Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Pharmaceutical and Biomedical Analysis
32 (2003) 753–765
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

An analysis of the SFSTP guide on validation of chromatographic bioanalytical methods: progresses and limitations

B. Boulanger [a], P. Chiap [b,*], W. Dewe [a], J. Crommen [b], Ph. Hubert [b]

[a] *Lilly Development Centre, Statistical and Mathematical Sciences, rue Granbompré, 11, B-1348 Mont-Saint-Guibert, Belgium*
[b] *Department of Analytical Pharmaceutical Chemistry, Institute of Pharmacy, University of Liège, CHU, B36, B-4000 Liège 1, Belgium*

Received 22 May 2002; received in revised form 23 September 2002; accepted 26 September 2002

Abstract

The Société Française des Sciences et Techniques Pharmaceutiques (SFSTP) published in 1997 a guide on the validation of chromatographic bio-analytical methods, which introduces new concepts in three different areas: stages of the validation, test of acceptability of a method and design of experiments to perform. In 'stages of validation', the SFSTP guide requires two phases to validate a method. The first phase, called 'prevalidation', is intended to (1) identify the model to use for the calibration curve; (2) evaluate the limits of quantitation; and (3) provide good estimates of the precision and bias of the method before designing the 'validation' phase per se. In the 'test of acceptability', the use of the interval hypotheses is envisaged by the SFSTP guide, not on the parameters of bias and precision, but on individual results by mixing mean bias and intermediate precision in a single test. The SFSTP guide also avoids the use of Satterthwaite's df for testing the acceptability. The reasons for those choices are discussed extensively. In 'design of experiments', much effort has been devoted to improving the quality of results by optimally designing and sizing the experiments to perform in validation. The rationale for using near D-optimal designs for the calibration curve is demonstrated and sample sizes are proposed to correctly size the validation experiments.

Keywords: Bioanalysis; Method validation; Statistics

1. Introduction

Before using an analytical method for quantitative determinations of drugs and their metabolites, an applicant laboratory must first demonstrate that the envisaged method fulfils a number of performance criteria. Since the publications of the 'Washington Conference' [1] and the ICH Guidelines on Validation of Analytical Methods Q2A and Q2B [2,3], which list the performance criteria to reach from a regulatory point of view, many laboratories have started to redesign their processes by involving analysts and statisticians, in order to define strategies that will allow the fulfilment of the regulatory requirements, while

* Corresponding author.
E-mail address: p.chiap@ulg.ac.be (P. Chiap).

0731-7085/03/$ - see front matter
doi:10.1016/S0731-7085(03)00182-1

being practicable and scientifically consistent. Some laboratories have probably been lucky in finding an easy way to reach the goals, most have certainly experienced, as we did, some frustrations while trying to cope with contradictory, sometimes scientifically irrelevant, requirements and definitions. As an indication of this difficulty to define reasonable practicable strategies to satisfy global regulatory requirements, laboratory constraints and scientific consistency, no guide has been published that entirely addresses that issue. For this reason, the 'Société Française des Sciences et Techniques Pharmaceutiques (SFSTP)' created in 1995 a Commission involving analysts and statisticians from the industry and the regulatory agencies with the objective of publishing a guide [4] that could be used by laboratories. The proposed guide has been validated in several real cases before being published and practical applications are now available [5,6] that provide the analyst, on the one hand, with a better understanding on the way to proceed and on the other hand, real data for qualifying his own computations that he could perform using a commercial spreadsheet.

The SFSTP guide does not constitute a final end point, but on the contrary, was envisaged as a large basis to pave the way for developments that are expected from readers and analysts that will practice the guide. On one hand, since the publication of the guide in 1997, members of the SFSTP Commission already have some modifications or warnings to propose in order to initiate a continuous process of improvements. On the other hand, many choices and decisions that have been taken in this guide constitute disruptive progresses compared to traditional ways to proceed in this area. Those choices must be clearly justified and understood because the guide is consistent as a whole and cannot be applied part by part. Finally, the SFSTP guide [4,7] does not cover all the topics or performance criteria imposed by the ICH, such as stability and robustness.

The objectives of the present article are precisely to identify and explain the progress permitted by the SFSTP guide, point out some of the limitations and suggest ways to overcome them.

2. Stages of validation

As pointed out by Smith and Sittampalam [8], the validation process involves four stages that are called by the authors 'Concept', 'Performance', 'Operational' and possibly 'Cross Validation'. Behind the new words proposed, it is of initial importance to understand that the validation is a permanent process that starts from the very beginning of the life of the method until its retirement. In the Concept or development phase, the analyst must identify and evaluate the impact of potential sources of variability that could later alter the global quality of the results. The objective today in development is no more to find a method that 'works', nor to elaborate smartly an analytical method whose quality will have to be evaluated in a later stage; the objective becomes to build results of quality by means of an analytical method. In other words, questions about the bias, precision and robustness must conduct the actions of the analyst developing a new method and no more focus its efforts only on some performance criteria, such as minimal resolution or maximal retention (migration) time in the case of chromatographic or electrophoretic methods. The ability of an analytical method to provide individual determinations of high quality, i.e. measurements close from the true content of a sample, should be the very endpoint every developer has to focus on.

The SFSTP guide unfortunately does not explicitly put a great emphasis on the development phase and might give the impression that the 'validation' is only seen as a sequence of experiments and calculations to perform to successfully reach an endpoint that is the documentation step. The SFSTP guide indirectly addresses the issue of the development since, as clearly stated, preliminary knowledge or a priori on the performance of the method must be available before properly starting the characterisation stage. This formal validation stage must be seen as a set of experiments that will confirm the regulatory agencies and the analyst himself that the method can indeed be used for its intended purpose. The validation phase can absolutely not be envisaged as a mean to estimate the performance of the method. If nothing or very little is known about the bias, the

precision, the range or the limits of quantitation before starting the validation itself, it is almost impossible or too expensive to specify the experiments to perform—i.e. selecting the levels of concentration, the number of runs, the number of replicates per run, the extreme concentration levels—while being able to give reasonable chance of success to end with a valid analytical method. Such an approach is counter-productive and unacceptable in an industrial perspective. If a developer proceeds to the validation stage with a sample size that is too small—with respect to the unknown performance, he takes the risk of increasing significantly the costs for his laboratory either by accepting as valid a truly non valid method (high non-productive cost in routine) or rejecting a truly valid method (non-productive cost in development and potential delays in a project). For this reason, the SFSTP guide warmly recommend to start with a 'Prevalidation' phase whose objectives are precisely to (1) identify the model to use for the calibration curve; (2) evaluate mainly the lower limit of quantitation; and (3) provide good estimates of the precision of the method for optimally sizing the 'validation' phase per se. Discussions could arise around the word 'prevalidation' that could be understood by some analysts as steps to perform but not to report necessarily. This 'prevalidation' must however be viewed as a real validation first phase and documented accordingly. During the 'prevalidation', the model to be used as calibration curve will be identified and the quality of fit will be assessed only at this stage. The experiments proposed are designed to consistently evaluate the adequacy of the model. In the second phase, called 'validation', the objective is to mimic the routine practice that is envisaged. The model will be used as is—the parameters will of course be estimated based on the new data—and no more investigation specific to the quality of fit will be conducted, the same way it should be carried out during routine. In this second step, the experiments are designed to focus on the estimation of the bias and precision of the method, not on the calibration curve. If the model identified in the 'prevalidation' is not adequate, then the bias and the precision in the 'validation' are impaired. The same reasoning applies to the limits of quantitation (LOQ); the range cannot be shortened after the validation results without impairing the global quality. For instance, the model identified in the previous phase has been demonstrated as adequate over the whole range and could induce bias if the range of application is changed.

As already stated, the knowledge of the bias, precision and limits of quantitation is mandatory for initiating a formal validation phase, but the proposed 'prevalidation' phase could be skipped if consistent estimates of the performance criteria are available to the analysts. That happens when methods are developed following a well-structured strategy, such as applying experimental design approaches. Unfortunately, the 'trial-and-error' approach is still widely used for developing new methods and so very little is known at the end of such a development process. In this last case, the 'validation' becomes unfortunately the very first opportunity to estimate the performance of the method.

On the other side, the validation of the method continues even after having successfully met all the requirements and documented the results. The SFSTP guide was precisely elaborated in that prospective since it provides to the analyst a controlled rate of failure during routine use. The total cost of the validation experiments proposed by the SFSTP guide could be perceived as more expensive than other classical approaches, but since a limited rate of failure is guaranteed at the end of the validation, the cost of routine use of the method can dramatically be reduced, which is more important. Unfortunately, the SFSTP commission does not make any recommendation for assessing the validity of the method after the 'validation', such as re-evaluating periodically the main performance criteria (bias, precision, calibration model, etc.) based on historical data obtained with the standard samples and the quality control samples. Besides the classical statistical process control (SPS) techniques that allow to detect rapidly and safely the occurrence of any problems, re-estimating periodically the criteria on large sets of data obtained in less controlled conditions—as opposed to well controlled conditions in validation—provides a less biased image of the quality of the method at very little cost. Regula-

tory documents, such as the ICH and the Washington Conference, impose the use of quality control samples (QCS) and clearly specify the limits of acceptance/rejection of each analytical runs separately, i.e. the 4-6-20 rule [9] that stated that at least four samples out of six should be measured within the $\pm 20\%$ acceptance limits and if two samples are observed outside those limits, they cannot occur at the same concentration level. Based on the SPC principles, those controls, if necessary, are not sufficient to guarantee the quality of a method since a non-negligible number of unacceptable runs still cannot be rejected and acceptable runs rejected. It is a matter of responsibility to continuously reassess the validity of a method after the formal validation.

As demonstrated by Kringle and Khan-Malek [10], the efficiency of the 4-6-20 rule largely depends on the way the acceptance tests are defined in the validation phase. As it will be seen later (observed versus true, parameters versus results) the SFSTP guide attempted to correctly address this issue in order to control and reduce the rate of false rejection and false acceptance if the method truly continues to behave over time as during the validation phase.

3. Observed versus true, parameters versus results

The interesting progress of the SFSTP guide is to recommend the use of the interval hypotheses approach already introduced by Hartman et al. [11] and coming directly from the bioequivalence paradigm [12] for assessing the acceptability of the bias and the precision in the same test. The use of the Interval Hypotheses tests was made possible since the establishment of limits of acceptance by the Washington Conference, i.e. the 80–120% limits at the LOQ and the 85–115% limits elsewhere for the bias and 20% (15%) for the precision.

In fact, the confused objective of the Washington Conference was to require that most, for instance 95%, new measurements made by an analytical method must fall within the 80–120% (or 85–115%) limits as suggested later in the same document by recommending the 4-6-20 rule. If the objective stated by the Washington Conference is correct and necessary, the way the requirements were formulated (i.e. bias $< 15\%$ and precision $< 15\%$) for the validation was erroneous for two main reasons.

First, if, from validation experiments, you end up with a method whose observed bias—an average over several measurements—is effectively within the 80–120% limit, let us say 119% and whose observed intermediate precision R.S.D. is $< 20\%$ acceptance limits, let us say 19%, then strictly according to the Washington Conference rule, this method can be accepted. However, in routine use, $\approx 50\%$ of the measurements will fall outside the acceptance limits and the 4-6-20 rule will fortunately reject most of such runs and less fortunately, also accept many suspect measurements. There is definitively a contradiction, or at least a confusion, between the requirements for validation and those for routine use. This was already pointed out by Hartmann et al. [12] and Kringle and Khan-Malek [10]. As stated by those authors [10], the main concern about the Washington Conference is: do the requirements of maximum 20% (15%) apply on the **observed** performance criteria (bias and precision) or on the **true** performance criteria? The difference is huge since in the first case the point estimate (e.g. the average bias: $\hat{\mu}_T$) must fall within the acceptance limit, while in the second case that is the (e.g. 95%) confidence intervals (e.g. CI in Eq. (3)) around the point estimate that must fall within the same acceptance limits. Reasonably, most will agree today that the acceptance limits apply on the **true** performance criteria as envisaged by Hartmann et al. [13]. To overcome this confusion, those authors did propose the use of the interval hypotheses approach whose general forms of the null hypotheses and alternate hypotheses can be written as follows:

$$H0:\ \theta_T \le \delta_L \ \text{ or } \ \theta_T \ge \delta_U$$

versus

$$Ha:\ \delta_L < \theta_T < \delta_U \tag{1}$$

where θ_T is the true parameter (i.e. true bias or true precision) for the test samples and δ_L, δ_U the lower and upper acceptance limits.

When the parameter of interest is the relative bias and the acceptance limits (−15%, 15%), the interval hypotheses in Eq. (1) becomes:

$$H0:\ (\mu_T - \mu)/\mu \leq -15\% \text{ or } (\mu_T - \mu)/\mu \geq 15\%$$

versus

$$Ha:\ -15\% < (\mu_T - \mu)/\mu < 15\% \quad (2)$$

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

Following the interval hypotheses approach, an analytical method will be accepted with respect to the relative bias if both null hypotheses are rejected in favour of the alternative hypotheses. As demonstrated by Schuirmann [12], testing the null hypothesis $H0$ at level $1-\alpha$ (let us say 95%) is operationally equivalent to computing the $1-2\alpha$ (i.e. the 90%) confidence intervals on the parameter of interest and accepts an analytical method at 95% level if the two-sided 90% confidence intervals are totally included within the acceptance limits.

The use of interval hypotheses does solve 'indirectly' the usual low power problem, as stated by Hartmann et al. [13]. Indeed, as can be seen in Eq. (3), the larger the sample size (n) use in validation, the smaller the confidence intervals on the average bias parameters, since the SD of the mean asymptotically converge to 0 when n tends to the infinite:

$$CI = \hat{\theta} \pm t_{(n-1,2\alpha)} \frac{\hat{\sigma}}{\sqrt{n}} \quad (3)$$

where $\hat{\theta}$ is the estimated value of the parameter, $t_{(n-1,2\alpha)}$ is the value of the student distribution with $n-1$ df at the 2α level and $\hat{\sigma}$ the estimated SD of the samples.

There are however logical concerns since with sample size (n) large enough, it is still possible to pass the validation step successfully without controlling correctly the rate of false rejection/acceptance in routine. Does the analytical method become better because more experiments are performed? Certainly not, only the variance of the mean of several measurements improves.

As long as the true bias is within the acceptance limits, let us say at 119%, making a method acceptable cannot be reduced to a problem of sample size that would provide confidence intervals on the parameter totally included within the acceptance limits. A similar reasoning could be applied to the precision estimates (but in a one-sided way instead of in a two one-sided way as for bias since there is only a maximal acceptance limit for the precision). Following that reasoning, with a sample size large enough, it could be possible—but expensive—to make a method accepted with a bias of 119% and a precision of 19% at the LOQ. But as previously noticed, in routine use, such a method could give $\approx 50\%$ of the measurement outside the acceptance limits and then be rejected by applying the 4-6-20 rule. Such a result is obviously not in line with the objectives the Washington Conference wanted to achieve.

The second reason of this contradiction between statistical achievements and objectives is to be found in the wrong equivalence that has been unconsciously established between individual results or measurements and validation criteria—bias and precision. Statistically speaking, it is abusive and incorrect to state that a method showing an acceptable bias and an acceptable precision will provide acceptable measurements. If a method provides acceptable individual results then necessarily does this method have acceptable validation criteria. But the reverse is not true: if an analytical method shows acceptable criteria—bias and precision—it does not necessarily imply that the method will provide acceptable individual results. This confusion and asymmetry in the reasoning led participants of the Washington Conference as well numerous authors after it to focus exclusively on the evaluation of the validation criteria, while the interest had to be exclusively on the future individual results. The objective of an analytical method is to provide acceptable measurements, i.e. to give measurements for each unknown sample close from their true value, not to have acceptable validation criteria. The objective of the validation is to provide users of the methods the guarantee—or probability—that each measurement on unknown samples is close enough from the true value. The evaluation of this guarantee requires estimating the parameters of bias and precision, but cannot be limited to this simple evaluation.

The most disruptive innovation of the SFSTP guide is to apply the interval hypothesis approach on the individual results and not on the parameters as erroneously carried out for validation purpose. It is then of first importance to basically understand that a method must be able to give 95% (or $1-\alpha\%$) of any new results within the acceptance limits before being considered as acceptable. For this reason, the SFSTP guide does envisage the use of the interval hypotheses test with the mean bias (relative bias or percent recovery) parameter and with the intermediate precision parameter estimate for computing the confidence interval mixing in one test both parameters. Expressed in terms of percent recovery, the interval hypotheses on individual results can then be written as follow:

$$H0:\ \mu_{new} \leq 85\%\mu \ \text{ or } \ \mu_{new} \geq 115\%\mu$$

versus

$$Ha: 85\%\mu < \mu_{new} < 115\%\mu \tag{4}$$

where μ is the true nominal value of a sample and μ_{new} is any new results estimated with the method. As indicated above, the method will be accepted if, whatever the concentration level, the classical shortest $(1-2\alpha)$ confidence intervals are included within the acceptance limits. The SFSTP guide proposes to compute by level of concentration the confidence intervals as follows:

$$CI_j = \hat{\mu}_j \pm t(0.1,\ v\ \text{df}) S_j(\text{IP}) \tag{5}$$

where $\hat{\mu}_j$ is the mean value of the results obtained at the concentration level j, $t(0.1,\ v\ \text{df})$ is the value of the student distribution with v df at the 10% level and $S_j(\text{IP})$ is the estimate of the intermediate precision SD calculated with an adequate nested model for the corresponding concentration level. The later request the use of a nested design as proposed by the SFSTP guide that imposes to have p independent runs and n replicates per run for each level of concentration. The confidence interval in Eq. (5) applies to the individual results and provides a better guarantee—but not perfect since it is assumed that the bias is known—the future individual results will be acceptable. Note that the bias and the precision are summed up in Eq. (5). This sum is also referred to as the total error. If one realises that the observed average $\hat{\mu}_j$ is in fact the true value μ of the sample plus an estimate of the average bias, then Eq. (5) implies that the greater the bias, the better should be the precision and vice-versa to make the method accepted. Their sum cannot be greater than a specific value depending on the risk (usually 5%) the user is ready to take.

The problem, and possibly one limitation of the SFSTP guide, is the number of df to use in Eq. (5). The SFSTP guide, for the sake of simplicity, suggest only the use of the within-run df : $p(n-1)$. Since the intermediate precision variance is equal to the sum of the two components (see Eq. (6)), there exist no exact distribution for such an estimate and therefore there exist no exact df [14].

$$S^2_{IP} = S^2_{\text{within-run}} + S^2_{\text{between-run}} \tag{6}$$

Several authors did propose ways to approximate the distribution of such a sum of variance components and the most widely used approximation is the one of Satterthwaite [15] who provides a way to compute the df in this case. Satterthwaite's df in our special case of interest with p runs and n replicates per run can be written as follows:

$$\text{Sat DF} = \frac{p(p-1)(n-1+F)^2}{(p-1)(n-1)+pF^2} \tag{7}$$

where

$$F = \frac{MS_{\text{between-run}}}{MS_{\text{within-run}}} \tag{8}$$

As can be seen from Eq. (7), for a selected design of experiment (p runs, n replicates per run), the Satterthwaite's df will only depend on the observed ratio F. Fig. 1 represents the values of Satterthwaite's df for different values of the F ratio for a design of experiment with four runs and four replicates per run. As can be seen from Fig. 1 and Eq. (7), the Satterthwaite's df are close from $p*n-1$ (the df of the total error) when the F ratio is close to 1 and then decreases continuously to reach asymptotically $p-1$ (the df of the between-run term) when F goes to infinite.

The question now is to examine what is lost by working with fixed $p(n-1)$ df, as suggested by SFSTP compared to the use of Satterthwaite's df, according to the values of the F ratio one can

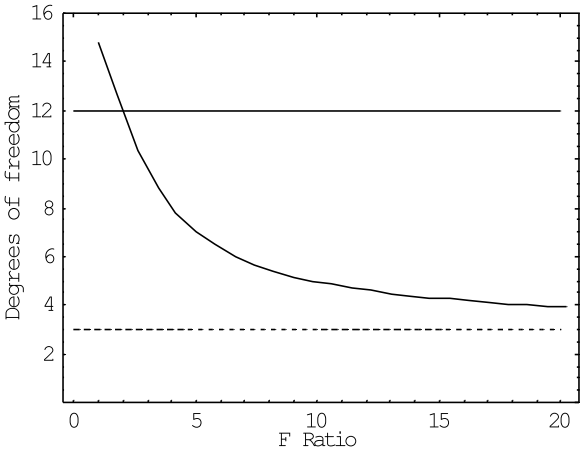

Fig. 1. Number of Satterthwaite's df as a function of the ratio F = MS between-run/MS within-run for a four run–four replicate nested designs. Horizontal line at 12 represents SFSTP recommendation for df for such a design. The dashed horizontal line at 3 represents the asymptotic value for Satterthwaite's df—in fact the $p-1$ df of the between-run variance—when ratio F goes to infinity.

expect to observe with chromatographic analytical methods. If the values of F are small, then the SFSTP approach is too liberal in the sense that it rejects more frequently a valid method than by using the Satterthwaite's df. For large F ratios, the SFSTP approach is however too conservative. But is that difference important?

Everyone who works with chromatographic bioanalytical methods, which is the purpose of the SFSTP guide, already pointed out that both the between-run and within-run variances are small (sometimes the between-run variance estimate is set to 0) around the lower LOQ and become larger with the concentration level. Two examples are based on real data obtained with two HPLC analytical methods (range: 25–1000 μg/ml) and are represented in Fig. 2 where the variance components have been modelled as functions of the concentration level.

As can be seen in Fig. 2(a, b), the two variance components effectively increase with the concentration level but they can relatively increase at different rates making the F ratio evolve in very different ways with the concentration levels (see Fig. 2c, d). Fig. 2(e, f) displays the corresponding df from Satterthwaite and are compared to those suggested by the SFSTP guide indicating a significant difference when the between-run variance component grow faster than the within-run variance component with the concentration. The difference between the two approaches is less important when both variances increase at comparable rates with the concentration. Whatever the difference observed with respect to the df, the most important impact has to be seen on the accuracy profile or on the function of the upper limit of confidence interval computed with Eq. (5) and assuming no bias. Fig. 2(g, h) displays this upper limit expressed in percentage of the concentration. As can be seen, the difference between both approaches is minimal and suggest that when between-run variance component increases faster than the within-run one, the lower LOQ is slightly overestimated using the SFSTP recommendation. Above the lower LOQ, the way to compute the df is without impact for accepting or not an analytical method. Stated differently the equation used for calculating the df only affect the estimate of the lower LOQ of a chromatographic analytical method, the Satterthwaite's method being preferred. As suggested by Fig. 2(c, d), in general the between-run variance at the lower LOQ is very small compared to the within-run variance and is frequently equal to 0. In fact, as demonstrated by Searle et al. [14], the probability of having negative estimates for the between-run variance increase dramatically once the ratio of the true between-run variance over the true within-run variance is smaller than 1 and becomes important when the ratio is smaller than 0.25. Having such small ratios is usually the rule at the lower LOQ. Then, at the LOQ, the most important components and probably the only component, is the within-run variance. For those practical reasons and for the sake of simplicity, the SFSTP guide recommends the use of the within-run variance df $p(n-1)$.

4. Design of experiments

Another major—but less apparent—progress of the SFSTP guide is the large part devoted to the design of experiments and the underlying use of theory of optimal design [16]. Except for few

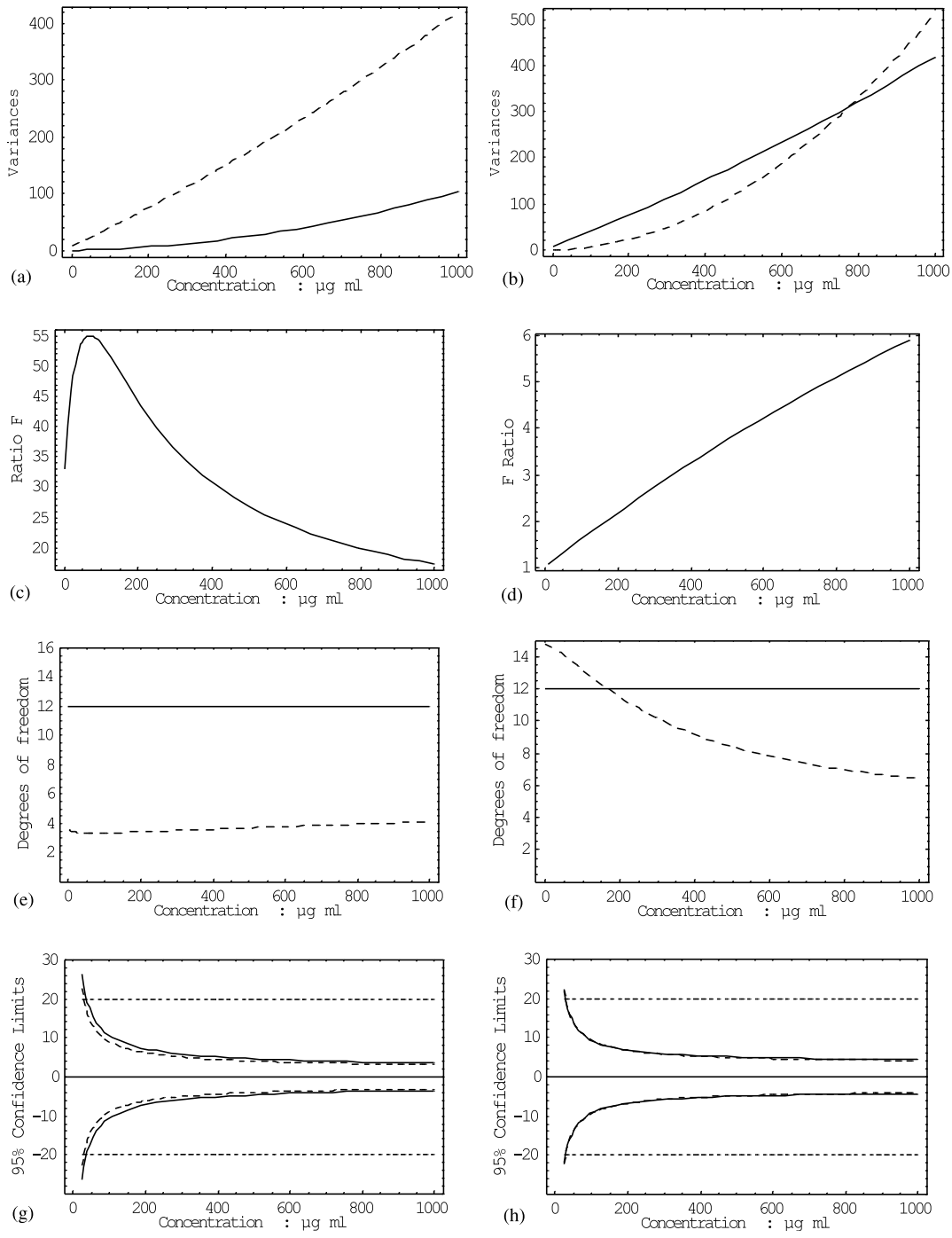

Fig. 2. Classical estimated values obtained with chromatographic bio-analytical methods of the between-run variance (a,b: dashed line) and the within-run variance (a,b: solid line) as a function of the concentration level. In (c, d) the corresponding F ratio as a function of the concentration levels. In (e, f) the df with Satterthwaite's approximation (dashed line) or according to the SFSTP guide (solid line). In (g, h) the precision profiles (expressed in CV) obtained using Satterthwaite's approximation for df (dashed line) or SFSTP df (solid line).

papers [17], very little is said in the literature about the dramatic gains in quality and cost that could be obtained by optimally designing experiments. Most analysts continue to think of experimental design in terms of sample size. This remains of course an important issue, but arises only after that the optimality of the design has been investigated, for example, selection of optimal concentration levels. In the opposite direction, minimal methodological recommendations made by the ICH in Q2B [2] or by the European Note EEC III/844/87 [18] regarding the experiments to perform are sub-optimal and so contradictory with their overall objective of quality measurements.

4.1. Design for the calibration curve

One simple topic where quality—bias and precision—of the results can be improved without additional cost is certainly the selection of the optimal concentration levels to be used for fitting the calibration curve. To figure out the impact of the selection of concentration levels on the precision, let us start with a basic theoretical example. Let us assume that the purpose is to daily calibrate a method that is truly linear in the original scale (e.g. from 10 to 1000 μg/ml) and that there are data enough to convince the analysts that this is effectively the case. The variance of the response using this method is homogenous across the whole range and the %R.S.D. at the 10-μg/ml concentration level—the assumed LOQ—is $\approx 7\%$. The bias of the method is assumed to be perfect across the range.

A first design that could be envisaged for establishing the calibration curve of such a method is the one recommended by the ICH [2] and the EEC Note [18] with at least five concentrations levels, for example six levels more or less equally spaced without repetition and covering the whole range (Design $1 = \{10, 200, 400, 600, 800, 1000\}$). Another design, also with six samples, consists of three different concentration levels—at the two extremes of the range and at the mid range—with duplicates (Design $2 = \{10, 10, 500, 500, 1000, 1000\}$). The third design, still with six samples, only has two different concentrations at the extremes but measured in triplicates (Design $3 = \{10, 10, 10, 1000, 1000, 1000\}$). Fig. 3 represents around the lower LOQ an approximation of the upper and lower confidence intervals [19] of the concentrations predicted using the calibration curve obtained with the three designs.

The two outside curves represent the upper and lower confidence intervals obtained with Design 1, while the two internal curves have been obtained with Design 3. Between those two extremes are the upper and lower limits when Design 2 is used for the calibration curve. As can be seen, the precision of the results is better with Design 3 than with Design 1, especially around the lower LOQ that is of first importance when supporting a pharmacokinetic study. Above 100 μg/ml, the difference in precision between the three designs is irrelevant. Moreover, if as suggested by the Washington Conference, the LOQ is the lowest concentration where 95% of the results fall within the (80–120%) limits, then the LOQ obtained when calibrating with Design 1 is 12 μg/ml instead of 10 μg/ml for Design 3, i.e. an improvement of 20% in this simple example. Stated differently, the selection of the concentration levels has an impact on the precision and the LOQ of the method and the gain is obtained here without additional cost.

In a statistical perspective, such a result is not a surprise since, in general, the smaller the variance of the estimated parameters—the slope and the

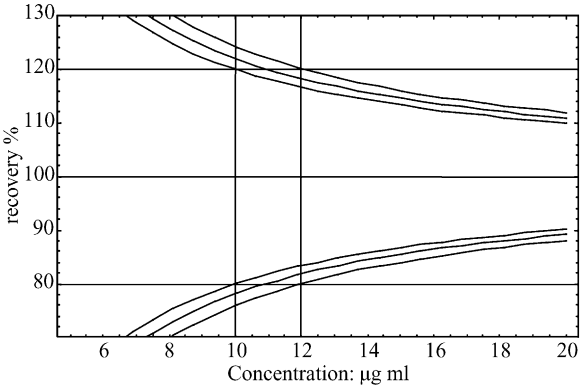

Fig. 3. The upper and lower 95% confidence intervals on a new prediction using linear calibration lines obtained with three different designs as a function of the concentration. Confidence intervals are expressed in the percent recovery scale.

intercept in the case of the simple linear model—the better the precision of the inverse prediction if everything else remains equal, i.e. the analytical error, the mean value of the parameters and the number of samples. As can be seen in Eq. (9), the approximation of the variance of a new inverse prediction is a function of the analytical error $\hat{\sigma}^2$, the number of calibration samples (n), the distance between this new prediction and the mean of concentration values of calibration samples ($\hat{X}_{\text{new}} - \bar{X}^2$), the estimated slope $\hat{b}$ and the estimated variance of the slope $\text{Var}\{\hat{b}\}$:

$$\text{Var}\{\hat{X}_{\text{new}}\} = \frac{1}{\hat{b}^2}\left[\hat{\sigma}^2 + \frac{\hat{\sigma}^2}{n} + (\hat{X}_{\text{new}} - \bar{X})^2 \text{Var}\{\hat{b}\}\right] \quad (9)$$

Since the analytical error $\hat{\sigma}^2$ and the slope $\hat{b}$ are two parts that cannot be modified in validation, the sample size n and the variance of the slope $\text{Var}\{\hat{b}\}$ are the two remaining elements on which it is still possible to act to improve the precision of measurements. For a fixed sample size ($n = 6$) as envisaged in the above example, it can be seen in Eq. (9) that significant improvements can be obtained by decreasing the estimated variance of the slope. As shown by Atkinson and Donev [16], the design that minimises the overall variance of the parameters is the D-optimal design. The D-optimal design is trivial to find in the case of the simple linear model since, as indicated in Eq. (10), it is obtained when the distance between the concentration values is maximal over the range envisaged. This is the case for $n = 6$ with Design 3.

$$\text{Var}\{\hat{b}\} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (10)$$

A contrario, the design suggested by the ICH and the EEC note is sub-optimal with respect to the precision of the results. However, the intended objective for proposing such a design was to force the analyst to demonstrate that the relationship between the concentration and the response is effectively linear. Indeed, the design proposed above is D-optimal if, as indicated, the relationship is known as being truly linear. If it is not the case, for example if the relationship can be modelled using a quadratic polynomial, then the D-optimal design becomes the Design 2 for $n = 6$. Intuitively, this design makes sense since it is at the middle of the range that the difference between the linear and the quadratic model is the greatest.

The problem in finding the optimal design that will maximise the accuracy of the results is first, to identify what is the most 'likely model' that could be used as calibration curve, second, to identify the most 'likely departure' to the previous model that could occur in routine use and third, to find a 'robust optimal' design related to this model and that could detect any departure to the main model while maximising the precision. A fourth problem that is added is the necessity to also find the 'lowest LOQ' that is impacted by the preliminary design used for identifying the model. Finally, be aware that for most bio-analytical methods with large range, the variance of the response will probably not be homogeneous across the range and weighted estimation strategies will have to be envisaged.

The SFSTP guide proposes an iterative strategy for solving those four problems during the pre-validation phase that could be summarised in the form of a 'rule-of-thumb':

1) Make some guesses on what could happen and identify the families of models that could be possible, such as linear model, log-linear model or quadratic polynomial model, for example. Not every model is likely depending on the method used and in the case of chromatographic methods, sigmoid models are, for example, not expected.
2) Make some guesses about the possible range for the method and the LOQ.
3) Select a design that is optimal with respect to the most complicated model envisaged in 1 and covering the range defined in 2.
4) To the optimal design selected in 3, add concentration levels below and above the a priori LOQ anticipated in 2 in order to locally figure out how the precision behaves. Those concentration levels close to the LOQ will act as a big point at the lowest (highest) extreme

of the range without impairing the overall optimality of the design.

5) Repeat at least three times each measurement and repeat this process over a minimum of three independent runs (or series), i.e. adjust sample size only after having identified an optimal design.
6) Find the best model within the family envisaged in 1 that fit the data while keeping all the concentration levels at this step.
7) Estimate the precision of the results using this model and eventually eliminate the lowest concentration level if the precision obtained at that level is too low.
8) Continue in 6 with the remaining concentration levels until that precision is acceptable across the updated range (7).
9) Consider the model found as the true model and do not re-evaluate its functional form in the validation phase.

This process suggested in the SFSTP guide, that could appear quite cumbersome, constitutes a good compromise between statistics and cost-effectiveness and provides results close to the ideal solution.

4.2. Design for estimating the variance components

Another type of design that is critical in the validation of analytical methods is the one used for estimating the different variance components, i.e. the within-run variance (repeatability variance) and the between-run variance and potentially the inter-laboratory variance; the intermediate precision variance being equal to the sum of the repeatability variance and the between-run variance (see Eq. (2)), the reproducibility variance being equal to the intermediate precision variance plus the inter-laboratory variance. Until Kringle and Khan-Malek [10] very little has been proposed in the literature that are in accordance with the new recommendations from the Washington Conference and the ICH. At least in the ICH it is indicated that 'the use of an experimental design (matrix) is encouraged'. From a statistical stand point, the use of a correctly sized experimental design is mandatory if the objective is to reduce the consumer risk as well as the producer risk. The SFSTP guide precisely does propose estimating the intermediate precision a way to optimally size a design for the validation phase that is reasonable, cost-effective while statistically meaningful and depending on the expected precision and bias of the method. The guide does not cover the estimation of the 'reproducibility' variance. The expected precision and bias are directly available at the end of the prevalidation phase, confirming, by the way, the important role that this preliminary phase plays in adequately powering the validation phase. The proposed minimal number of runs and replicates to perform within each run are reported in Table 1. They have been obtained by simulation assuming a maximal relative bias of 2% and using the interval hypotheses approach described earlier. Note that the proposed sample sizes indicated in Table 1 are sample sizes by level of concentration and apply only to the quality control samples (QCS) that must be prepared independently of the calibration samples. At least three concentration levels are requested for the QCS and the intermediate precision has to be estimated by concentration level. Ideally, the levels to be selected must be at the two extremes of the range (0 and 100%) and around the mid-range (50%). The SFSTP guide proposes four levels: at the lowest LOQ, three times the LOQ, mid-range (50%) and at 80% of the range. The choice of 80 instead of 100% is a matter of documentation and cannot be statistically justified. The argument for having QCS at 80% is that, according to regulatory documents, no extrapolation is allowed outside the range. Having QCS at 100% will of course provide $\approx$ 50% of the measurements above the maximal determination limit and this could be seen as an extrapolation. The argument arises probably from a wrong interpretation of the regulatory documents. Extrapolation has to be avoided for unknown samples, but remain acceptable and even recommended, when it applies to a sample whose nominal value is known and within the range. With QCS at 80% of the range, it is impossible to guarantee that at a concentration close to the upper end of the range, the precision and bias are still acceptable. Recommending QCS at 100% of the range during the validation phase

Table 1
Recommended numbers of runs and replicates per run as a function of the relative between-run and within-run SD estimated in prevalidation

Relative between-run SD	No. runs	Relative within-run SD				
		4%	5%	6%	7%	8%
		No. replicates per run				
4%	3	4	4	5	6	–
	4	4	4	4	5	9
	5	4	4	4	5	5
5%	3	4	4	4	5	–
	4	4	4	4	6	–
	5	4	4	4	5	8
	6	4	4	4	4	5
6%	3	4	4	6	10	–
	4	4	4	6	7	–
	5	4	4	5	7	–
	6	4	4	5	5	6
7%	3	6	8	–	–	–
	4	4	4	6	–	–
	5	4	4	5	7	–
	6	4	4	5	7	9
8%	4	9	–	–	–	–
	5	6	8	–	–	–
	6	4	5	8	–	–

and during the routine use is certainly an improvement to the original SFSTP guide.

The rationale for having QCS at three times the LOQ is different: most analysts do suspect that, using the 4-6-20 rule, the rate of rejection in routine use will be too high at the LOQ and having QCS located slightly above the LOQ is perceived as a 'back-up' that will allow the restriction of the range occasionally in routine without loosing measurements falling above this limit. As seen earlier, the LOQ is defined as the lowest concentration where 95% of the future measurements will fall within the acceptance limits (80–120%). If correctly estimated, only $\approx 5\%$ of the QCS will fall outside the acceptance and this 'back-up' level is certainly of little use.

The SFSTP guide also strongly recommends that the different runs must be performed in conditions as different as possible and reflecting the way the method will potentially be used in routine. It is indicated that, if envisaged, different apparatus, operators or any other environmental source of variations must cover from run to run. It is a matter of protecting the analyst himself against a too high rate of rejection of runs in routine. Validating the method in very uniform conditions will increase the chance to pass successfully the documentation or characterisation step, but routinely the cost of use of the method could increase in an uncontrolled manner. This is certainly not the objective for an analyst developing and validating a method.

References

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

[2] Food and Drug Administration, International Conference on Harmonization, Fed. Regist. 60 (1995) 11260–11262.
[3] Food and Drug Administration, International Conference on Harmonization, Fed. Regist. 62 (1997) 27463–27467.
[4] E. Chapuzet, N. Mercier, S. Bervoas-Martin, B. Boulanger, P. Chevalier, P. Chiap, D. Grandjean, Ph. Hubert, P. Lagorce, M. Lallier, M.C. Laparra, M. Laurentie, J.C. Nivet, S.T.P. Pharm. Pratiq. 7 (1997) 169–194.
[5] E. Chapuzet, N. Mercier, S. Bervoas-Martin, B. Boulanger, P. Chevalier, P. Chiap, D. Grandjean, Ph. Hubert, P. Lagorce, M. Lallier, M.C. Laparra, M. Laurentie, J.C. Nivet, S.T.P. Pharm. Pratiq. 8 (1998) 81–107.
[6] P. Chiap, Ph. Hubert, B. Boulanger, J. Crommen, Anal. Chim. Acta 391 (1999) 227–238.
[7] P. Hubert, Ph. Chiap, J. Crommen, B. Boulanger, E. Chapuzet, N. Mercier, S. Bervoas-Martin, P. Chevalier, D. Grandjean, P. Lagorce, M. Lallier, M.C. Laparra, M. Laurentie, J.C. Nivet, Anal. Chim. Acta 391 (1999) 135–148.
[8] W.C. Smith, G.S. Sittampalam, J. Biopharm. Statistics 8 (1998) 509–532.
[9] R.O. Kringle, Pharm. Res. 11 (1994) 556–560.
[10] R. Kringle, R. Khan-Malek, Statistical intervals for nested random samples in assay validation, Midwest Biopharmaceutical Statistics Workshop, Muncie, IN, May 24, 1994.
[11] C. Hartmann, J. Smeyers-Verbeke, W. Pennick, XX, Y. Van der Heyden, P. Vankeerberghen, D.L. Massart, Anal. Chem. 67 (1995) 4491–4499.
[12] D.J. Schuirmann, J. Pharmacokin. Biopharm. 15 (1987) 657–680.
[13] C. Hartmann, D.L. Massart, R.D. McDowall, J. Pharm. Biomed. Anal. 12 (1994) 1337–1343.
[14] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, Wiley, New York, 1992.
[15] F.E. Satterthwaite, Biometr. Bull. 2 (1946) 110–114.
[16] A.C. Atkinson, A.N. Donev, Optimum Experimental Design, Oxford University Press, Oxford, 1992.
[17] D.M. Rocke, G. Jones, Technometrics 39 (1997) 162–170.
[18] Explanatory Note EEC III/844/87 EN-Final, August 1989.
[19] S.-C. Chow, J.-P. Liu, Statistical design and analysis in pharmaceutical sciences: validation, process controls and stability, in: Statistics Textbooks and Monographs, vol. 143, Marcel Dekker, New York, 1995.