Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International 165 (2007) 216–224
Forensic Science International
www.elsevier.com/locate/forsciint

Validation of new methods

Frank T. Peters [a,*], Olaf H. Drummer [b], Frank Musshoff [c]

[a] Department of Experimental and Clinical Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, University of Saarland, Building 46, D-66421 Homburg (Saar), Germany
[b] Victorian Institute of Forensic Medicine and Department of Forensic Medicine, Monash University, 57-83 Kavanagh Street, Southbank, Melbourne 3006, Australia
[c] Institute of Legal Medicine, Rheinische Friedrich-Wilhelms-University, Stiftsplatz 12, D-53111 Bonn, Germany

Received 30 January 2006; accepted 10 May 2006
Available online 16 June 2006

Abstract

Reliable analytical data are a prerequisite for correct interpretation of toxicological findings in the evaluation of scientific studies, as well as in daily routine work. Unreliable analytical data 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. Therefore, new analytical methods to be used in forensic and/or clinical toxicology require careful method development and thorough validation of the final method. This is especially true in the context of quality management and accreditation, which have become matters of increasing relevance in analytical toxicology in recent years. In this paper, important considerations in analytical method validation will be discussed which may be used as guidance by scientists wishing to develop and validate analytical methods.

Keywords: Validation; Bioanalysis; Clinical and forensic toxicology

1. Introduction

Reliable analytical data are a prerequisite for correct interpretation of toxicological findings in the evaluation of scientific studies, as well as in daily routine work. In scientific work, unreliable results may lead to over- or underestimation of effects, to false interpretations, and to unwarranted conclusions. If such errors are not obvious, they might remain undetected during a scientific study or case investigation. Unless officially contested by other experts in the field, they may thus be multiplied within the scientific community, or become part of the accepted general knowledge in a certain area of research and thus cause further misinterpretations. In routine daily work in clinical and forensic toxicology, unreliable analytical data 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. Because of all these considerations the importance of high quality analytical data cannot be overestimated.

The basis for high quality data is reliable analytical methods. Therefore, new analytical methods to be used in clinical and/or forensic toxicology require careful method development to be followed by a thorough validation of the final method. At this point, it must be stated that the quality of an analytical method largely depends on method development and not so much on its validation. Therefore, it is imperative that the method itself is also fit for purpose. However, only validation can objectively demonstrate the inherent quality of an analytical method by fulfilment of minimum acceptance criteria and thus prove its applicability for a certain purpose. Since such objective performance data are essential for assessment of a method's quality, the importance of validation, at least of routine analytical methods, cannot be overestimated. This is especially true in the context of quality management and accreditation, which have become matters of increasing relevance in analytical toxicology in recent years. Not surprisingly, peer-reviewed scientific journals have reacted by increasing requirements concerning method validation.

In this paper, important considerations in analytical method validation will be discussed. These considerations may be used as guidance by scientists wishing to develop and validate new analytical methods. These will then also be suitable for publication in peer-reviewed journals.

* Corresponding author. Tel.: +49 6841 16 26430; fax: +49 6841 16 26051.
E-mail address: frank.peters@uniklinikum-saarland.de (F.T. Peters).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2006.05.021

2. Method validation

Owing to the importance of method validation in the whole field of analytical chemistry, a number of guidance documents on this subject have been issued by various international organizations and conferences [1–7]. All of these documents are important and potentially helpful for any method validation. However, only few specifically address analysis of drugs, poisons, and/or their metabolites in body fluids or tissues [1,5,7]. Of these, the most influential guidance documents are the reports on the conference on ''Analytical methods validation: bioavailability, bioequivalence and pharmacokinetic studies'' held in Washington in 1990 (Conference Report) [1] and the follow-up conference in 2000 (Conference Report II) [5], in which experiences and progress since the first conference were discussed. Both of these reports were published by Shah et al. and had an enormous impact on validation of bioanalytical methods in the pharmaceutical industry. Because of the close relation to bioanalysis in the context of bioavailability, bioequivalence and pharmacokinetic studies, Conference Report II is probably also the most useful guidance paper for bioanalytical method validation in clinical and forensic toxicology. It was therefore also used as basis for the guidance document recently issued by the German-speaking Society of Toxicological and Forensic Chemistry (GTFCh) [7], and to the authors' knowledge, the only available comprehensive guideline specifically addressing method validation in analytical toxicology.

Besides these official guidance documents, a number of review articles have been published on the topic of analytical method validation [8–13]. Again, all of these papers are interesting and helpful for any method validation, while only part of them specifically address analysis of drugs, poisons, and/or their metabolites in body fluids or tissues [8,9,13]. This includes the excellent review on validation of bioanalytical chromatographic methods which includes detailed discussions of theoretical and practical aspects [8]. The other two deal with the implications of bioanalytical method validation in clinical and forensic toxicology [9] and with theoretical and practical aspects in method validation using LC–MS(/MS) [13]. The latter also describes a proposed experimental design for validation experiments, as well as statistical procedures for calculating validation parameters.

2.1. Validation parameters

Analytical methods in clinical and forensic toxicology may either be used for screening and identification of drugs, poisons and/or their metabolites in biological fluids or tissues, for their quantification in these matrices, or for both. For quantitative bioanalytical procedures, there is a general agreement, that at least the following validation parameters should be evaluated: selectivity, calibration model (linearity), stability, accuracy (bias), precision (repeatability, intermediate precision) and the lower limit of quantification (LLOQ). Additional parameters which may be relevant include limit of detection (LOD), recovery, reproducibility, and ruggedness (robustness) [1,5,8,14]. For qualitative procedures, a general validation guideline is currently not available [12], but there seems to be agreement that at least selectivity and the LOD should be evaluated and that additional parameters like precision, recovery and ruggedness (robustness) might also be important [3,12,15,16]. For methods using LC–MS, experiments for assessment of possible matrix effects (ME), i.e. ion suppression or ion enhancement, should always be part of the validation process, particularly if they employ electrospray ionisation (ESI) [5,13,17–19].

Another issue often raised in the context of analytical method validation is measurement uncertainty. However, measurement uncertainty is not a validation parameter of itself. It can be obtained from validation data, most notably bias and precision, or from routine QC data obtained during application of the analytical method. For this reason, it is not covered in detail here. Interested readers are referred to the EURACHEM/CITAC guide to quantifying uncertainty in analytical measurement [20].

2.1.1. Selectivity

In the Conference Report II [5], selectivity was defined as ''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.''. It should be noted that the term specificity is often used interchangeably with selectivity, although in a strict sense this is not correct [21]. One approach to establish method selectivity is to prove the lack of response in blank matrix [1,5,7–9,11,14], i.e. that there are no signals interfering with the signal of the analyte(s) or the IS. The second approach is based on the assumption that for merely quantitative procedures, small interferences can be accepted as long as accuracy (bias) and precision at the LLOQ remain within certain acceptance limits [4,6,8,9,11]. However, in clinical and forensic toxicology, analysis is often mainly performed to prove the intake of a substance and qualitative data are, therefore, most important. Here, the approach to prove selectivity by absence of interfering signals seems much more reasonable [9].

While the requirement established by the Conference Report [1] to demonstrate selectivity by analysing at least six different sources of blank matrix has become state-of-the-art during the last decade, the probability of relatively rare interferences remaining undetected is rather high when only analysing such a small number of matrix blanks [8]. For this reason, it has been proposed to evaluate at least 10–20 sources of blank samples [22], which seems reasonable regarding the great importance of selectivity in the field of analytical toxicology and that even relatively rare matrix interferences are not unlikely to occur if large numbers of samples are analysed in routine application.

Because samples from clinical and forensic toxicology cases often contain many different drugs, poisons and/or their metabolites, it may also be important to check for possible interferences from other xenobiotics which may be expected to be present in authentic samples [9]. This can be accomplished by analysing blank samples spiked with possibly interfering

compounds at their highest expectable concentrations. Another way to exclude interference from other drugs or their metabolites is to check authentic samples containing these but not the analyte of interest. This approach is preferable, if the suspected interfering substance is known to be extensively metabolized, as it also allows excluding interferences from such metabolites, which are usually not available as pure substances.

Stable-isotope-labelled analogues of the target analytes are often used as internal standard (IS) in MS-based methods. They can ideally compensate for variability during sample preparation and measurement, but still be differentiated from the target analyte by mass spectrometric detection. However, isotopically labelled compounds may contain the non-labelled compound as an impurity or their mass spectra may sometimes contain fragment ions with the same mass-to-charge ratios (m/z) as the monitored ions of the target analyte. In both cases, the peak area of the analyte peak would be overestimated, thus compromising quantification. The absence of such interference caused by the IS should be checked by analysing so-called zero samples, i.e. blank samples spiked with the IS. In a similar way as described above, the analyte might interfere with a stable-isotope-labelled IS. This even becomes a principle problem with deuterated analogues when the number of deuterium atoms of the analogue or one of its monitored fragments is three or less [23]. Blank samples spiked with the analyte at the upper limit of the calibration range, but without IS, can be used to check for absence of such interferences.

An important problem that may affect the IS, even though in strict sense not because of insufficient analytical selectivity, is the use of therapeutic drugs as IS, as it is often described in literature [17]. This practice is not acceptable, because even if a drug is not marketed in a certain country, its presence in the specimen to be analysed can never be fully excluded in times of increasing international travel and globalization. In this case, the signal of the IS will be overestimated, inevitably leading to underestimation of the analyte concentration in the sample. In order to avoid such problems, the chosen IS must never be a therapeutic drug. In methods employing mass spectrometry, the IS should always be chosen from the pool of available isotopically labelled compounds. These are available in a large variety of structures with different physicochemical properties so it should not be a problem to find an appropriate IS in this pool of compounds. The only drawback of the isotopically labelled analogues as IS is that they may contribute to the analyte signal to a certain extent if the number of isotope labels is low (see above).

2.1.2. Calibration model (linearity)

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 response (in bioanalytical methods mostly the area ratio of analyte versus IS) must be investigated. A detailed discussion on the strategy for selecting an appropriate calibration model is beyond the scope of article, so only the most important aspects are discussed here. Interested readers will find more details in Refs. [8,13].

There is general agreement that for bioanalytical methods, calibrators should be matrix-based, i.e. prepared by spiking of blank matrix. Calibrator concentrations must cover the whole calibration range [5,7–9,13] and should be evenly spaced across it [6–8]. Most guidelines require a minimum of five to eight concentration levels [3–7,9,11,14] and some specify that at least two to six replicates should be analysed per level [6–8]. Generally, it is advisable to use fewer concentration levels with more replicates than vice versa [8]. After a sufficient number of calibration levels have been measured with a sufficient number of replicates, it is necessary to find a mathematical model that adequately describes the relationship between analyte concentration in the samples and response. Usually, linear models are preferable, but, if necessary, the use of non-linear (e.g. second order) models can be used. In many publications in the field of analytical toxicology, the use of a linear ordinary, i.e. un-weighted, least squares regression model is reported, which is not appropriate in many cases. Ordinary least squares regression models are only applicable for homoscedastic data sets, i.e. if there is homogeneity of variances across the calibration range. As a rule of thumb, this can be expected for calibration ranges spanning not more than one order of magnitude. However, most methods in analytical toxicology span at least two or three orders of magnitude which are usually associated with significant heteroscedasticity. In such cases the data should mathematically be transformed or a weighted least squares model should be applied [4–6,8]. Usually, the factors $1/x$ or $1/x^2$, i.e. the inverse of the concentration or the inverse of the squared concentration, respectively, adequately compensate for heteroscedasticity.

The appropriateness of the chosen calibration model should always be confirmed by statistical tests for model fit. For details see Ref. [24]. The widespread practice of evaluating a calibration model via its coefficients of correlation or determination is not acceptable from a statistical point of view [8]. 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 [8].

2.1.3. Accuracy (bias) and precision

In a strict sense, the accuracy of a method is affected by systematic (bias) as well as random (precision) error components [8], but the term is often used to describe only the systematic error component, i.e. in the sense of bias [1,2,5,14,21]. In the following, the term accuracy will be used in the sense of bias, which will be indicated in brackets. Bias is ‘‘the difference between the expectation of the test results and an accepted reference value’’ [25]. It is usually expressed as a percent deviation from the accepted reference value. 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’’ [2]. It is usually measured in terms of imprecision, expressed as an absolute or relative standard deviation (R.S.D.) and does not relate to reference

values. ''Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability is also termed intra-assay precision'' [2]. Within-run or within-day precision are also often used to describe repeatability. ''Intermediate precision expresses within-laboratories variations: different days, different analysts, different equipment, etc.'' [2]. In a strict sense intermediate precision is the total precision under varied conditions, whereas so-called inter-assay, between-run or between-day precision only measure the precision components caused by the respective factors (see below). However, the latter terms are not clearly defined and obviously often used interchangeably with each other and also with the term intermediate precision. ''Reproducibility expresses the precision between laboratories (collaborative studies, usually applied to standardization of methodology)'' [2]. Reproducibility only has to be studied, if a method is supposed to be used in different laboratories. Unfortunately, some authors also use the term reproducibility for within-laboratory studies at the level of intermediate precision, which should be avoided to prevent confusion.

Precision and bias can be estimated from the analysis of QC samples under specified conditions. The QC samples should ideally be obtained from an independent source rather than produced by the same person(s) performing the validation. This includes preparing a batch(s) from a different batch of standard. This approach more properly assesses any bias due to weighing, dilution or even impurities in standards.

QC samples should be evaluated at least near the extremes of the calibration range, but preferably also near the middle [1,3,5,7,8,11,14]. In clinical and forensic toxicology, it seems reasonable to also quantify concentrations above the highest calibration standard after appropriate dilution or after reduction of sample volumes as described by Dadgar et al. [26]. The acceptance criteria for precision (15% R.S.D., 20% R.S.D. near LLOQ) and accuracy (bias within $\pm$ 15% of the accepted reference value, within $\pm$ 20% near LLOQ) specified by the Conference Reports have been widely accepted in bioanalysis [1,5].

There are two principle approaches for precision data. In the first approach, five [1,5] to ten [4] independent determinations are performed for each concentration level under each repeatability and intermediate precision conditions. In this straightforward approach, the corresponding precision data can simply be calculated as R.S.D.s of the values obtained under these stipulated conditions. The corresponding bias values can be calculated as the percent deviations of the observed mean values from the respective reference values. This yields two bias values for each concentration level, one from the mean value of the repeatability experiment (within-day or within-run bias), and one from the mean of the factor-different intermediate precision experiment (between-day or between-run bias).

In the second approach, replicates at each concentration level are analysed on each of a number of different occasions, e.g. duplicate measurements at each concentration level on each of eight days [8] or in a number of successive runs [6]. Using one-way analysis of variance (ANOVA) with the varied factor (e.g. day, run) as the grouping variable, such experimental designs allow separate calculation of repeatability and of the precision component caused by the grouping variable from the same data sets, as well as the calculation of factor-different intermediate precision as the combination of the previous two. For details of calculations see Refs. [7,13,27–29]. In this approach, the bias values can be calculated as the percent deviation of the grand means at each concentration level from the respective reference values. In the opinion of the authors, the second approach is preferred, because more information can be derived from this approach with a comparable number of analyses. Furthermore, the repeatability estimate obtained in this approach is based on data acquired not only on a single but on several occasions and thus more likely reflect the actual repeatability performance of the method.

Unfortunately, in many publications, a mixture of the abovementioned approaches is used. The experimental design is the same as in the second approach, with replicates being analysed on several occasions, but calculations are performed as in the first approach [4]. Only the replicates from the first occasion are considered when estimating repeatability, which is statistically correct but still a waste of valuable information, because the other occasions are not taken into account. The calculation of intermediate precision as the (R.)S.D. of all observations is even more problematic. This way of calculation treats all observations as independent, which they are in fact not, because they were obtained in groups on several occasions. This is not only statistically incorrect, but leads to underestimation of intermediate precision [27]. For these reasons, the mixed approach should not be used in any validation study. Furthermore, publications describing validated assays should not only report the experimental design and results of precision studies but also a detailed description on how the results were calculated. Otherwise comparison between precision data in different publications will not be possible.

2.1.4. 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) [2,5]. There are different approaches for the determination of LLOQ. In the first and probably most practical approach, the LLOQ is defined as the lowest concentration of a sample that can still be quantified with acceptable precision and accuracy (bias) [1–3,5,8]. In the Conference Reports [1,5], the acceptance criteria for these two parameters at LLOQ are 20% R.S.D. for precision and $\pm$20% for bias. In the second and also very common approach, the LLOQ is estimated based on the signal-to-noise ratio (S/N) [3,14]. S/N can 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. A third approach to estimate the LLOQ is the concentration that corresponds to a response that is k-times greater than the estimated S.D. of blank samples [3], where a k-factor of 10 is usually applied. The concentration is obtained by dividing this response by the slope of the calibration curve. This approach is only applicable for

methods where $S.D._{bl}$ 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 analysed with a selective method. Finally, the LLOQ can be estimated using a specific calibration curve established using calibration samples containing the analyte in the range of LLOQ. For details and calculations, see Refs. [3,7]. One must not use the calibration curve over the whole range of quantification for this determination, because this may lead to overestimation of the LLOQ.

2.1.5. Upper limit of quantification (LLOQ)

The ULOQ 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 [5].

2.1.6. Limit of detection (LOD)

Quantification below LLOQ is by definition not acceptable [1–3,5,8]. Therefore, below this value a method can only produce semiquantitative or qualitative data. However, particularly in analytical toxicology, it can be very important to know the LOD of the method, which can be defined as the lowest concentration of analyte in a sample which can be detected but not necessarily quantified as an exact value [2] or as the lowest concentration of an analyte in a sample, that the bioanalytical procedure can reliably differentiate from background noise [5]. The approaches most often applied for estimation of the LOD are basically the same as those described for LLOQ with the exception of the approach using precision and accuracy data, which cannot be used here for obvious reasons. In contrast to the LLOQ determination, for LOD a S/N or *k*-factor equal to or greater than three is usually chosen [3,4,6,8,11]. At this point it must be noted that all these approaches only evaluate the pure response of the analytes. In toxicology, however, unambiguous identification of an analyte in a sample requires more complex acceptance criteria to be fulfilled. Such criteria have recently been reviewed by Rivier [16]. Especially in forensic toxicology and doping control, it would certainly be more appropriate to define the LOD as the lowest concentration of analyte in a sample, for which specific identification criteria can still be fulfilled.

2.1.7. Stability

In Conference Report II stability was defined as follows [5]: ''The chemical stability of an analyte in a given matrix under specific conditions for given time intervals''. Stability of the analyte during the whole analytical procedure is a prerequisite for reliable quantification. Unless data on analyte stability are available in the literature, full validation of a method must include stability experiments for the various stages of analysis including storage prior to analysis. For 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 storage period at least as long as the one expected for authentic samples [1,5,8,14,21,22,26]. Since samples are often frozen and thawed, e.g. for reanalysis, freeze/thaw stability of analytes should be evaluated over at least three freeze/thaw cycles at two concentrations in triplicate [1,5,8,22]. In-process or bench-top stability is the stability of analyte under the conditions of sample preparation (e.g. ambient temperature over time needed for sample preparation). This type of stability should be evaluated to find out, if preservatives have to be added to prevent degradation of analyte during sample preparation [5,8,22]. Finally, instability cannot only occur in the sample matrix, but also in processed 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 [5,8,22,26]. A detailed account of experimental designs and statistical evaluations of stability experiments is beyond the scope of this article and can be found in Refs. [8,22,26].

2.1.8. Recovery

As already mentioned above, recovery is not among the validation parameters regarded as essential for method validation. 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 [5,8,26]. It can be calculated as the percentage of the analyte response after sample workup compared to that of a solution containing the analyte at a concentration corresponding to 100% recovery. Therefore, absolute recoveries can usually not be determined if the sample workup includes a derivatization step, as the derivatives are often not available as reference substances. Nevertheless, some guidance documents request the determination of the recovery at high and low concentrations [7,14] or even specify that the recovery should be greater than 50% [7].

In LC–MS(–MS) analysis, a different experimental design must be used to determine recovery, because part of the change of the response in prepared samples in comparison to respective standard solutions might be attributable to ME. In the validation of LC–MS(–MS), it is therefore more appropriate to perform the recovery experiments together with ion suppression/enhancement experiments as described below.

2.1.9. Ruggedness (robustness)

Ruggedness is a measure for the susceptibility of a method to small changes that might occur during routine analysis, e.g. small changes of pH values, mobile phase composition, temperature, etc. Full validation must not necessarily include ruggedness testing, but it should be performed if a method is to be transferred to another laboratory [2,3,21,30,31]. Furthermore, it can be very helpful during the method development/pre-validation phase, as problems that may occur during validation are often detected in advance. A detailed account and helpful guidance on experimental designs and evaluation of ruggedness/robustness tests can be found in Ref. [31].

2.1.10. Matrix effects (ion suppression/enhancement)

Suppression or enhancement of analyte ionisation by co-eluting compounds is a well known phenomenon in LC–MS(–MS) analysis mainly depending on the sample matrix, the sample preparation procedure, the quality of chromatographic separation, mobile phase additives, and ionisation type [18,19,32–39]. While ESI has been reported to be much more prone to such effects, they may also occur with atmospheric pressure chemical ionisation (APCI) [18,19,32–35,37,38]. It is obvious that ion suppression, as well as ion enhancement, may affect validation parameters such as LOD, LLOQ, linearity, precision and/or bias – the latter three especially in the absence of an isotopically labelled IS. Therefore, studies of ion suppression/enhancement should be an integral part of the validation of any LC–MS(–MS) method. In the literature, two approaches have been used to study ion suppression/enhancements. In the first approach, a solution of the analyte is constantly infused into the eluent from the column via post-column tee connection using a syringe pump. The continuous post-column infusion leads to a constant signal in the detector, unless compounds that elute from the column suppress or enhance ionisation, which would lead to a decreased or increased detector response, respectively [32–36,40–42]. A comprehensive strategy for the second approach was recently published by Matuszewski et al. [19]. This paper provides excellent guidance on how to perform and evaluate studies on ME in LC–MS(–MS) analysis. The principle approach involves determination of peak areas of analyte in three different sets of samples, one consisting of neat standards (set 1), one prepared in blank matrix extracts from different sources and spiked after

Table 1
Summary of proposed experiments, evaluations, and acceptance criteria for full validation of new analytical methods intended for routine use

Validation parameter	Experiments	Evaluation	Acceptance criteria
Selectivity	Analysis of at least 6, preferably 10–20 sources of blank matrix; analysis of 1–2 zero samples (blank matrix + internal standard); analysis of blank samples spiked with compounds expected to be present in real samples, if applicable; analysis of real samples with suspected interference but without analyte, if applicable	Checking for interfering signals	Absence of interfering signals
Calibration model	At least 4–6 concentration levels; analysis of at least 5–6 replicates per level	Identification/elimination of outliers; analysis of behavior of variances across calibration range; evaluation of weighting factor in case of heteroscedastic data; evaluation of linear model; evaluation of non-linear model, if linear model not applicable; statistical test of model fit	Statistical model fit; acceptable accuracy and precision data
Accuracy (bias) and precision	QC samples at low, medium and high concentrations relative to calibration range; analysis of duplicates on each of 8 different days	Calculation of bias as percent deviation of mean calculated value from nominal value; calculation of precision data using ANOVA approach	Bias within ±15% of nominal value (±20% near LLOQ); precision within 15% R.S.D. (20% near LLOQ)
LOD	Analysis of spiked samples with decreasing concentrations of analyte	Checking for compliance with identification criteria; alternatively, evaluation of S/N	Compliance with identification criteria; alternatively, S/N $\geq$ 3
LLOQ	Use of accuracy and precision experiments with QC samples with analyte concentration near LLOQ; alternatively, analysis of spiked samples with decreasing concentrations of analyte	Use of accuracy and precision data of QC samples with analyte concentration near LLOQ; alternatively, evaluation of S/N in spiked samples	Compliance with accuracy and precision criteria near LLOQ, see above; alternatively, S/N $\geq$ 10
Recovery/extraction efficiency	Low and high concentrations relative to calibration range; extraction of 5–6 spiked samples, preferably using different blank matrix sources; analysis of 5–6 100% controls	Calculation of recovery as percentage of response in extracted samples as compared to control samples; calculation of respective R.S.D.	Acceptable sensitivity; reproducible recovery
Processed sample stability	Processing of samples containing low and high analyte concentrations; repeated injection of processed samples at certain time intervals	Linear regression analysis of response plotted vs. injection time	Negative slope significantly different from 0 indicates instability
Freeze/thaw stability	QC samples at low and high concentrations relative to calibration range; analysis of six replicates before (control) and six replicates after three freeze/thaw cycles (treatment)	Calculation of percentage of mean concentration in treated samples as compared to mean concentration in control samples; calculation of respective 90% confidence interval	Stability assumed when mean of treated samples within 90–110% of control samples and 95% CI within 80–120% of control mean

For additional validation experiments on matrix effect to be performed for LC–MS(/MS)-based methods see Section 2.1.10.

extraction (set 2), and one prepared in blank matrix from the same sources but spiked before extraction (set 3). From these data, one can then calculate the ME (ion suppression/enhancement) as a percentage of the response of set 2 samples in relation to those of set 1 samples, the recovery as a percentage the response of set 3 samples in relation to that of set 2 samples, and finally the so-called process efficiencies as a percentage of the response of set 3 samples in relation to set 1 samples. For details on the experimental design and calculations see Ref. [19].

With two well-established procedures for studying ME, the question arises which one of them is best suited for validation studies. In the authors' opinion, the post-column infusion experiments are very useful during method development, because it provides information on the retention times where ion suppression/enhancement is expected, which can then be avoided by optimizing the separation system. For a validation study, the alternative approach seems to be more suitable, because it yields a quantitative estimation of ME and their variability and is thus more objective. However, no matter which approach is used in validation studies, it is essential to evaluate several sources of blank matrix [18,19] just as it has been described for the selectivity experiments.

2.2. Experimental design for full method validation

While quite a number of guidance documents and reviews on analytical method validation are available in the literature, there is little guidance of practical aspects of designing validation experiments. A paper on a rational experimental design for bioanalytical methods validation was first published by Wieling et al. [43]. This experimental design was later modified by Peters based on considerations discussed in his review article on method validation [9]. The modified design has become the basis for method validation in a series of methods published by the working group of Maurer in the last few years [44–50]. A detailed description of the latest version of this experimental design, which also contains ion suppression/enhancement experiments is described in Ref. [13]. A summary of proposed experiments, evaluations, and acceptance criteria for a full validation study is given in Table 1. It is recommended to start the validation studies with the selectivity and ion suppression/enhancement experiments, because if any of these two parameters are not acceptable, major changes of the method might be required. If the method is found to be selective and free of relevant ME, the processed sample stability should be assessed to ensure stability of processed samples under the conditions on the autosampler tray during analysis of large batches of samples. If the processed samples are stable, one can proceed to the linearity experiments and evaluation of the calibration model (linearity experiments). After establishing an appropriate calibration model, the early validation phase is complete and the main validation phase can be started, in which bias and precision as well as freeze/thaw stability are evaluated.

2.3. Validation of methods to be used for single case analysis

The full validation of a new analytical method is associated with a considerable workload. The experimental design described in Ref. [13] comprises a total number of more than 200 sample injections, although the number of samples has been reduced to the minimum number required for sound statistics and reliable estimations of validation parameters. Such extensive validation studies are certainly justified if an analytical method is to be used routinely, e.g. for analysis of

Table 2
Summary of proposed experiments, evaluations, and acceptance criteria for validation of new analytical methods to be used in single case studies or for analysis of rare analytes

Validation parameter	Experiments	Evaluation	Acceptance criteria
Selectivity	Analysis of at least six sources of blank matrix; analysis of 1–2 zero samples (blank matrix + internal standard)	Checking for interfering signals	Absence of interfering signals
Calibration model	At least 4–5 concentration levels; analysis of duplicates at each level	Evaluation of linear model; evaluation of non-linear model, if linear model not applicable	Acceptable accuracy and precision data
Accuracy (bias) and precision	QC samples at low and high concentrations relative to calibration range; analysis of 5–6 replicates per level under repeatability conditions	Calculation of bias as percent deviation of mean calculated value from nominal value; calculation of precision data as R.S.D.	Bias within ±15% of nominal value (±20% near LLOQ); precision within 15% R.S.D. (20% near LLOQ)
LOD	Analysis of spiked samples with decreasing concentrations of analyte	Checking for compliance with identification criteria; alternatively, evaluation of S/N	Compliance with identification criteria; alternatively, S/N $\geq$ 3
LLOQ	Use of accuracy and precision experiments with QC samples with analyte concentration near LLOQ; alternatively, analysis of spiked samples with decreasing concentrations of analyte	Use of accuracy and precision data of QC samples with analyte concentration near LLOQ; alternatively, evaluation of S/N in spiked samples	Compliance with accuracy and precision criteria near LLOQ, see above; alternatively, S/N $\geq$ 10

For additional validation experiments on matrix effect to be performed for LC–MS(/MS)-based methods see Section 2.1.10.

common drugs of abuse in driving under the influence of drugs cases in forensic toxicology, for therapeutic drug monitoring of antidepressants in clinical toxicology, or for analysing samples from pharmacokinetic studies in research projects. Furthermore, extensive validation is required if the method is to be published.

For methods used for analysis of rare analytes or in publications of case reports, it is certainly acceptable to reduce the extent of validation experiments. A summary of proposed experiments, evaluations, and acceptance criteria in such situations is given Table 2. Here, it would not be necessary to perform stability and recovery experiments. In addition, the number of concentration levels and replicates in the linearity experiments might be kept at minimum, e.g. four to five levels with duplicate measurements at each level. In some situations it might even be sufficient to use one point calibration. Finally, it should be acceptable to limit the precision and accuracy experiments to evaluation of repeatability and bias at two concentration levels with five or six replicates per level. However, a minimum of validation experiments are necessary even in the mentioned situation, because the reliability of single case data must be ensured, considering that only such data are available for some analytes.

2.4. Validation in analysis of partly decomposed or putrefied post-mortem specimens

A special situation in the context of method validation is post-mortem analysis of partly decomposed or putrefied samples since the composition of such specimens may vary considerably from case to case [51]. It is questionable if the validation parameters acquired in a validation study using matrix samples from one or several post-mortem cases would be representative for others. For this reason, a useful approach would be to use the method of standard addition [52], in which calibration and quantification are performed in the sample matrix of the case in question. The standard error of the predicted concentration in the sample might be used as a rough estimation of precision. In these post-mortem cases accurate quantification is not necessary since the degradation of the specimen has already affected the drug concentration. Hence, all such data should be reported as approximate no matter how well the analysis has been conducted.

3. Conclusions

Validation of new methods in analytical toxicology is an integral part of quality assurance, accreditation, and publication. Methods intended for routine use or publication must be fully validated to objectively demonstrate their applicability for the intended use. For methods used for analysis of rare analytes or in single cases, the number of validation experiments and parameters to be evaluated may be reduced, but a minimum validation is necessary to ensure sufficient quality of the results. Only for post-mortem analysis, method validation is problematic, because not only is acquisition of representative validation data virtually impossible owing to varying composition of samples but also the concentration of drug is no longer what it was at the time of death.

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, Analytical methods validation: bioavailability, bioequivalence and pharmacokinetic studies. Conference report, Pharm. Res. 9 (1992) 588–592.
[2] International Conference on Harmonization (ICH), Validation of analytical methods: definitions and terminology, ICH Q2 A, 1994.
[3] International Conference on Harmonization (ICH), Validation of analytical methods: methodology, ICH Q2 B, 1996.
[4] EURACHEM, The Fitness for Purpose of Analytical Methods—A Laboratory Guide to Method Validation and Related Topics, 1998.
[5] V.P. Shah, K.K. Midha, J.W. Findlay, H.M. Hill, J.D. Hulse, I.J. McGilveray, G. McKay, K.J. Miller, R.N. Patnaik, M.L. Powell, A. Tonelli, C.T. Viswanathan, A. Yacobi, Bioanalytical method validation—a revisit with a decade of progress, Pharm. Res. 17 (2000) 1551–1557.
[6] M. Thompson, S.L.R. Ellison, R. Wood, Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC technical report), Pure Appl. Chem. 74 (2002) 835–855.
[7] F.T. Peters, M. Hartung, M. Herbold, G. Schmitt, T. Daldrup, F. Musshoff, Anlage zu den Richtlinien der GTFCh zur Qualitätssicherung bei forensisch-toxikologischen Untersuchungen, Anhang C: Anforderungen an die Durchführung von Analysen, 1.Validierung, Toxichem. Krimtech. 71 (2004) 146–154. http://www.gtfch.org/tk/tk71_3/Peters1.pdf.
[8] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, R.D. McDowall, Validation of bioanalytical chromatographic methods (review), J. Pharm. Biomed. Anal. 17 (1998) 193–218.
[9] F.T. Peters, H.H. Maurer, Bioanalytical method validation and its implications for forensic and clinical toxicology—a review (review), Accred. Qual. Assur. 7 (2002) 441–449.
[10] N.A. Epshtein, Validation of HPLC techniques for pharmaceutical analysis, Pharm. Chem. J. (Translation of Khimiko-Farmatsevticheskii Zhurnal) 38 (2004) 212–228.
[11] I. Taverniers, M. De Loose, E. Van Bockstaele, Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance, TrAC, Trends Anal. Chem. 23 (2004) 535–552.
[12] E. Trullols, I. Ruisanchez, F.X. Rius, Validation of qualitative analytical methods, TrAC, Trends Anal. Chem. 23 (2004) 137–145.
[13] F.T. Peters, Method validation using LC-MS, in: A. Polettini (Ed.), Applications of Liquid Chromatography–Mass Spectrometry in Toxicology, Pharmaceutical Press, London, in press.
[14] W. Lindner, I.W. Wainer, Requirements for initial assay validation and publication in J. Chromatography B (editorial), J. Chromatogr. B 707 (1998) 1–2.
[15] C. Jimenez, R. Ventura, J. Segura, Validation of qualitative chromatographic methods: strategy in antidoping control laboratories, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 767 (2002) 341–351.
[16] L. Rivier, Criteria for the identification of compounds by liquid chromatography–mass spectrometry and liquid chromatography–multiple mass spectrometry in forensic toxicology and doping analysis, Anal. Chim. Acta 492 (2003) 69–82.
[17] H.H. Maurer, Advances in analytical toxicology: current role of liquid chromatography–mass spectrometry for drug quantification in blood and oral fluid (review), Anal. Bioanal. Chem. 381 (2005) 110–118.
[18] T.M. Annesley, Ion suppression in mass spectrometry, Clin. Chem. 49 (2003) 1041–1044.
[19] B.K. Matuszewski, M.L. Constanzer, C.M. Chavez-Eng, Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC–MS/MS, Anal. Chem. 75 (2003) 3019–3030.
[20] EURACHEM/CITAC, Quantifying Uncertainty in Analytical Measurement, 2000.
[21] H.T. Karnes, G. Shiu, V.P. Shah, Validation of bioanalytical methods, Pharm. Res. 8 (1991) 421–426.

[22] D. Dadgar, P.E. Burnett, Issues in evaluation of bioanalytical method selectivity and drug stability, J. Pharm. Biomed. Anal. 14 (1995) 23–31.
[23] M.J. Bogusz, Large amounts of drugs may considerably influence the peak areas of their coinjected deuterated analogues measured with APCI-LC–MS (letter), J. Anal. Toxicol. 21 (1997) 246–247.
[24] W. Penninckx, C. Hartmann, D.L. Massart, J. Smeyers-Verbeke, Validation of the calibration procedure in atomic absorption spectrometric methods, J. Anal. Atom. Spectrom. 11 (1996) 237–246.
[25] International Organization for Standardization (ISO), Accuracy (Trueness and Precision) of Measurement Methods and Results, ISO/DIS 5725-1 to 5725-3, 1994.
[26] D. Dadgar, P.E. Burnett, M.G. Choc, K. Gallicano, J.W. Hooper, Application issues in bioanalytical method validation, sample analysis and data reporting, J. Pharm. Biomed. Anal. 13 (1995) 89–97.
[27] J.S. Krouwer, R. Rabinowitz, How to improve estimates of imprecision, Clin. Chem. 30 (1984) 290–292.
[28] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Internal method validation, in: B.G.M. Vandeginste, S.C. Rutan (Eds.), Handbook of Chemometrics and Qualimetrics. Part A, Elsevier, Amsterdam, 1997, p. 379.
[29] NCCLS, Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline, NCCLS Document EP5-A, 1999.
[30] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, R.D. McDowall, Validation of bioanalytical chromatographic methods, J. Pharm. Biomed. Anal. 17 (1998) 193–218.
[31] Y. Vander-Heyden, A. Nijhuis, J. Smeyers-Verbeke, B.G. Vandeginste, D.L. Massart, Guidance for robustness/ruggedness tests in method validation, J. Pharm. Biomed. Anal. 24 (2001) 723–753.
[32] S. Souverain, S. Rudaz, J.-L. Veuthey, Matrix effect in LC-ESI–MS and LC-APCI–MS with off-line and on-line extraction procedures, J. Chromatogr. A 1058 (2004) 61–66.
[33] C.R. Mallet, Z. Lu, J.R. Mazzeo, A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts, Rapid Commun. Mass Spectrom. 18 (2004) 49–58.
[34] H.R. Liang, R.L. Foltz, M. Meng, P. Bennett, Ionization enhancement in atmospheric pressure chemical ionization and suppression in electrospray ionization between target drugs and stable-isotope-labeled internal standards in quantitative liquid chromatography/tandem mass spectrometry, Rapid Commun. Mass Spectrom. 17 (2003) 2815–2821.
[35] R. Dams, M.A. Huestis, W.E. Lambert, C.M. Murphy, Matrix effect in bioanalysis of illicit drugs with LC–MS/MS: influence of ionization type, sample preparation, and biofluid, J Am. Soc. Mass Spectrom. 14 (2003) 1290–1294.
[36] C. Muller, P. Schafer, M. Stortzel, S. Vogt, W. Weinmann, Ion suppression effects in liquid chromatography-electrospray-ionisation transport-region collision induced dissociation mass spectrometry with different serum extraction methods for systematic toxicological analysis with mass spectra libraries, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 773 (2002) 47–52.
[37] R. King, R. Bonfiglio, C. Fernandez-Metzler, C. Miller-Stein, T. Olah, Mechanistic investigation of ionization suppression in electrospray ionization, J. Am. Soc. Mass Spectrom. 11 (2000) 942–950.
[38] B.L. Ackermann, M.J. Berna, A.T. Murphy, Recent advances in use of LC/MS/MS for quantitative high-throughput bioanalytical support of drug discovery, Curr. Top. Med. Chem. 2 (2002) 53–66.
[39] H. Sun, W. Naidong, Narrowing the gap between validation of bioanalytical LC–MS–MS and the analysis of incurred samples, Pharm. Technol. 27 (2003), pp. 74,76,78,80,82,84,86.
[40] W. Naidong, H. Bu, Y.L. Chen, W.Z. Shou, X. Jiang, T.D. Halls, Simultaneous development of six LC–MS–MS methods for the determination of multiple analytes in human plasma, J. Pharm. Biomed. Anal. 28 (2002) 1115–1126.
[41] F. Streit, M. Shipkova, V.W. Armstrong, M. Oellerich, Validation of a rapid and sensitive liquid chromatography-tandem mass spectrometry method for free and total mycophenolic acid, Clin. Chem. 50 (2004) 152–159.
[42] B. Fan, M.G. Bartlett, J.T. Stewart, Determination of lamivudine/stavudine/efavirenz in human serum using liquid chromatography/electrospray tandem mass spectrometry with ionization polarity switch, Biomed. Chromatogr. 16 (2002) 383–389.
[43] J. Wieling, G. Hendriks, W.J. Tamminga, J. Hempenius, C.K. Mensink, B. Oosterhuis, J.H. Jonkman, Rational experimental design for bioanalytical methods validation. Illustration using an assay method for total captopril in plasma, J. Chromatogr. A 730 (1996) 381–394.
[44] F.T. Peters, T. Kraemer, H.H. Maurer, Drug testing in blood: validated negative-ion chemical ionization gas chromatographic-mass spectrometric assay for determination of amphetamine and methamphetamine enantiomers and its application to toxicology cases, Clin. Chem. 48 (2002) 1472–1485.
[45] C. Kratzsch, A.A. Weber, F.T. Peters, T. Kraemer, H.H. Maurer, Screening, library-assisted identification and validated quantification of fifteen neuroleptics and three of their metabolites in plasma by liquid chromatography/mass spectrometry with atmospheric pressure chemical ionization, J. Mass Spectrom. 38 (2003) 283–295.
[46] F.T. Peters, S. Schaefer, R.F. Staack, T. Kraemer, H.H. Maurer, Screening for and validated quantification of amphetamines and of amphetamine- and piperazine-derived designer drugs in human blood plasma by gas chromatography/mass spectrometry, J. Mass Spectrom. 38 (2003) 659–676.
[47] C. Kratzsch, O. Tenberken, F.T. Peters, A.A. Weber, T. Kraemer, H.H. Maurer, Screening, library-assisted identification and validated quantification of twenty-three benzodiazepines, flumazenil, zaleplone, zolpidem and zopiclone in plasma by liquid chromatography/mass spectrometry with atmospheric pressure chemical ionization, J. Mass Spectrom. 39 (2004) 856–872.
[48] H.H. Maurer, O. Tenberken, C. Kratzsch, A.A. Weber, F.T. Peters, Screening for, library-assisted identification and fully validated quantification of twenty-two beta-blockers in blood plasma by liquid chromatography–mass spectrometry with atmospheric pressure chemical ionization, J. Chromatogr. A 1058 (2004) 169–181.
[49] V. Habrdova, F.T. Peters, D.S. Theobald, H.H. Maurer, Screening for and validated quantification of phenethylamine-type designer drugs and mescaline in human blood plasma by gas chromatography/mass spectrometry, J. Mass Spectrom. 40 (2005) 785–795.
[50] F.T. Peters, N. Samyn, C. Lamers, W. Riedel, T. Kraemer, G. de Boeck, H.H. Maurer, Drug testing in blood: validated negative-ion chemical ionization gas chromatographic–mass spectrometric assay for enantioselective determination of the designer drugs MDA, MDMA (Ecstasy) and MDEA and its application to samples from a controlled study with MDMA, Clin. Chem. 51 (2005) 1811–1822.
[51] O.H. Drummer, Postmortem toxicology of drugs of abuse, For. Sci. Int. 142 (2004) 101–113.
[52] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Standard addition method, in: B.G.M. Vandeginste, S.C. Rutan (Eds.), Handbook of Chemometrics and Qualimetrics. Part A, Elsevier, Amsterdam, 1997, p. 207.