Journal of Chromatography B, 877 (2009) 2235–2243

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

A risk-based analysis of the AAPS conference report on quantitative bioanalytical methods validation and implementation☆

Bruno Boulanger[a], Eric Rozet[b,1], François Moonen[c], Serge Rudaz[d], Philippe Hubert[b,*]

[a] UCB Pharma, Exploratory Statistics, Braine L'Alleud, Belgium
[b] Laboratory of Analytical Chemistry, CIRM, Department of Pharmacy, University of Liège, B36, B-4000, Liège, Belgium
[c] Arlenda S.A., CHU, B36, B-4000, Liège, Belgium
[d] School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, 1211 Geneva, 4, Switzerland

ARTICLE INFO

Article history:
Received 28 January 2009
Accepted 12 June 2009
Available online 18 June 2009

Keywords:
Validation
Tolerance interval
Decision rule
Accuracy
Total error
Acceptance criteria
Risk
Fit-for-purpose

ABSTRACT

The 3rd American Association of Pharmaceutical Scientists (AAPS)/Food and Drug Administration (FDA) Bioanalytical workshop in 2006 concluded with several new recommendations regarding the validation of bioanalytical methods in a report published in 2007. It was aimed to conciliate or adapt validation principles for small and large molecules and an opportunity to revisit some of the major decision rules related to acceptance criteria given the experience accumulated since 1990. The purpose here is to provide a "risk-based" reading of the recommendations of 3rd AAPS/FDA Bioanalytical Workshop. Five decision rules were compared using simulations: the proposed pre-study FDA and Total Error Rules, the rules based on the β-Expectation Tolerance and β-γ-Content Tolerance Interval and, finally, the 4-6-20 rule for in-study acceptance of runs. The simulation results demonstrated that the β-Expectation Tolerance Rule controls appropriately the risk. The β-γ-Content Tolerance Interval was found to be too conservative, depending on the objective, and to lead to a high rate of rejection of procedures that could be considered as acceptable. On the other side, the FDA and the AAPS/FDA workshop Total Error Rule, combined or not, did not achieve their intended objective. With these rules, the risk is high to deliver results in study that would not meet the targeted acceptance criteria. This can be explained because, first, there is confusion between the quality of a procedure and the fitness of purpose of the results it could produce and, second, between the initial performances of a procedure, for example evaluated during pre-study validation and the quality of the future results.

1. Introduction

The American Association of Pharmaceutical Scientists (AAPS) and the Food and Drug Administration (FDA) Bioanalytical Workshops organized in 1990 and 2000 workshops that permit significant improvements and progresses in the practices for the validation of bioanalytical methods. The 3rd AAPS/FDA Bioanalytical Workshop, held on May 1–3, 2006, in Arlington, VA, concluded with several new recommendations published in 2007 [1] that do not override the current FDA guidance [2]. The need of a 3rd Conference was driven by two major elements: first the confirmation of new technologies in analytical sciences and second, the divergent evolution of methodologies and technologies for conventional low molecular weight analytes and macromolecules, the later becoming more and more important in the development of new drug substances and products.

This conference was an opportunity to understand the major differences that exist between the quantitative determination of low and large molecular weights molecules from a practical or laboratory point of view while identifying and recognizing the need to define harmonized approaches to demonstrate the validity of quantitative bioanalytical methods. Both types of procedures, chromatographic and ligand-binding assays (LBA), are mainly used in the development and commercialization of drugs for quantifying the active substance, their degradation products, metabolites or the excipients. The assessment of quality of a result should be independent of the technology used to obtain it and from the size of the molecule to quantify. This was a major challenge, since in this type of debate objectives and practices to reach the objectives are usually embedded and confounded. Stated differently, the standard of quality to achieve is more driven by what the technology permits rather than by the needs. This attempt to achieve harmonization of

☆ This paper is part of a special issue entitled "Method Validation, Comparison and Transfer", guest edited by Serge Rudaz and Philippe Hubert.

* Corresponding author at: Laboratory of Analytical Chemistry, Institute of Pharmacy, University of Liège, CHU, B36, B-4000 Liège 1, Belgium.
Fax: +32 4 366 43 17.
E-mail address: Ph.hubert@ulg.ac.be (P. Hubert).
[1] F.R.S.-FNRS Postdoctoral Researcher (Belgium).

1570-0232/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2009.06.019

approaches reflects clearly a need, and proposals have already been made by other organizations such as International Conference on Harmonization (ICH) in 2005 [3] or, with extensive details and practical proposals, by the Société Française des Sciences et Techniques Pharmaceutiques (SFSTP) in 2003 [4], 2004 [5] and 2006 [6,7].

The 3rd Workshop was also an opportunity to revisit some of the major statistical methodologies already proposed earlier by numerous authors [8–12]. This was a legitimate objective given the technological progresses observed in computers and software since 1990. Today all analysts can access to easy-to-use or automated statistical software and therefore, there is no more reason to maintain statistical recommendations that can be handled manually but that do not accomplish their intended objective: giving confidence that each measurement can be fully trusted. Since 1990, many publications have proposed appropriate approaches to answer in statistically sound manner to the various questions and issues encountered in analytical sciences. A recent review of nearly 20 years of recommendations on analytical methods validation can be found in Rozet et al. [13].

The aim of the present study was to give a "risk-based" perspective on the proceedings of 3rd AAPS/FDA Bioanalytical Workshop by comparing the proposals with the main statistical recommendations and methodologies published since Wilks' paper in 1942 [14]. An excellent review presenting the state-of-art in the field of "Statistical method in biological assay" was already published in a book by Finney in 1978 [15]. Outside the regulatory world of pharmaceutical industry, the ISO organization already published statistical recommendations and methodologies dedicated to the analysis of validation studies [16–22].

The statistical interpretation of the results obtained and the final decision about the validity or invalidity of an analytical procedure and the associated risk should depend on consistent and adequate definition of the criteria assessed. This leads to highly critical consequences since the validated analytical method will be daily used in routine analysis (including batch release, stability assessment, establishment of shelf life, pharmacokinetic or bioequivalence studies) for making decision about the utmost business and public health consequences.

The focus of the 3rd AAPS/FDA Bioanalytical Workshop, as reflected in the report, was exclusively on the procedures themselves, the technologies, their condition of use, the preparation of samples, the suitability tests to perform, the calibration standards, as well as on the performance of the procedures. Giving limits or guidelines for the good practice of analytical procedures is necessary and valuable, but may not be sufficient. Not including from the beginning the customers in the debate and in the definition of criteria is not an optimal quality practice. All quality systems must be built by bringing together producers and customers for finding agreements to ensure that the technology developed will deliver what customers need. When dealing with analytical procedures, the customer need is to obtain high quality results that can be trusted to make critical decisions.

The consequences of disconnecting the producers and customers of the results in the writing of the guidelines may be of two types: first, the focus is on the performance of the procedures instead on the quality of the results, the very objective of a procedure. Second, the limits of acceptance are defined based on current technological capabilities instead of being based on the meaning, the use and the consequences of the results themselves. The first consequence impacts the risk of poor quality results, while the latter make potentially procedures and therefore results not fit-for-purpose. In the "risk-based" perspective analysis proposed here we will only discuss the risk of poor quality results assuming that the limits of acceptance have been appropriately determined in accordance with the use of the results.

2. Quality of future results instead of performance of procedures

What matters, in fine, are the results generated by a procedure, not the procedure *intrinsically*. This will seem obvious to most practitioners and all official documents always start by general comments such as "*The quality of these studies* [bioequivalence, pharmacokinetics (PK), and toxicokinetics], *which are often used to support regulatory filings*, *is directly related to the quality of the underlying bioanalytical data*" [2]. The ICH Q2R1 [3] starts by stating that "*The objective of validation of an analytical procedure is to demonstrate that it is suitable for its intended purpose*." The very purpose and objective of an analytical procedure is to ensure that each of the unknown quantity that the laboratory will have to quantify is accurately determined. Note that the accuracy of a measurement is taken here in its original sense and should be differentiated from the bias of the procedure [13]. For example, when a pharmacokineticist analyses the results of a study, what matter is the accuracy of each concentration level used to fit the pharmacokinetic model. The fact that the results come from an LC–MS/MS or an ELISA method is not taken into account in pharmacokinetic analyses or models and is basically ignored by the users of the results; this remain in the area of expertise of the analysts that should apply best practices. In addition, the bias and the precision of the procedure are also not taken into account into the analyses of the results; those performance criteria again remain in the area of the analyst even if performances have an impact on the accuracy of each future individual result. What matters the most to users of the results is to know how far the result is from the unknown true concentration of analyte in the sample, i.e., the results accuracy. Therefore the objective of validation of a bioanalytical procedure is to give to the user of the results, as well as to the regulatory bodies, guarantees that every single measure that will be performed in routine will – likely – be close enough to the unknown "true value" of the sample, i.e., that the measure will be accurate enough given its intended use. Everyone will agree with such objectives for the pre-study validation phase of a bioanalytical method. As reminded in the ICH Q2R1 [3]: "*The objective of the analytical procedure should be clearly understood since this will govern the validation characteristics which need to be evaluated*." The difficulty is to demonstrate the objective is reached. From a statistical perspective there are two issues. The first one is the confounding between the objective and the performances of the procedure that produce the results. The underlining assumption almost unanimously and unconsciously made is that "good" methods provide "good" results. Indeed, procedures with good performances, such as an acceptable bias and precision, do not necessarily provide "good" results. However, good results are necessarily generated by procedure with acceptable performances. The second issue is that decisions are based on past performances, not on the prediction of the quality of future results. At the outcome of method validation, what should be known is if during its future routine application, the procedure will generate reliable results.

3. Decision rules and control of risks

The conference proceeding [1] insists about the performance criteria that need to be examined, estimated and reported. The decision rule regarding the ligand-binding assays (LBA) for the pre-study validation part states: "*For a method to be considered acceptable, it is recommended that both the interbatch imprecision (%CV) and the accuracy, expressed as absolute mean bias (%RE) be ±20% (25% at LLOQ and ULOQ). As an additional constraint to control method error, it is recommended that the target total error (sum of the absolute value of the %RE [accuracy] and precision [%CV] be less than ±30% [±40% at the LLOQ and ULOQ]*)." Later, when referring to

in-study validation for ligand-binding assays, the same document indicates: "*The recommended standard curve acceptance criteria for macromolecule LBAs are that at least 75% of the standard points should be within 20% of the nominal concentration (%RE of the back-calculated values*)" while regarding the QCs, "*At least 4 of the 6 QCs must be within 20% of the nominal value*." For chromatographic procedures the rule remains the same as stated in FDA 2001 [2]:"*Acceptance criteria: At least 67% (4 out of 6) of QC samples should be within 15% of their respective nominal value, 33% of the QC samples (not all replicates at the same concentration) may be outside 15% of nominal value*". The spirit from those proposed decision rules, as written, is that it is expected that each result coming from unknown sample is likely to fall within the acceptance limits. This is indirectly assessed in routine, on a run-by-run basis, by the observed proportion of QC or standard – at least 67% or 75% – falling within the limits. Note that the objective is the likelihood each result from unknown sample is within the limit. The proportion of QC samples is not the objective. QC samples are only used as a surrogate to ensure conditions of use of the analytical procedure were acceptable for the run. If the QC samples are acceptable, then the original estimate of this likelihood remains applicable for the unknown samples. The proportion of QC samples within acceptance limits during a run is a poor estimate of the quality of the results of that run, particularly when a limited number of QC samples are used. It only indicates that the procedure performances are as expected during the run. The assumption made here is that if the bias (expressed in %RE) and the precision (expressed in %CV) are both better than 20% (15% for chromatographic procedures) during the pre-study validation, then results in the future are likely to fall with the same [−20%, 20%] ([−15%, 15%] respectively) acceptance limits. However the criteria on individual results cannot be the same as the criteria on a mean of many results.

More formally, the objective of a quantitative analytical procedure is to be able to quantify as accurately as possible each of the future unknown quantities that the laboratory will have to determine. In other terms, what all analysts expect from an analytical procedure is that the difference between the measurement or observation (X) and the unknown "true value" μ_T of the test sample be small or inferior to an acceptance limit λ a priori defined (Eq. (1)):

$$-\lambda < X - \mu_T < \lambda \Leftrightarrow |X - \mu_T| < \lambda \quad (1)$$

The acceptance limit λ can be different depending on the use of the results, it is the fit-for-purpose principle. The proposed λ by regulatory documents, so far is 20% for LBA and 15% for chromatographic procedures. The adequacy of those proposed limits will not be discussed here.

Therefore, the aim of the validation phase is to generate enough information to have guarantees that the analytical method will provide, in future routine analysis, results that will likely be close to the true value without being affected by other elements present in the sample, assuming everything else remain reasonably similar, e.g. the conditions of use of the procedure have not been dramatically changed as compared to pre-study phase. This is assessed by the QC samples.

As already mentioned [5,11,13], the difference between a measurement X and its true value μ_T is composed of a systematic error (bias or trueness as opposed to accuracy) and a random error (variance or precision). The true values of these performance parameters are themselves unknown but can be estimated based on the (pre-study) validation experiments and the reliability of these estimates depends on the adequacy of these experiments (design, size). It is important to underline that the performances are just estimated, with uncertainty, and therefore these estimates are not the true performances at all.

This implies that the objective of the pre-study validation is, first, to provide estimates of the performance of the analytical procedure and, second, using that information to predict the probability of each future result during the in-study phase to be accurate, i.e., to fall within the acceptance limits. If the predictive probability of accurate future results is high—say greater than 0.75, the procedure will then be used in-study for routine determination of unknown samples. Formally it means that, given the estimates of bias δ_M and precision (standard deviation) σ_M of the procedure, the objective of the validation phase is to evaluate whether the predictive probability π of future results that will fall within the acceptance limits is greater than a predefined minimal proportion, say π_{min}. This can be expressed mathematically as follows (Eq. (2)):

$$\hat{\pi} = E_{\hat{\delta}_M, \hat{\sigma}_M}\{P[|X - \mu_T| < \lambda] | \hat{\delta}_M,\ \hat{\sigma}_M\} \geq \pi_{min} \quad (2)$$

As indicated in the conference report or in the FDA 2001 guide [2], suggested values for π_{min} are 0.667 or 0.75: "At least 67% (4 out of 6) of QC samples should be within 15% of their respective nominal value" or ". . ..at least 75% of the standard points should be within 20% of the nominal concentration. . .".

A practically simple and old solution already proposed by other authors [5,12,13,23,24] to deal with the objective as formulated in Eq. (2), is to compute the Tolerance Intervals [25,26] that should be totally included within the acceptance limits [−20%, 20%] ([−15%, 15%] respectively). The Tolerance Interval will compute where, likely, in the future the results will lie. This is easy to interpret and to represent. The use of the β-Expectation Tolerance Interval has been shown as equivalent to the Uncertainty criteria in the ISO and related regulations [27] and equivalent to the Bayesian predictive interval for a single observation [28,29,30].

Some authors [12] recommend the β-γ-Content Tolerance Interval (see [12,24] for details), which serves another objective as the one proposed here. The subtle differences of interpretation will be developed hereafter.

The β-Expectation Tolerance Interval is defined as follows [25] (Eq. (3)):

$$E_{\hat{\mu}_M, \hat{\sigma}_M}\{P_X(\hat{\mu}_M - k_E\hat{\sigma}_M < X < \hat{\mu}_M + k_E\hat{\sigma}_M | \hat{\mu}_M,\ \hat{\sigma}_M)\} = \beta \quad (3)$$

while the β-γ-Content Tolerance Interval, also called the β-Content γ-Confidence Tolerance Interval, is defined by Eq. (4):

$$P_{\hat{\mu}_M, \hat{\sigma}_M}\{P_X(\hat{\mu}_M - k_C\hat{\sigma}_M < X < \hat{\mu}_M + k_C\hat{\sigma}_M | \hat{\mu}_M,\ \hat{\sigma}_M) > \beta\} = \gamma \quad (4)$$

The difference is rather important and the choice depends on the real question. The β-Expectation Tolerance Interval can be interpreted as the interval where each result has $p = \beta$ chance or predictive probability to fall in the future [28,29], while the β-γ-Content Tolerance Interval can be interpreted as the interval where there is γ confidence that a proportion β will lie in the future. The first one is about to the predictive probability of a single future measurement while the second one is about the proportion of future measurements, such as needed for QC or standard samples. The confusion in the literature comes from the confounding between probability and proportion. Asymptotically, or when a large number of results has been obtained, if each result has a probability β to fall within the interval, then β% of them will be within this interval: the proportion, aftermath, estimate this probability in frequentist statistics. It is this probability for each result of unknown samples that is the very criteria for quality, not where a proportion of future QC samples may fall in the future. Having, in a run, 5 QC results out of 6 within the acceptance limits certainly does not imply that there is 83% of the results of that run that are within the acceptance limits and therefore accurate. Five out of six is a promising indication that the run was performed appropriately, but nothing else. Adding a confidence on this future proportion, as introduced by the β-γ-Content Tolerance Interval, to guarantee

that most runs are likely to be accepted based on a proportion type of rule is good business practices to minimize costs but is not relevant to the very objective of validation. In reality, a proportion of future results has limited interest, only the probability of each single result matters: decisions will be made based on the quality of each of them, not based on the proportion of them being within the acceptance limits, since they remain unknown samples, as opposed to QC or standard samples. The other point of discussion, whatever the type of interval, is the selection of an appropriate Acceptance Level $\pi_{\min}$ that will guarantee results of quality. This is also linked to the number of runs to be rejected in routine.

4. Comparing decision rules for chromatographic assays

To compare and illustrate the capabilities of the various rules proposed to declare if an analytical procedure will be able to perform as expected (i.e., declared as valid), simulation studies have been performed. The four Validation Decision Rules compared are:

1. FDA rule: the mean observed bias (%RE) must be less than 15% and the intermediate precision (%CV) should also be less than 15%.
2. Total Error Rule: the mean bias (%RE) plus (+) the intermediate precision (%CV) must be less than 15%.
3. The β-Expectation Tolerance Interval must be included within the [−15%, 15%] acceptance limits. Various values of β will be envisaged.
4. The β-γ-Content Tolerance Interval must be included within the [−15%, 15%] acceptance limits. Various values of β will be envisaged but γ, the confidence, will always be fixed at 95%.

Procedures with "true" bias, intra-run and inter-run precision have been simulated by using Eq. (5):

$$X_{ij} = \delta + \mu_T + \alpha_i + \varepsilon_{ij} +, \quad \alpha_i \sim N(0,\ \sigma_A^2), \quad \varepsilon_{ij} \sim N(0, \sigma_E^2), \quad \frac{\sigma_A^2}{\sigma_E^2} = 1 \tag{5}$$

with δ being the bias and σ_A^2, σ_E^2 the between-run and the within-run variance, respectively.

The ratio inter-run variance over intra-run variance is set equal to 1, a value largely met for analytical procedures. The Intermediate Precision (IP) variance is the sum of the intra-run precision and the inter-run precision variances, i.e., $IP_X = \sigma_A^2,\ \sigma_E^2$. For the sake of simplicity, μ_T is set to 100 in the simulations. Those virtual procedures are only envisaged at one (concentration or quantity) level μ_T and therefore locally evaluated for their ability to be more or less on target. In reality a range of concentration or quantities is covered.

Since the real true performances (bias, intermediate precision) are unknown, procedures with true (relative) bias ranging from −20% to +20% and with intermediate precision ranging from 4% to 20% have evenly been chosen at random in the "space" of performance. 20,000 different procedures have been simulated according to this procedure and have been "virtually" run.

For each virtual procedure simulated the true probability π a measure to fall with the acceptance limits [−λ, +λ] is computed as follows ($\mu_T = 100$) under normality assumption (Eq. (6)):

$$\begin{aligned} \pi &= P(X - \mu_T < \lambda) - P(X - \mu_T < -\lambda) \\ &= \phi\left(\frac{\lambda - \delta}{\sqrt{\sigma_A^2 + \sigma_E^2}}\right) - \phi\left(\frac{-\lambda - \delta}{\sigma_A^2 + \sigma_E^2}\right) \end{aligned} \tag{6}$$

Out of those 20,000, 6944 or about 35% of the virtual procedures do have a true probability π to provide accurate results greater than 0.67.

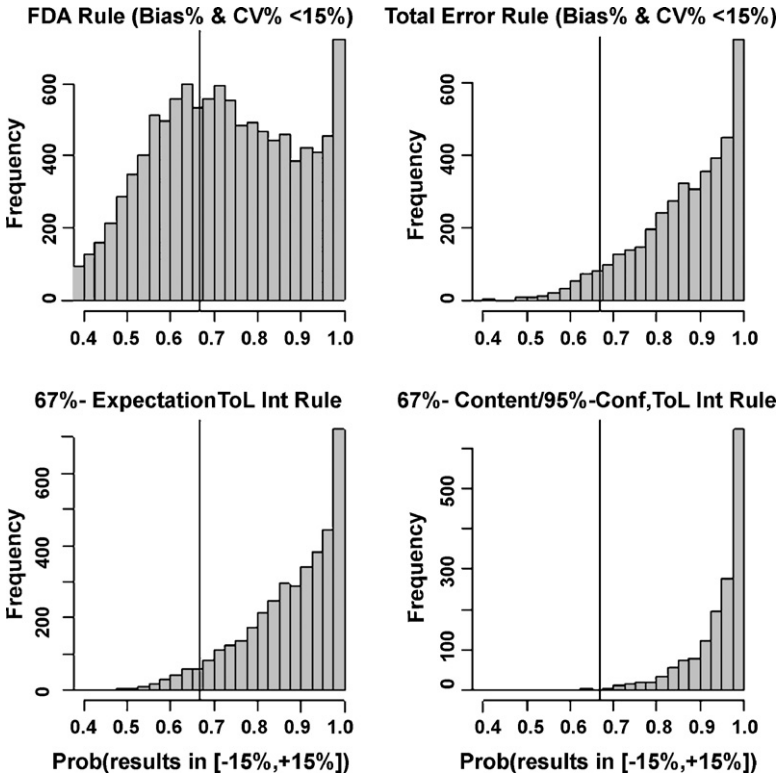

Fig. 1. Distribution of (virtual) procedures out of 20,000 that passed one of the four Validation Decision Rules as a function of the true probability of results obtained after the validation phase that are within the acceptance limits [−15%, +15%].

First a "validation phase" that includes $J = 6$ runs with $I = 6$ independent replicates within each run was performed and the various decision rules applied using the $J \times I = 36$ measurement generated that way. See Appendix A for computational details.

If a procedure passed a rule, then the procedure was used in virtual "routine" for 500 additional runs of 36 measurements per run. Each run included in addition 6 QC samples and the proportion of runs with 4 or more QC samples within the acceptance limits [−15%, 15%] was summarized to estimate the percentage of runs accepted using the 4-6-15 rule in routine after the pre-study validation. The true probability of each virtual procedure declared as valid at the pre-study stage is also reported to evaluate the ability of each decision rule to reject inappropriate procedures and select the good ones.

If β, a minimum quality level, is fixed to 67% for pre-study decision rule, as suggested by the statement "*Acceptance criteria: At least 67% (4 out of 6) of QC samples should be within 15% of their respective nominal value*" then Fig. 1 shows, for each of the four rules envisaged, the distribution of selected procedures according to their quality levels, i.e., the true probability a result will fall within the acceptance limits. The frequency is the number of (virtual) procedures out of 20,000 that passed one of the four Validation Decision Rules defined above. The first observation is that the FDA rule clearly accepts a larger number of procedures (10,871 out of 20,000 = 54%) than the three other rules. As direct consequence of this, given virtual procedures are selected at random, 39% of the procedures claimed as "valid" using the FDA decision rule do in fact have a true probability to produce results within the acceptance limits smaller than 0.67, the intended minimum quality level targeted. It means that using this rule there is only about 60% chance that the selected procedure will provide results whose probability to be in the acceptance limits is greater than or equal to 0.67. Symmetrically there are about 40% of the procedures declared as "valid" when using the FDA rule, whose individual results have a probability smaller than 0.67 to be within the acceptance limits. The risk of inaccurate results is therefore high and without control with this commonly recommended and used rule. The Total Error Rule, however, like the β-Expectation Tolerance Interval appears to be more appropriate when the quality level is targeted to be 67%. Both decision rules accept respectively only 20% and 18% of the 20,000 virtual procedures. With the Total Error Rule there is only 6% of validated procedures whose true probability of accurate results are slightly inferior to 0.67. The same applies to the β-Expectation Tolerance Interval with about 5% chances to select procedure that may not meet the minimal quality level, a risk acceptable. In addition, with these two rules, about 55% of the truly valid procedures (true $\pi_{\min} > 0.67$) are declared as valid. The fact that the Total Error Rule and the β-Expectation Tolerance Interval Rule are very comparable when the quality level is set to 67% is intuitively consistent: the interval around the mean + or − one standard deviation, according to the normal distribution, contains 67% of the values. With the β-γ-Content Tolerance Interval (67%-content and 95%-confidence) as rule, however, only 8% of the 20,000 virtual procedures are declared as valid and the risk to keep a procedure whose true probability to provide accurate results being smaller than 0.67 drop below 1%. But this is too conservative since there are about 78% of truly valid procedures that are declared as not valid and not retained using this decision rule. The major drawback of the last rule is that there is very little chance of accepting procedures, even an acceptable one. The use of the β-γ-Content Tolerance Interval is therefore not recommended for this purpose as previously mentioned, i.e., the rule is for the proportion of results, not for the probability a result, whatever the result, is in the acceptance limits. This latter is not economically appropriate because it limits drastically the number of truly valid methods that may be selected as such.

Fig. 2 shows the distribution of proportion of accepted runs out of 500 runs simulated for same procedures declared as valid after the pre-study phase as in Fig. 1, still using the same quality level 67%. The frequency is the number of (virtual) procedures out of 20,000 that have been simulated. When using the FDA rule, there are very high chances to select and use in routine procedures showing a low proportion of runs being accepted in routine. If the run acceptance

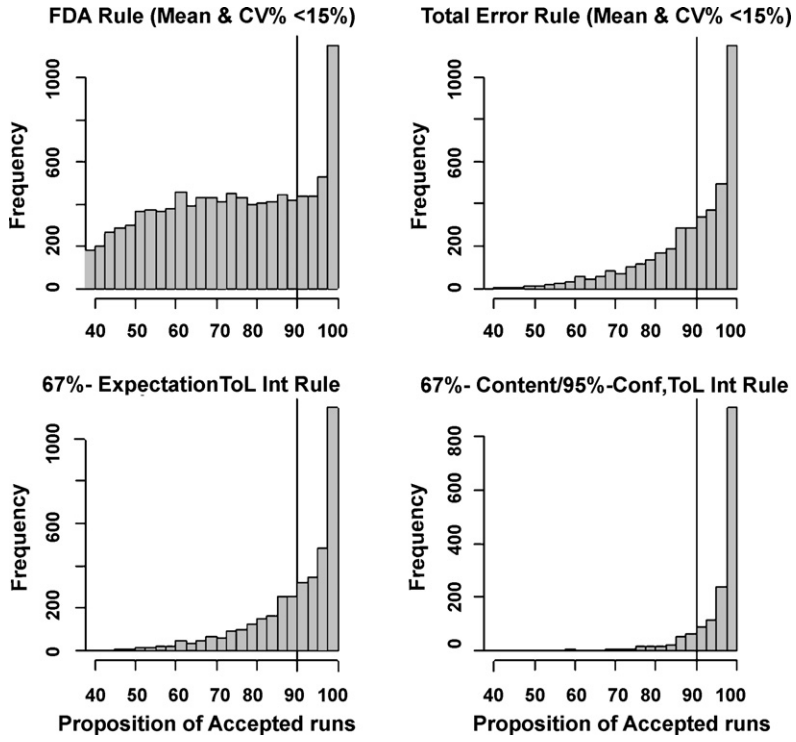

Fig. 2. Distribution of (virtual) procedures out of 20,000 that passed successfully one of the four Validation Decision Rules as a function of proportion of runs accepted using the 4-6-15 rule during the in-study validation. The vertical bar represents the 90% arbitrary proportion of run accepted.

rule prevents releasing poor quality results, the FDA rule is nevertheless not recommendable from an economical point of view. Using as a simple rule-of-thumb that ideally 90% of the runs performed in routine should be accepted to bring confidence in the results obtained, then with the FDA rule, about 76% of the "valid" procedures do not achieve this minimum proportion of 90% of runs accepted as opposed to 42% when using the Total Error Rule, 36% with the β-Expectation Tolerance Interval Rule and about 5% for the few procedures that passed the β-γ-Content Tolerance Interval rule. This latter result about the β-γ-Content Tolerance Interval is as expected since the confidence γ was set to 95%: therefore only 5% of the runs have less than 67% (β) of QC samples within the acceptance limits.

As already suggested, increasing the quality level $\pi_{\min}$ to 80%, instead of 67%, is a way to give appropriate guarantees that both 90% of the runs will be accepted using the 4-6-15 rule in routine and that each result has a high probability, namely 0.8, to be accurate. That way two objectives are reached with one stone: more quality results and higher economical efficiency of the laboratories since less runs will need to be reprocessed.

Fig. 3 shows the distribution of the "valid" procedures, out of 20,000, as a function of the proportion of accepted runs after decision to use them has been made using the β-Expectation Tolerance Interval Rule, for 4 values of β: 67%, 80%, 90% and 95%. Clearly, as expected, the chances to have a high number of runs accepted in routine are directly linked to the quality level targeted using the tolerance interval. When β, the quality level, is set to 80%, on average 95% of the runs are accepted and there is less than 2% chances a procedure declared as valid in pre-study phase will result in less than 90% of the runs being accepted. Increasing the value of β to 90% or 95% obviously will increase the future proportion of accepted runs, to 96% and 99% respectively, but will dramatically decrease the chances to accept a procedure during the validation phase from 18% to 11%, 6% and 3% respectively. With the β-Expectation Tolerance Interval a trade-off between today's acceptance of procedures and tomorrow's acceptance of runs can be found. The same applies to the accuracy of the future results that will be produced when using this rule for decision (Fig. 4).

Fig. 4 shows the distribution of true probabilities of a result to be within the acceptance limits when the β-Expectation Tolerance Interval Rule is envisaged for declaring a procedure valid for various quality level β levels ranging from 67%, 80%, 90% to 95%. As it appears in Fig. 4, with this rule there are guarantees that only procedures that are likely to provide results with a predefined quality level are selected appropriately, i.e., without being too liberal or too conservative. There is only 5–10% risks to have procedures whose quality level may be slightly below the targeted quality level β. This is an acceptable risk that can be improved by increasing the number of experiments to perform during the validation phase to have better estimates of procedure's performance.

5. Comparing decision rules for ligand-binding assays

Still in the proceeding of the conference, there is a new proposal and subtle change that has been introduced specifically for the ligand-binding assays (LBA). The new acceptance criteria states: "*For a method to be considered acceptable, it is recommended that both the interbatch imprecision (%CV) and the accuracy, expressed as absolute mean bias (%RE) be ±20% (25% at LLOQ and ULOQ). As an additional constraint to control method error, it is recommended that the target total error (sum of the absolute value of the %RE [accuracy] and precision [%CV]) be less than ±30% [±40% LLOQ and ULOQ].* The proceeding states that "*The additional constraint of total error allows for consistency between the criteria for pre-study method validation and*

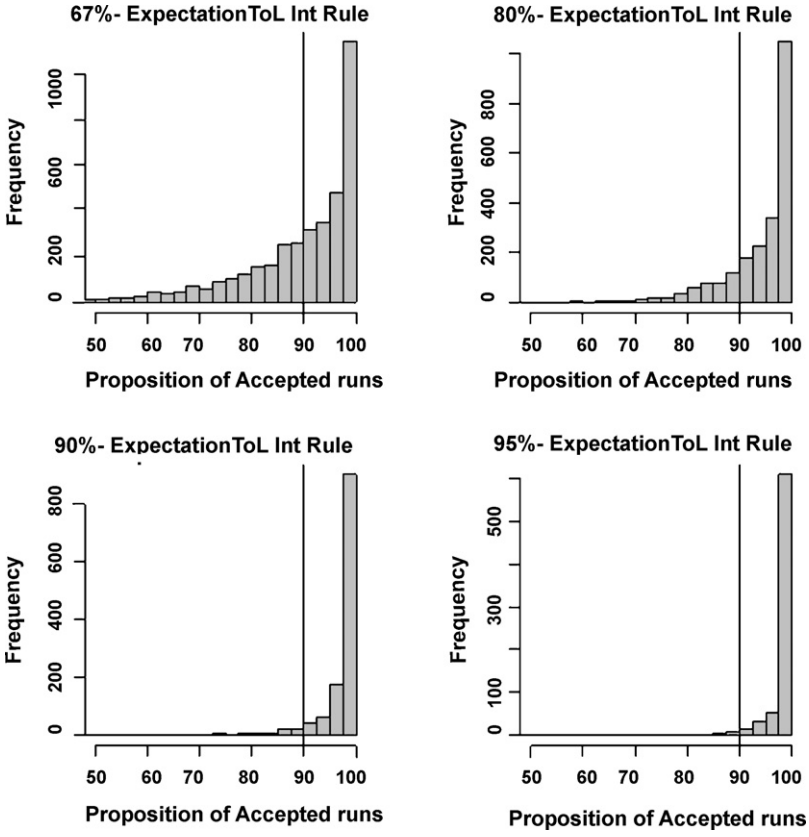

Fig. 3. Distribution of (virtual) procedures out of 20,000 that passed successfully the β-Expectation Tolerance Interval Rule for values of β: 67%, 80%, 90% and 95% as a function of proportion of runs accepted using the 4-6-15 rule during the in-study validation.

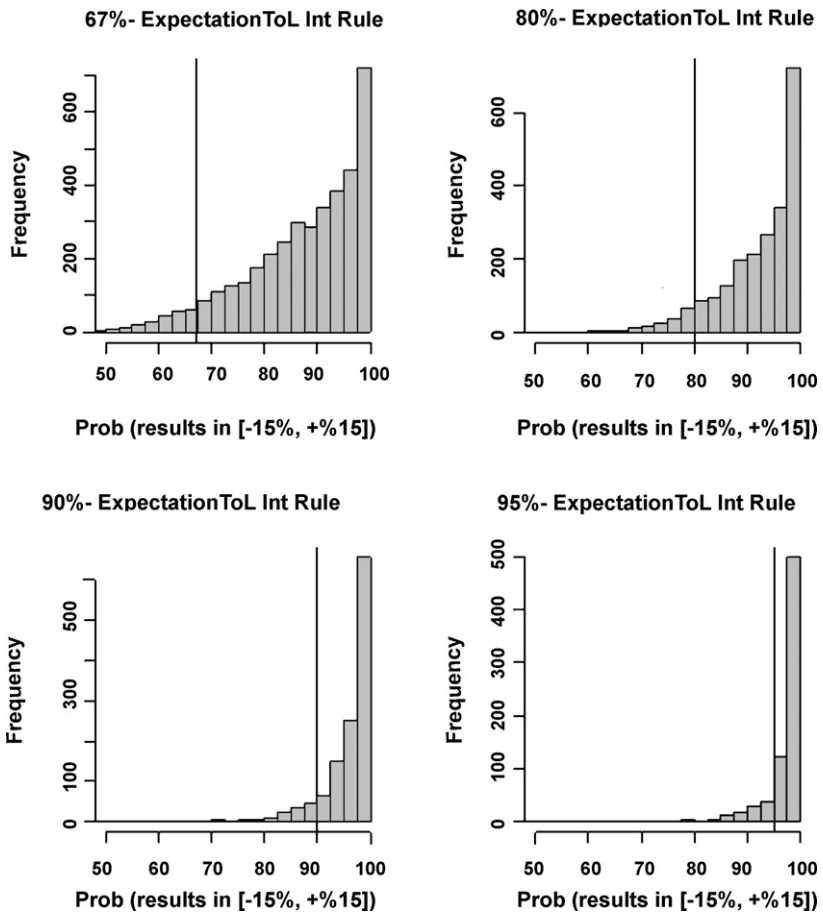

Fig. 4. Distribution of (virtual) procedures out of 20,000 that passed the β-Expectation Tolerance Interval Rule for different values of β: 67%, 80%, 90% and 95% as a function of the true probability of having accurate results, i.e., within the acceptance limits [−15%, +15%]. The vertical line is the corresponding β value chosen.

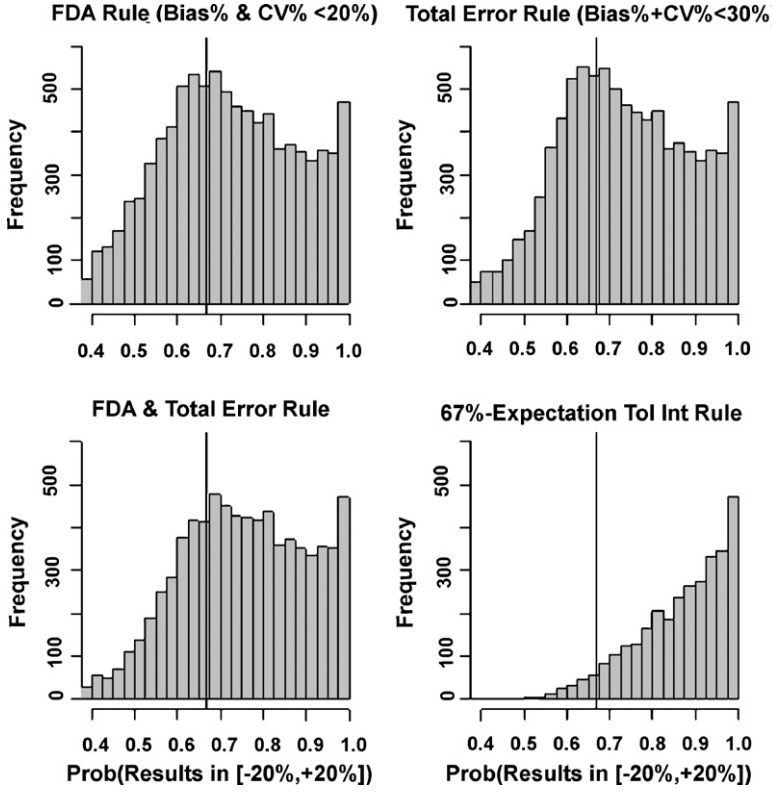

Fig. 5. Distribution of (virtual) procedures, out of 15,000, that passed one of the four Validation Decision Rules as a function of the true probability of producing accurate results, i.e., within the acceptance limits [−20%, +20%]. The vertical bar is the 67% quality level.

in-study batch acceptance". Let us investigate by simulation if the announced objective is met. For the LBA, the four pre-study decision rules that will be compared with respect to their performances are:

1. FDA rule: the mean observed bias (%RE) must be less than 20% and the intermediate precision (%CV) should also be less than 20%
2. Total Error Rule: the sum of the mean bias (%RE) and the intermediate precision (%CV) must be less than 30%.
3. Combined Rule: FDA rule and Total Error Rule should be satisfied.
4. The β-Expectation Tolerance Interval must be included within the [−20%, 20%] acceptance limits. β = 67% will be envisaged.

As for the simulated procedures in the previous section, the ratio inter-run variance over intra-run variance is set equal to 1. Since in the real world the true performances (bias, intermediate precision) are unknown, procedures with true (relative) bias ranging from −30% to +30% and with intermediate precision ranging from 6% to 30% have been evenly chosen at random in the "space" of performance. 15,000 procedures selected according to this procedure have been "virtually" run. First a "validation phase" that includes 6 runs with 6 independent replicates within each run was performed and the various decision rules applied. If a procedure passed a rule, then the procedure was use in virtual "routine" for 500 runs of 36 measurements per run. In addition, 6 QC samples per run were performed and the proportion of runs with 4 or more QC within the acceptance limits [−20%, 20%] was computed to estimate the percentage of runs accepted using the 4-6-20 rule.

As can be seen in Fig. 5, the Total Error Rule for LBA procedures here is performing as poorly as the conventional FDA rule. This result may contrast with the nice properties of the similar decision rule as noticed about the procedures in the previous section that could be assimilated to chromatographic procedures (see Fig. 1). But the major difference here is that, because of misunderstanding about the concept of Total Error, the acceptance limits for the Total Error have been expanded from 20% to 30%, while with the chromatographic procedures the 15% limits was kept for all rules. Therefore by enlarging to 30%, and even by combining both FDA (at 20%) and Total Error (at 30%) rules, the risk is high – about 30% – to select procedures with true low quality levels and thus producing results that have low probability to be accurate. This unfortunate proposal comes from the recurrent confusion that exists between bias and accuracy, between performance of procedures and accuracy of results. The proposal is presented as an improvement to increase the quality of future results, but in fact this proposal has no effect on the quality of the procedures that could be claimed as valid and therefore this proposal does not reduce the risk of producing inaccurate results.

6. Conclusion

The proposed rules based on the observed bias and precision fail to achieve their intended objective and risks are very high to release results of low quality. This risk is hidden by the fact that more procedures can pass the decision rules and therefore is largely perceived as an appropriate rule for procedures, but it is not an appropriate set of rules for the results themselves. The simulation studies highlight, by examining the quality of results and runs, the undesired consequences of pre-study decision rules that have not been carefully examined and tailored for their purpose. The performances of an analytical procedure are not known after validation studies, they are simply estimated with uncertainty. Due to the important and pivotal role played by quantitative procedures, statistically and risk sounds methods for making decisions should be applied. The matter is not always obvious and straightforward to understand, indeed, but the complexity of the concepts is not a valid excuse for using inappropriate methods. Proper statistical solutions for this purpose exist and have been proposed in the literature since 1942 and today's computational facilities make those proven approaches available to all analysts. In the world of analytical sciences, it is not the procedure that is released: it is the guarantee that it will produce quality results in the future that is released. The simulation studies presented here clearly show that no such guarantee is available using the regulatory rules such as the FDA rule and the AAPS/FDA workshop interpretation of Total Error Rule as defined in [1]. As anticipated and known from a theoretical point of view, the β-Expectation Tolerance Interval is providing the waited guarantees about future results given past estimated performances. Customer's risk and producer's risk are even duly balanced. The risks are chosen and known by the analyst upfront. Patients, industry and regulatory bodies will all benefit when moving into that direction.

Acknowledgements

The authors are very grateful to the anonymous reviewers for providing important comments that led to significant improvements of this article. A research grant from the Belgium National Fund for Scientific Research (FRS-FNRS) to E. Rozet is gratefully acknowledged.

Appendix A. Computational details

When the true nominal value μ_T is set to 100 for the sake of simplicity, the following statistics were estimated on the 36 measurements of pre-study validation phase for obtaining and applying four pre-study rules:

$$\hat{\delta} = \frac{1}{IJ}\sum_{ij} X_i - \mu_T = \hat{\mu} - \mu_T, \text{ is the estimated bias.}$$

$$MSE = \frac{1}{J(I-1)}\sum_{ij}(X_{ji} - \bar{X}_{.i})^2, \quad MSA = \frac{1}{J-1}\sum_{j} I(\bar{X}_{.i} - \bar{X})^2$$

$$\hat{\sigma}_X^2 = \frac{1}{J(I-1)}MSA + \left(1 - \frac{I}{J}\right)MSE,$$

is the estimatedintermediate variance.

J is the number of runs and I the number of independent replicates within each run. For the simulation I = 6 and J = 6.

The Tolerance Intervals were estimated as follows:

$$TI = [\hat{\mu} - k\hat{\sigma}_X, \quad \hat{\mu} + k\hat{\sigma}_X]$$

When the β-Expectation is considered, then:

$$k = k_E = t_{f,(1+\beta)/2}\sqrt{1 + 1 - \frac{1}{N_E}} \text{ with}$$

$$f = \frac{(MSA + (J-1)MSE)^2}{(1/(I-1))MSA^2 + ((J-1)/I)MSE^2}$$

the degrees of freedom of the t distribution

When the β-γ-Content Tolerance Interval is considered, then:

$$k = k_C = z_{(1+\beta)/2}\sqrt{1 + 1/N_E} \times \sqrt{1 + 1/\hat{\sigma}_X^2\sqrt{H_1^2(1/J)^2 MS_A^2 + H_2^2(1-1/J)^2 MS_E^2}}$$

with $H_1 = \dfrac{1}{F_{1-\gamma;1/J;\infty}} - 1, \quad H_2 = \dfrac{1}{F_{1-\gamma;1-1/J;\infty}} - 1$

$N_E = I(MSA + (J-1)MSA)/MSA$, the effective sample size. The four pre-study validation rules become:

FDA rule: $\hat{\sigma}_X^2/\hat{\mu} < \lambda$ and $\hat{\delta} < \lambda$

Total Error Rule: $\hat{\sigma}_X^2/\hat{\mu} + \hat{\delta} < \lambda$

β-Expectation Tolerance Interval Rule: $[\hat{\mu} - k_E\hat{\sigma}_X, \hat{\mu} + k_E\hat{\sigma}_X] \subset [-\lambda, +\lambda]$

β-γ-Content Tolerance Interval Rule: $[\hat{\mu} - k_C\hat{\sigma}_X, \hat{\mu} + k_C\hat{\sigma}_X] \subset [-\lambda, +\lambda]$

LBA FDA & Total Error Rule: $\hat{\sigma}_X^2/\hat{\mu} < 20\% \cap \hat{\delta} < 20\% \cap \hat{\sigma}_X^2/\hat{\mu} + \hat{\delta} < 30\%$

with $\lambda = 15\%$ or 20%

References

[1] C.T. Viswanathan, S. Bansal, B. Booth, A.J. DeStefano, M.J. Rose, J. Sailstad, V.P. Shah, J.P. Skelly, P.G. Swann, R. Weiner, AAPS J. 9 (2007) E30.

[2] Guidance for Industry: Bioanalytical Method Validation, US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), 2001.

[3] International Conference on Harmonization of Technical Requirements for registration of Pharmaceuticals for Human Use, Topic Q2 (R1): Validation of Analytical Procedures: Text and Methodology, IFPMA, Geneva, 2005.

[4] Ph. Hubert, J.-J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.-A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, J. STP Pharma Pratiques 13 (2003) 101.

[5] Ph. Hubert, J.-J. Nguyen-Huu, B. Boulanger, E. Chapuzet, P. Chiap, N. Cohen, P.-A. Compagnon, W. Dewe, M. Feinberg, M. Lallier, M. Laurentie, N. Mercier, G. Muzard, C. Nivet, L. Valat, J. Pharm. Biomed. Anal. 36 (2004) 579.

[6] Ph. Hubert, J.-J. Nguyen-Huu, B. Boulanger, E. Chapuzet, N. Cohen, P.-A. Compagnon, W. Dewe, M. Feinberg, M. Laurentie, N. Mercier, G. Muzard, L. Valat, J. STP Pharma Pratiques 16 (2006) 28.

[7] Ph. Hubert, J.-J. Nguyen-Huu, B. Boulanger, E. Chapuzet, N. Cohen, P.-A. Compagnon, W. Dewe, M. Feinberg, M. Laurentie, N. Mercier, G. Muzard, L. Valat, J. STP Pharma Pratiques 16 (2006) 87.

[8] D.L. Massart, B.G. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics, Elsevier, Amsterdam, 1988, p. 115.

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

[10] R.O. Kringle, Pharm. Res. 11 (1994) 556.

[11] B. Boulanger, P. Chiap, W. Dewe, J. Crommen, Ph. Hubert, J. Pharm. Biomed. Anal. 32 (2003) 753.

[12] D. Hoffman, R. Kringle, J. Biopharm. Stat. 15 (2005) 283.

[13] E. Rozet, A. Ceccato, C. Hubert, E. Ziemons, R. Oprean, S. Rudaz, B. Boulanger, P. Hubert, J. Chromatogr. A 1158 (2007) 111.

[14] S.S. Wilks, Ann. Math. Stat. 13 (1942) 400.

[15] D.J. Finney, Statistical Method in Biological Assay, Charles Griffin, London, 1978.

[16] ISO 5725, Application of the Statistics-Accuracy (Trueness and Precision) of the Results and Methods of Measurement, Parts 1 to 6, International Organization for Standardization (ISO), Geneva, Switzerland, 1994.

[17] ISO 3534-1, Statistics—Vocabulary and Symbols, International Organization for Standardization (ISO), Geneva, Switzerland, 2006.

[18] ISO VIM. DGUIDE 99999.2, International Vocabulary of Basic and General Terms in Metrology (VIM), 3rd edition, ISO, Geneva, 2006 (under approval step).

[19] Eurachem, The Fitness for Purpose of Analytical Methods, 1998.

[20] Analytical Methods Committee, AMC Technical Brief 13 (2003). http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

[21] Codex Alimentarius Commission, Procedural Manual, 15th edition, FAO, Rome, 2005.

[22] B. Boulanger, W. Dewe, P. Hubert, B. Govaerts, C. Hammer, F. Moonen, AAPS Third Bioanalytical Workshop, May 1–3, Arlington, VA, 2006, http://www.aapspharmaceutica.com/meetings/meeting.asp?id=64.

[23] B. Boulanger, V. Devanaryan, W. Dewé, W. Smith, in: A. Dimitrienko (Ed.), SAS Book on Pre-clinical Statistics, SAS Institute, 2007.

[24] B. Boulanger, W. Dewé, A. Gilbert, B. Govaerts, M. Maumy, Chemometr. Intell. Lab. Syst. 86 (2007) 198.

[25] R.W. Mee, Technometrics 26 (1984) 251.

[26] R.W. Mee, Communication in Statistics—Theory and Methods 17 (1988) 1465.

[27] M. Feinberg, B. Boulanger, W. Dewe, P. Hubert, Anal. Bioanal. Chem. 380 (2004) 502.

[28] J. Aitchison, J. R. Stat. Soc., Ser. B 26 (1964) 161.

[29] M. Hamada, L.M. Moore, V. Johnson, Technometrics 46 (2004) 452.

[30] I. Guttman, Statistical Tolerance Regions: Classical and Bayesian, Griffin's Statistical Monographs and Courses, London, 1969.