical procedure is a prerequisite for reliable quantification. Therefore, full validation of a method must include stability experiments for the various stages of analysis including storage prior to analysis.

4.5.1. Long-term stability

The stability in the sample matrix should be established under storage conditions, i.e. in the same vessels, at the same temperature and over a period at least as long as the one expected for authentic samples [3, 4, 6, 10, 11, 13, 14].

4.5.2. Freeze/thaw stability

As samples are often frozen an thawed, e.g. for reanalyis, the stability of analyte during several freeze/thaw cycles should also be evaluated. The Conference reports require a minimum of three cycles at two concentrations in triplicate, which has also been accepted by other authors [3, 6, 13, 14, 17].

4.5.3. In-process stability

The stability of analyte under the conditions of sample preparation (e.g. ambient temperature over time needed for sample preparation) is evaluated here. There is general agreement, that this type of stability should be evaluated to find out, if preservative have to be added to prevent degradation of analyte during sample preparation [3, 6, 14].

4.5.4. Processed sample stability

Instability cannot only occur in the sample matrix, but also in prepared samples. It is therefore important to also test the stability of an analyte in the prepared samples under conditions of analysis (e.g. autosampler conditions for the expected maximum time of an analytical run). One should also test the stability in prepared samples under storage conditions, e.g. refrigerator, in case prepared samples have to be stored prior to analysis [3, 4, 6, 14, 17].

For more details on experimental design and statistical evaluation of stability experiments see references [3, 4, 6].

Stability can be tested by comparing the results of QC samples analyzed before (comparison samples) and after (stability samples) being exposed to the conditions for stability assessment. It has been recommended to perform stability experiments at least at two concentration levels (low and high) [3, 4, 6, 17]. For both, comparison and stability samples, analysis of at least six replicates was recommended [6]. Ratios between comparison samples and stability samples of 90-110% with 90% confidence intervals within 80-120% [6] or 85-115% [3] were regarded acceptable. Alternatively, the mean of the reference samples can be tested against a lower acceptance limit corresponding to 90% of the mean of the comparison samples [2, 6].

4.6. Recovery

As already mentioned above, recovery is not among the validation parameters regarded as essential by the Conference reports. Most authors agree, that the value for recovery is not important, as long as the data for LLOQ, (LOD), precision and accuracy (bias) are acceptable [1, 2, 4, 6, 10, 14]. It can be calculated by comparison of the analyte response after sample workup with the response of a solution containing the analyte at the theoretical maximum concentration. Therefore absolute recoveries can usually not be determined if the sample workup includes a derivatization step, as the derivatives are usually not available as reference substances. Nevertheless, the guidelines of the Journal of Chromatography B require the determination of the recovery for analyte and internal standard at high and low concentrations [11].

4.7. Ruggedness (Robustness)

Ruggedness is a measure for the susceptibility of a method to small changes, that might occur during routine analysis like small changes of pH values, mobile phase composition, temperature etc. Full validation must not necessarily include ruggedness testing; it can however be very helpful during the method development/prevalidation phase, as problems that may occur during validation are often detected in advance. Ruggedness should be tested, if a method is supposed to be transferred to another laboratory [6-8, 10].

5. Implications for Forensic and Clinical Toxicology

Almost all of the above mentioned publications referred to bioanalytical methods for bioavailability, bioequivalence or pharmacokinetic studies. This field is of course very closely related to forensic and clinical toxicology, especially if only routine methods are considered. Therefore, it seems reasonable to base the discussion concerning method validation in toxicological analysis on the experiences and consensus described above and not to start the whole discussion anew. In the following, possible implications for forensic and clinical toxicology will be discussed.

5.1. Terminology

As already mentioned above, there are several sets of terminology in literature. It is therefore strongly recommended to adopt in principle one of these sets for validation in clinical and forensic toxicology and add slight modifications, where it seems necessary. The definitions established by the ICH seem to be a reasonable choice as they are consensus definitions of an international conference and easily available on the homepage of ICH (www.ich.org).

5.2. Validation parameters

5.2.1. Selectivity (Specificity)

During pharmakokinetic studies (therapeutic) drugs are usually ingested under controlled conditions. Therefore, there is no need to prove the ingestion of this drug. Due to this fact the selectivity evaluation can be based on the acceptability of precision and accuracy data at the LLOQ. This approach is quite problematic for forensic and clinical toxicology, where analysis is often performed to prove ingestion of an (illicit) substance and therefore, qualitative data are also important. Here the approach to prove selectivity by absence of signals in blank samples makes much more sense. The confinement of Conference Report II [14] to only study one source of blank matrix for methods employing MS detection does not seem reasonable for toxicological applications because of the great importance of selectivity in this field. However, discussion is needed how many sources of blank samples should be analyzed and if this should depend on the detection method.

It seems reasonable to also check for interferences from other xenobiotics, that can be expected to be present in authentic samples (e.g. other drugs of abuse for methods to determine MDMA, other neuroleptics for methods to determine olanzapine). This can be accomplished by spiking these possible interferents at their highest expectable concentrations into blank matrix and checking for interferences after analysis.

5.2.2. Calibration model

The use of matrix based calibration standards seems also important in toxicological analysis, in order to account for matrix effects during sample workup and measurement (e.g. by chromatographic methods). Consensus should be reached on how many concentration levels and how many replicates per level should be analyzed. From our own experience six levels with six replicates each seems reasonable. Weighted calibration models will generally be the most appropriate in toxicological analysis, as concentration ranges of analytes in toxicological samples are usually much greater than in samples for pharmacokinetic studies. Homoscedasticity, a prerequisite for unweighted models, can however only be expected for small calibration ranges.

5.2.3. Accuracy (precision and bias)

There is no obvious reason to evaluate these parameters in another way than has been described above. Due to the often higher concentration ranges, it might be reasonable to also validate the analysis of QC samples containing concentrations above the highest calibration standard after dilution or after reduction of sample volumes, as it has been described by Wieling et al. [17] and Dadgar et al. [4]. The latter has also described the use of QC samples with concentrations below those of the lowest calibration standard using greater sample volumes.

5.2.4. Limits

The same approaches and criteria as those described above (4.4.) could be used. All approaches have been described to lesser or greater extent in international publications, especially for the determination of LOD. Nevertheless, it seems important to reach consensus on this matter at least for forensic and clinical toxicology, as reliable detection of a substance is one of the most important issues in toxicological analysis. At this point it must be stressed that for the estimation of LOD and LLOQ via a special calibration curve, the calibration samples must only contain the analyte at concentrations close to LOD and LLOQ. Use of the calibration curve over the whole range may lead to overestimation of these limits.

5.2.5. Stability

The greatest problems encountered during stability testing for bioanalytical methods in forensic and clinical toxicology is the fact, that there is great number of different sampling vessels. Furthermore, the used anticoagulants also differ. Both facts make it difficult to assess long-term stability, as the workload to analyze all possible combinations of vessels and anticoagulants is of course far to great. However, for some analytes relevant for forensic and clinical toxicology (e.g. cocaine, GHB) stability problems with differnt sampling vessels have been reported. Therefore, the relevance of this parameter for forensic and clinical toxicology has to be discussed extensively. Agreement on a single type of vessels to use for sampling of toxicological samples would probably be the easiest solution. Another problem is the fact, that storage conditions prior to arrival in the laboratory are not known. So this matter will also have to be discussed.

5.2.6. Recovery

Recovery does not seem to be a big issue for forensic and clinical toxicologists as long as precision, accuracy (bias), LLOQ and especially LOD are satisfactory. However, during method development one should of course try to optimize recovery.

5.2.7. Ruggedness

There is no obvious reason to treat this matter differently than described above (4.7.).

6. Conclusion

There are only a few principle differences concerning validation of bioanalytical methods in the fields of pharmacokinetic studies and forensic and clinical toxicology. Therefore, it seems reasonable to base the discussion on validation in the field of toxicology on the experiences and consensus already existing in the closely related field of pharmacokinetic studies for registration of pharmaceuticals and focus the discussion on those parameters, which are of special importance for toxicologists, i.e. selectivity, LOD, LLOQ and stability.

7. References

[1] Bressolle F, Bromet PM, Audran M (1996) Validation of liquid chromatographic and gas chromatographic methods. Applications to pharmacokinetics. J.Chromatogr.B 686:3-10.

[2] Causon R (1997) Validation of chromatographic methods in biomedical analysis. Viewpoint and discussion. J.Chromatogr.B 689:175-180.

[3] Dadgar D, Burnett PE (1995) Issues in evaluation of bioanalytical method selectivity and drug stability. J.Pharm.Biomed.Anal. 14:23-31.

[4] Dadgar D, Burnett PE, Choc MG, Gallicano K, Hooper JW (1995) Application issues in bioanalytical method validation, sample analysis and data reporting. J.Pharm.Biomed.Anal. 13:89-97.

[5] Hartmann C, Massart DL, McDowall RD (1994) An analysis of the Washington Conference Report on bioanalytical method validation. J.Pharm.Biomed.Anal. 12:1337-1343.

[6] Hartmann C, Smeyers-Verbeke J, Massart DL, McDowall RD (1998) Validation of bioanalytical chromatographic methods. J.Pharm.Biomed.Anal. 17:193-218.

[7] International Conference on Harmonization (ICH). Va lidation of Analytical Methods: Definitions and Terminology. ICH Q2 A. 1994.

[8] International Conference on Harmonization (ICH). Validation of Analytical Methods: Methodology. ICH Q2 B. 1996.

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

[10] Karnes HT, Shiu G, Shah VP (1991) Validation of bioanalytical methods. Pharm.Res. 8:421-426.

[11] Lindner W, Wainer IW (1998) Requirements for initial assay validation and publication in J. Chromatography B [editorial]. J.Chromatogr.B 707:1-2.

[12] NCCLS. Evaluation of Precision Performance of Chlinical Chemistry Devices; Approved Guideline. NCCLS document EP5-A. 1999.

[13] Shah VP, Midha KK, Dighe S, McGilveray IJ, Skelly JP, Yacobi A, Layloff T, Viswanathan CT, Cook CE, McDowall RD, Pittman KA, Spector S (1992) Analytical methods validation: bioavailability, bioequivalence and pharmacokinetic studies. Conference report. Pharm.Res. 9:588-592.

[14] Shah VP, Midha KK, Findlay JW, Hill HM, Hulse JD, McGilveray IJ, McKay G, Miller KJ, Patnaik RN, Powell ML, Tonelli A, Viswanathan CT, Yacobi A (2000) Bioanalytical method validation--a revisit with a decade of progress. Pharm.Res. 17:1551-1557.

[15] U.S.Department of Health and Human Services, Food and Drug Administration. Guidance for Industry, Bioanalytical Method Validation. http://www.fda.gov/cder/guidance/index.htm. 2001.

[16] W.Penninckx, C.Hartmann, D.L.Massart, J.Smeyers-Verbeke (1996) Validation of the Calibration Procedure in Atomic Absorption Spectrometric Methods. J.Anal.At.Spectrom. 11:237-246.

[17] Wieling J, Hendriks G, Tamminga WJ, Hempenius J, Mensink CK, Oosterhuis B, Jonkman JH (1996) Rational experimental design for bioanalytical methods validation. Illustration using an assay method for total captopril in plasma. J.Chromatogr.A 730:381-394.