Validation of qualitative microbiological test methods

MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1663 Published online 21 November 2014 in Wiley Online Library

Validation of qualitative microbiological test methods

Pieta C. IJzerman-Boon[a]* and Edwin R. van den Heuvel[b,c]

This paper considers a statistical model for the detection mechanism of qualitative microbiological test methods with a parameter for the detection proportion (the probability to detect a single organism) and a parameter for the false positive rate. It is demonstrated that the detection proportion and the bacterial density cannot be estimated separately, not even in a multiple dilution experiment. Only the product can be estimated, changing the interpretation of the most probable number estimator. The asymptotic power of the likelihood ratio statistic for comparing an alternative method with the compendial method, is optimal for a single dilution experiment. The bacterial density should either be close to two CFUs per test unit or equal to zero, depending on differences in the model parameters between the two test methods. The proposed strategy for method validation is to use these two dilutions and test for differences in the two model parameters, addressing the validation parameters specificity and accuracy. Robustness of these two parameters might still be required, but all other validation parameters can be omitted. A confidence interval-based approach for the ratio of the detection proportions for the two methods is recommended, since it is most informative and close to the power of the likelihood ratio test.

Keywords: accuracy; detection proportion; specificity; false positives; limit of detection; generalized most probable number estimator

1. INTRODUCTION

New or alternative microbiological test methods (e.g. a rapid sterility test) must be validated before they are used in practice. Regulatory guidelines, such as the European Pharmacopoeia (EP) 5.1.6 [1] and the United States Pharmacopoeia (USP) <1223> [2], describe requirements for a successful validation study. They give recommendations on design and analysis of experimental studies and provide minimal acceptance criteria. However, the EP and USP have different views on the necessary validation parameters. For qualitative tests, they both require specificity, limit of detection, and robustness, but the EP also requires accuracy and precision, although they only discuss accuracy, and the USP requires repeatability and ruggedness, although they only discuss ruggedness.

The EP and USP define specificity in similar ways. It is the ability to detect the required range of micro-organisms that may be present in the test sample. They also mention that extraneous matter in the test system (e.g. growth medium) should not interfere with the test. However, there is no clear description of the proposed experiment nor do they provide criteria for accepting the alternative method.

The limit of detection (LOD) is defined in both EP and USP as the lowest number of micro-organisms in a test sample that can be detected under the stated experimental conditions. It refers to the number of micro-organisms in the original sample before any dilution or incubation steps, and not to the number of micro-organisms present at the time of testing. Both guidelines recommend to determine first an inoculum that provides at least 50% of the samples showing growth in the pharmacopoeial or compendial method and then to test repeated samples (at least five) with both methods at this inoculum. Their proportions of positive test samples should be compared with a chi-square test and this test should not be significant. The USP suggests a second approach where they make use of a serial dilution experiment from which the most probable number (MPN) can be calculated [3]. The suggested criterion is that the 95% confidence intervals on the MPN for both methods should overlap.

Both the EP and USP define robustness as a measure of the capacity of the alternative method to remain unaffected by small but deliberate variations in method parameters. They recognize that robustness of the alternative method need not be compared with the pharmacopoeial method. Both guidelines feel that robustness is best suited for evaluation by the supplier of the equipment, unless critical parameter settings are modified by the user. The USP mentions that general criteria cannot be set a priori and that they should be tailored for each method.

The EP proposes for accuracy to study the degree of agreement between the alternative and pharmacopoeial method, because a side-by-side comparison of the methods on identical samples with a low probability of failure requires too many tests to demonstrate equivalence. They suggest to determine the false positive and false negative rates for the alternative method against the

[a] *Center for Mathematical Sciences - Europe, MSD, Oss, The Netherlands*

[b] *Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands*

[c] *Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*

**Correspondence to: Pieta C. IJzerman-Boon, Center for Mathematical Sciences - Europe, MSD, Oss, The Netherlands.*
E-mail: pieta.ijzerman@merck.com

pharmacopoeial method using a standardised, low-level inoculum. They specify that the frequency of recovery (true positives) of the alternative method is at least as high as for the pharmacopoeial method.

The USP describes ruggedness as a measure of precision, which coincides with the definition in chemical method validation (ICH Q2(R1) [4]), where the test result may be affected by a variety of normal test conditions (e.g. analysts, reagents, instruments, etc.). They do not provide any experimental settings to be able to estimate this measure of precision nor provide any criteria to be able to judge the level of precision, but instead they leave the investigation of ruggedness to the supplier.

The purpose of this article is to demonstrate that under certain statistical assumptions, the only parameters that must be validated for an alternative qualitative method are specificity and accuracy. They should be studied in a comparison with respect to the compendial method. This conclusion follows from a statistical formulation of microbiological detection mechanisms and an investigation of the optimal experiment to evaluate these parameters. Robustness may still be required, but this is outside our scope, because we believe that robustness is not a separate parameter, but rather an investigation of the stability of the validation parameters. Our view on method validation for microbiological qualitative tests differs from the view of the guidelines, but our goal is not to discredit the guidelines, it is rather to provide a statistical perspective that would complement and improve the guidelines.

The paper is organized as follows. The next section presents the statistical model for detection of single organisms for qualitative tests. Three specific choices of detection mechanisms, which fall within this general description, have been published in literature before (e.g. [3,5–8]). The detection mechanisms are used to formulate the expected proportions of positive test samples in specific validation experiments. Section 3 then describes likelihood-based methods for statistical inference and uses asymptotics to optimize the validation experiment. Simulations supporting the proposed strategy for validation are presented in Section 4. The final section links the parameters of our detection mechanisms to the validation parameters and discusses our theory in relation to the current guidelines.

2. STATISTICAL DETECTION OF MICRO-ORGANISMS

Consider a dilution i from which multiple test samples may be collected for validation purposes. Let Y_{hij} be the number of micro-organisms in test sample j from dilution i intended for microbiological method h (alternative: $h = 1$, pharmacopoeial: $h = 2$). In principle, multiple dilutions may be used ($i = 1, ..., m$), and the number of test samples collected for each method may differ per dilution ($j = 1, ..., n_{hi}$). It is assumed that the total number of organisms in dilution i is N_i and the average number of organisms per test unit is λ_i. The relation between N_i and λ_i is given by $\lambda_i = v \cdot N_i/V_i$, with V_i the volume of dilution i and v the volume of each test sample. The parameter λ_i in dilution i may be referred to as the *bacterial density* of dilution i.

Given the number of micro-organisms N_i in dilution i, the marginal distribution of Y_{hij} is binomial with parameters N_i and p_i, with $p_i = v/V_i$. This distribution is true only when the micro-organisms in the dilution are randomly distributed throughout the dilution, that is, no clotting, no repelling, or any other systematic or dynamic positioning of the organisms in the dilution. In case the proportion p_i is small, the binomial distribution is close to the Poisson distribution with parameter $N_i p_i \approx \lambda_i$; see [3]. Additionally, if we would view N_i random with a Poisson distribution, the marginal distribution of Y_{hij} would be Poisson too. Therefore, we will assume that Y_{hij} has a Poisson distribution with parameter λ_i. Furthermore, we will assume that $Y_{hi1}, ..., Y_{hin_{hi}}$ are independent although they are correlated in practice. This correlation can be neglected when the volume V_i is large with respect to $v(n_{1i} + n_{2i})$, the total volume collected from dilution i [7].

2.1. Detection Mechanisms

When a test sample j from dilution i is tested with microbiological method h, we would obtain an outcome Z_{hij} indicating the absence or presence of micro-organisms in the test sample ($Z_{hij} \in \{0, 1\}$). It seems reasonable to assume that the outcome of a test sample is affected by the number of organisms Y_{hij} present in the test sample. Thus, the *detection mechanism* of microbiological method h can be described by a conditional probability

$$\pi_h(y) = P\left(Z_{hij} = 1 | Y_{hij} = y\right), \tag{1}$$

with $y \in \{0, 1, 2, ..., N_i\}$ the number of organisms present in the test sample. In practice, the detection mechanism for both microbiological methods is typically unknown and a validation study is performed to provide more knowledge on these mechanisms, in particular on the difference or similarity of $\pi_1(\cdot)$ and $\pi_2(\cdot)$ for the alternative and compendial method.

In the past, several assumptions have been made directly or indirectly on the detection mechanism in (1). For instance, many decades ago, it was assumed that the *limit of detection* of growth-based sterility tests (e.g. current pharmacopoeial method) was equal to one (e.g. [3,9]). This would imply that $\pi_2(y) = 1_{[1,\infty)}(y)$, with $1_A(y)$ equal to one when $y \in A$ and zero otherwise. Recently, this assumption has been relaxed to limits of detection L larger than one [7], resulting in a detection mechanism for microbiological method h equal to

$$\pi_h(y) = 1_{[L_h,\infty)}(y), \tag{2}$$

with $L_h \in \{1, 2, ...\}$ the limit of detection of microbiological method h. We will refer to this specific formulation in (2) as the *deterministic mechanism*. Indeed, when the number of micro-organisms in a test sample is at least equal to the limit of detection L_h, the microbiological method will detect the presence of micro-organisms with 100% certainty and when the number is lower than this limit of detection, the test sample is considered sterile with again 100% certainty.

The deterministic mechanism does not involve any stochastic component, thus another assumption for (1) is the *binomial mechanism*. This was introduced for the accuracy of enumeration tests [8] and it assumes that each micro-organism is detected with a certain a priori probability $\theta_h \in [0, 1]$, i.e.

$$\pi_h(y) = \begin{cases} 1_{[1,\infty)}(y) & \text{if } \theta_h = 1 \\ 1 - (1 - \theta_h)^y & \text{if } \theta_h < 1 \end{cases}. \tag{3}$$

In this approach, the a priori probability θ_h is referred to as the *detection proportion* of method h and represents the probability

of detecting a single micro-organism, that is, $\pi_h(1) = \theta_h$. Thus, in case the detection proportion is equal to one, the binomial mechanism reduces to the deterministic mechanism with a limit of detection of one.

The detection mechanisms (2) and (3) do not allow for false positive test results. Indeed, the probability of a positive (or contaminated) test result is equal to $\pi_h(0) = 0$ for both mechanisms. This assumption may be realistic for growth-based methods, because these methods cannot detect non-viable micro-organisms, but it may not be true for direct microbiological methods that would also detect particles or other cell material. This means that we should consider detection mechanisms with $\pi_h(0) = \eta_h$, with $\eta_h \in [0, 1)$. The binomial mechanism is then extended to the *zero-deflated binomial mechanism*

$$\pi_h(y) = \begin{cases} \eta_h + (1-\eta_h)\,1_{[1,\infty)}(y) & \text{if } \theta_h = 1 \\ 1-(1-\eta_h)(1-\theta_h)^y & \text{if } \theta_h < 1 \end{cases}. \tag{4}$$

The term zero-deflated is used because the number of negative test results (i.e. detecting zero organisms) is reduced due to the false positives. The parameter θ_h of method h is still referred to as the detection proportion of method h, although $\pi_h(1) = 1-(1-\eta_h)(1-\theta_h) \neq \theta_h$ when $\eta_h > 0$.

Note that the detection mechanisms in (3) and (4) are not new and have been used outside microbiology in for instance bioassays for insect viruses [6]. For these bioassays, the susceptibility of an insect for a virus and the unknown ingestion of virus particles both contribute to the mortality probability of an insect. The susceptibility would correspond to our detection proportion θ_h and the virus particles would correspond to the unknown number of organisms y in our test sample. The false positive rate η_h in (4) would then correspond to control mortality, that is, the probability of death of an insect that is unrelated to the virus. Moreover, our zero-deflated detection mechanism is essentially a special case of Abbott's formula [5] for the effectiveness of an insecticide.

2.2. Proportions of Positive Test Results

In microbiology, it is very difficult to spike samples with an exact number of micro-organisms. Thus, the detection mechanism in (1) cannot be estimated directly from different (spiked) levels of y. Even stronger, it is also impossible to spike exact numbers of N_i in dilution i, although they might be known approximately when some kind of special reference material (e.g. BioBalls™, [10]) is used. This is the reason for constructing multiple dilutions (serial or non-serial dilution experiments) with different bacterial densities. Thus, we can only estimate expected proportions of positive (or negative) test results for each of the different dilutions, that is, we obtain only information on the mean $\mu_{hi} = \mathbb{E}\left[\pi_h\left(Y_{hij}\right)\right]$.

Under the assumption that the number of micro-organisms Y_{hij} is Poisson distributed with parameter λ_i, the expected proportion μ_{hi} for detection mechanism (2) is given by the Poisson probability

$$\mu_{hi} = 1 - \sum_{y=0}^{L_h-1}\left[\frac{\lambda_i^y}{y!}\exp(-\lambda_i)\right]. \tag{5}$$

The parameters L_h and λ_i are in principle unknown, although the parameter λ_i is sometimes assumed to be known approximately when appropriate reference material is used.

Estimation of the limit of detection together with the bacterial density has been discussed elsewhere for the setting of non-serial dilutions, which were also fully tested in order to measure all spiked organisms [7,11]. This approach has generalized the MPN method [3], because it simultaneously estimates the bacterial density λ_i and the limit of detection L_h, instead of just the bacterial density. This method has been referred to as the most probable limit (MPL) of detection [12], but it assumes detection mechanism (2).

For detection mechanisms (3) and (4), the expected proportion is given by

$$\mu_{hi} = 1 - (1-\eta_h)\exp(-\theta_h\lambda_i), \tag{6}$$

with essentially three unknown parameters θ_h, η_h, and λ_i. The zero-deflated binomial detection mechanism extends the MPN method with two parameters, because the MPN method assumes and uses a Poisson probability of $1-\exp(-\lambda_i)$ for dilution i, not taking into account the two additional parameters (θ_h, η_h) for method h. Clearly, when the false positive rate would be neglectable ($\eta_h \approx 0$), the expected proportion in (6) reduces to the Poisson probability $1-\exp(-\theta_h\lambda_i)$, which is still a generalization of the MPN and is called the *one-hit model* in bioassays for insecticides [6].

The expected proportion in (6) identifies a serious problem for the estimation of the three parameters involved. The parameters θ_h and λ_i appear in the (zero-deflated) Poisson probability as a product. This demonstrates that the parameters θ_h and λ_i are not identifiable and only the product $\xi_{hi} = \theta_h\lambda_i$ can be estimated, in addition to the false positive rate η_h. Thus, an experiment with just one method makes it impossible to estimate the three parameters. Note that this issue of identifiability is unrelated to the number of dilutions. Indeed, in case multiple dilutions are used their densities may be proportional to the first dilution, that is, $\lambda_i = \lambda_1/d^{i-1}$ with d the dilution factor for serial dilutions. Each dilution would then help to estimate $\theta_h\lambda_1$ but not to separate θ_h and λ_1. Moreover, an experiment with both methods, even when the detection proportions θ_1 and θ_2 are different, will not help in the estimation of all five parameters η_1, η_2, θ_1, θ_2, and λ_i. An increased value for the bacterial density can easily be compensated with lower values of θ_1 and θ_2, as long as the ratio between θ_1 and θ_2 remains constant. Only if this ratio would be known or if the detection proportion of the compendial method would be known (e.g. equal to one), the density λ_i can be estimated. On the other hand, the ratio of θ_1 and θ_2 can always be estimated in the general setting of the expected proportion in (6). As a consequence, it would be possible to test the individual or combined null hypotheses $\eta_1 = \eta_2$ and $\theta_1 = \theta_2$ with an appropriate experiment. A natural test statistic for this null hypothesis would be the likelihood ratio test, because the method of maximum likelihood (ML) seems the most natural method of estimation (Section 3).

An interesting observation is that if the detection mechanism of method h is of the binomial type ($\eta_h \approx 0$), but the deterministic detection proportion (5) is fitted to the experimental data, then the limit of detection L_h would be most likely estimated with one ($L_h = 1$) and the bacterial density λ_i would be estimated with an estimate for $\xi_{hi} = \theta_h\lambda_i$. Indeed, the deterministic proportion in (5) substituted with these obtained estimates would become an estimate of the binomial proportion $1-\exp(-\theta_h\lambda_i)$ from (6) with $\eta_h \approx 0$. This would imply that the MPL estimate of the bacterial density λ_i in (5) is equal to the true density multiplied with the detection proportion θ_h. Thus, only when this detection

proportion would be one, the MPL method of [7] would estimate the density correctly. [11] have emphasized that in addition to a limit of detection of one, a high ratio between the MPL estimate and the assumed true density is important. They called this the *recovery* of spiked micro-organisms, which now appears to be directly related to the detection proportion for the binomial detection mechanism.

The expected proportions in (5) and (6) have been obtained by choosing a specific form of the detection mechanism in (1), but μ_{hi} can also be modelled directly with a curve of the form $\mu_{hi} = g(\alpha_h + \beta_h \log(\lambda_i))$, with $g : \mathbb{R} \to (0,1)$ an inverse link function, α_h and $\beta_h > 0$ microbiological method-specific parameters, and λ_i the expected concentration or bacterial density for dilution sample i. This approach is common in quantal response bioassays. In this form, the expected proportion would typically be equal to zero for a blank concentration, thus a generalization that would make it possible to include also false positives is the following

$$\mu_{hi} = \eta_h + (1 - \eta_h)\, g(\alpha_h + \beta_h \log(\lambda_i)). \tag{7}$$

This model has essentially four unknown parameters $\eta_h, \alpha_h, \beta_h$, and λ_i. In case β_h would be equal to one and the link function is of the complementary log-log form, the expected proportion in (7) becomes the zero-deflated binomial in (6), but in microbiology, the logistic model is frequently applied (e.g. [13]), although often incorrectly. Indeed, the general form in (7) also has the disadvantage that it is non-identifiable, because the product $\beta_h \log(\lambda_i)$ has similar issues as the product $\theta_h \lambda_i$ discussed earlier. Furthermore, the linear combination $\alpha_h + \beta_h \log(\lambda_i)$ introduces another issue, because the unknown concentration makes it possible to shift the parameter $\beta_h \log(\lambda_i)$ with a certain value that would then be compensated in the intercept α_h. Only when the bacterial density λ_i is known, it is possible to fit model (7) with logistic regression approaches [14]. Another issue is the lack of interpretation of the general form (7) in terms of an underlying detection mechanism $\pi_h(\cdot)$, in particular for the logistic model. It is unknown which detection mechanism $\pi_h(\cdot)$ would lead to (7) with a logistic function. This makes the general form less suitable for further study, and we will therefore focus on the special case of the zero-deflated binomial mechanism.

3. LIKELIHOOD-BASED INFERENCE

Assuming a proportion of positive test results $\mu_{hi} = \mu_{hi}(\theta_h, \eta_h, \lambda_i)$ of the form (6) for method h at dilution i, the null hypothesis $H_0 : \theta_1 = \theta_2 \wedge \eta_1 = \eta_2$ would indicate that both microbiological test methods are identical. This null hypothesis induces the null hypotheses $H_0 : \mu_{1i} = \mu_{2i} = \mu_i$ for all dilutions i. Considering just one dilution and defining the number of positive test samples tested with method h for dilution i by $Z_{hi\cdot} = \sum_{j=1}^{n_{hi}} Z_{hij}$, the ML estimator for the proportion of positive test results for dilution i is equal to $\hat{\mu}_i = (Z_{1i\cdot} + Z_{2i\cdot})/(n_{1i} + n_{2i}) \equiv \bar{Z}_{\cdot i\cdot}$. The ML estimator for the proportion μ_{hi} under the alternative hypothesis of $\mu_{1i} \neq \mu_{2i}$ is given by $\hat{\mu}_{hi} = Z_{hi\cdot}/n_{hi} \equiv \bar{Z}_{hi\cdot}$. The likelihood ratio test statistic for dilution i is then given by

$$\begin{aligned} LRT_i = 2 \sum_{h=1}^{2} \big(& Z_{hi\cdot} \left(\log \bar{Z}_{hi\cdot} - \log \bar{Z}_{\cdot i\cdot} \right) \\ & + (n_{hi} - Z_{hi\cdot}) \left(\log(1 - \bar{Z}_{hi\cdot}) - \log(1 - \bar{Z}_{\cdot i\cdot}) \right) \big). \end{aligned} \tag{8}$$

Note that in case one of the estimates $\bar{Z}_{hi\cdot}$ equals zero or one, the corresponding term in (8) involving this estimate is set equal to zero, because $\lim_{x \downarrow 0} x \log x = 0$.

The asymptotic distribution of likelihood ratio statistic LRT_i is derived by approximating this test statistic with a second-order Taylor expansion around μ_i. Because the zero-order and first-order terms vanish under the null hypothesis, the asymptotic behavior of LRT_i is determined by the second-order term of the Taylor expansion. When the sample sizes for the two methods are equal, that is, $n_{1i} = n_{2i} = n$, the second-order term becomes

$$\frac{n}{2} \left(\frac{\hat{\mu}_{1i} - \mu_i}{\sqrt{\mu_i(1 - \mu_i)}} - \frac{\hat{\mu}_{2i} - \mu_i}{\sqrt{\mu_i(1 - \mu_i)}} \right)^2. \tag{9}$$

Because $\sqrt{n}(\hat{\mu}_{hi} - \mu_i)$ converges to $N(0, \mu_i(1 - \mu_i))$ under the null hypothesis, and the sum of two normal distributions converges to a normal distribution again, the second-order term in (9) converges to a χ^2- distribution with 1 degree of freedom. This asymptotic result is a well-known result due to [15].

Under the alternative hypothesis, we can derive the asymptotic distribution when we consider local alternatives, that is, $\mu_{hi} = \mu_i + n^{-1/2}\Delta_{hi}$ with $\mu_i = (\mu_{1i} + \mu_{2i})/2$, the average of the two true expected proportions of positive results. The Taylor expansion of the likelihood ratio test LRT_i with respect to μ_i remains the same. Furthermore, by writing $\hat{\mu}_{hi} - \mu_i = \hat{\mu}_{hi} - \mu_{hi} + \mu_{hi} - \mu_i$, it can be seen that $\sqrt{n}(\hat{\mu}_{hi} - \mu_i)$ converges to $N(\Delta_{hi}, \mu_i(1 - \mu_i))$. Hence, the second-order term in (9) now converges to a non-central χ^2-distribution with 1 degree of freedom, and with non-centrality parameter

$$nc = \frac{(\Delta_{1i} - \Delta_{2i})^2}{2\mu_i(1 - \mu_i)} = \frac{n(\mu_{1i} - \mu_{2i})^2}{2\mu_i(1 - \mu_i)}. \tag{10}$$

The non-central χ^2-distribution can be used to determine the optimal value λ_O for the bacterial density λ_i for dilution i and the minimal sample size n_O to achieve a prespecified power with the likelihood ratio test under the alternative hypothesis. The optimal values λ_O and n_O will depend on the parameters θ_1, θ_2, η_1, and η_2. However, the structure of the non-centrality parameter in (10), a product of n with a function of (μ_{1i}, μ_{2i}), implies that the bacterial density can be optimized independently of the sample size. Furthermore, because each dilution would select the same optimal bacterial density λ_O, a multiple dilution experiment does not contribute to the power of the likelihood ratio test.

For this single dilution experiment, the optimal bacterial density λ_O was numerically determined for different settings of the parameters θ_1, θ_2, η_1, and η_2, and the power was calculated for a sample size of n equal to 150, 200, and 250 (Table I). We assumed a perfect pharmacopoeial method ($\theta_2 = 1$, $\eta_2 = 0$), a detection proportion of $\theta_1 = 0.7$ for the alternative method, and different values for the false positive rate η_1. The value of 0.7 was chosen because the guidelines suggest for quantitative methods that the recovery of the alternative method compared with the pharmacopoeial method should at least be 70%. Table I demonstrates that approximately 200 samples are needed at the optimal λ_O to detect this difference in detection proportions with 80% power when neither of the two methods has false positives.

The numerical calculations demonstrated further that the optimal density λ_O is either close to two or equal to zero. When the difference in the false positive rates would dominate the likelihood ratio test more than the difference in detection proportions,

Table I. Asymptotic power at optimal bacterial densities and at a density $\lambda = 2$, for $\theta_2 = 1$, $\eta_2 = 0$, and $\theta_1/\theta_2 = 0.7$.

		Power (%) at λ_O			Power (%) at $\lambda = 2$		
η_1	λ_O	$n = 150$	$n = 200$	$n = 250$	$n = 150$	$n = 200$	$n = 250$
0.00	1.84	69.0	81.0	88.7	68.8	80.8	88.6
0.01	1.90	67.2	79.3	87.4	67.1	79.2	87.3
0.02	1.95	65.4	77.6	86.0	65.3	77.6	86.0
0.03	2.01	63.5	75.9	84.5	63.5	75.9	84.5
0.04	0.00	69.7	81.5	89.1	61.7	74.1	83.0
0.05	0.00	79.2	89.3	94.7	59.8	72.2	81.3

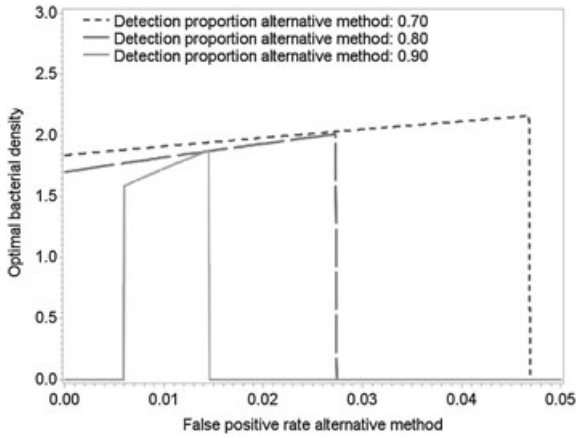

Figure 1. Optimal density values for different settings of the alternative method and a compendial method with $\eta_2 = 0.01$ and $\theta_2 = 0.95$.

the optimal density is zero, otherwise it is close to two. Other settings provide a similar picture as illustrated in Figure 1. In this setting, the compendial method had a false positive rate $\eta_2 = 0.01$ and a detection proportion $\theta_2 = 0.95$. The optimal bacterial density λ_O is either in the range of 1.5 to 2.5 (not all data shown) or equal to zero for settings of practical interest.

Note that an optimal density for estimation of the detection proportion θ_h for a binomial detection mechanism (i.e. no false positives) would be $\lambda = 1.59/\theta_h$ for method h [16,17] when the density λ can be set precisely. This differs from our optimal concentration λ_O. The reason is that λ_O is determined by maximizing the power for a comparison of two methods instead of minimizing the standard error of the parameter estimate for a single method, which is a different optimization problem. Our optimal bacterial density depends on all four parameters θ_1, θ_2, η_1, and η_2, but even if we would assume a binomial detection mechanism, we would still end up with a different optimal value λ_O, because it would depend on both detection proportions θ_1 and θ_2. Moreover, when the detection proportions would be equal, the power is reduced to the significance level and thus independent of the bacterial density, that is, under the null hypothesis there is no optimal concentration. On the other hand, it is expected that the optimal density for testing two detection proportions for the binomial mechanism is in the range of $1.59/\theta_1$ and $1.59/\theta_2$, because this optimal density would most likely provide small standard errors on both parameter estimates, which makes testing most powerful asymptotically. This may explain our observed range of 1.5 to 2.5 for the optimal bacterial density.

Although the optimal density depends on the parameters θ_1, θ_2, η_1, and η_2, choosing a value of λ equal to two as a general strategy, even if it is unequal to the optimal bacterial density, may still give a reasonable power, although it does not give the optimal power. Table I presents the asymptotic power values under this strategy of $\lambda = 2$ for a single dilution. The results show that if the optimal density would be zero, then the use of a bacterial density with $\lambda = 2$ would lead to a drop in power. The drop can be substantial. For $\eta_1 = 0.05$ and a sample size of 150 test samples per method, the power reduces from approximately 80% to 60%. Otherwise, when the optimal value λ_O is in the range of 1.5 to 2.5, the power at $\lambda = 2$ is quite similar to the optimal power.

Thus, when the false positive rates are known to be close to zero or almost equal, an appropriate strategy to test the null hypothesis $H_0 : \theta_1 = \theta_2$ is the use of a single dilution with a bacterial density around $\lambda = 2$. Spiking this level does not have to be very precise, because deviations from $\lambda = 2$ have only small effect on the power. When the false positive rates are not known or known to be quite different, a single dilution with $\lambda = 2$ might be insufficient to test the null hypothesis $H_0 : \theta_1 = \theta_2 \wedge \eta_1 = \eta_2$. Even if this single dilution would have a high power, it does not give any information about the underlying parameters.

Considering the fact that there exist roughly two optimal dilutions (either at $\lambda_O = 0$ or at $\lambda_O = 2$), we suggest to perform validation experiments with only two dilutions: a blank dilution ($\lambda_1 = 0$) and additionally a dilution with a bacterial density around two ($\lambda_2 = \lambda \approx 2$). For this two dilution experiment, we can estimate the false positive rate η_h and the product of the detection proportion and the bacterial density $\xi_h = \theta_h\lambda$ with ML. The ML estimators are given by

$$\hat{\eta}_h = \hat{\mu}_{h1}, \qquad \hat{\xi}_h = \log(1 - \hat{\mu}_{h1}) - \log(1 - \hat{\mu}_{h2}). \tag{11}$$

with $\hat{\mu}_{h1}$ the estimated proportion of positive test results at the blank dilution for method h and $\hat{\mu}_{h2}$ the estimated proportion of positive tests results at the non-blank dilution. The estimator $\hat{\xi}_h$ in (11) can be viewed as a *generalized MPN estimator*. It has two differences compared with the original MPN. First, there is an additional first term, which can be viewed as a downward correction in case there is a false positive rate. Indeed, if the false positive rate is zero, then the proportion of positive results in the blank dilution would equal zero, and the first term would disappear. Secondly, it is not the bacterial density itself that is estimated, but the product of the detection proportion and the density ($\theta_h\lambda$ for method h). If the detection proportion would be one, which corresponds to the assumption in [3], we get back the original MPN

estimator for a single non-blank dilution. Note that the estimator does not exist if all samples are tested positively in either the blank or the non-blank dilution or in both.

Approximate 100(1 − α)% confidence intervals for proportions or their differences have been discussed in literature extensively [18,19]. It is well-known that the simple large sample approximation suffers from aberrations, in particular for proportions close to the boundary of zero or one. Instead, Wilson's approach for proportions or differences in proportions can be used, which is computationally simple and performs better [18,19]. Thus, we propose to use Wilson's approach for confidence intervals on η_h and on $\eta_1 - \eta_2$.

Table II. Simulated coverage probabilities, type I error rates and powers related to the false positive rates for the experiment with two dilutions (blank and $\lambda = 2$).

					Powers (%)	
		Coverages (%) of CIs			Based on CI	LRT
η_1, η_2	n	η_1	η_2	$\eta_1 - \eta_2$	$\eta_1 - \eta_2$	LRT_1
0.005, 0.005	150	95.5	95.5	99.7	0.3	5.2
	200	90.8	92.1	99.6	0.4	5.4
	250	96.2	96.5	99.7	0.3	8.6
0.03, 0.005	150	94.9	96.4	96.9	29.3	53.8
	200	94.5	91.8	97.0	43.7	60.5
	250	94.6	96.5	97.5	59.5	71.0
0.03, 0.03	150	95.2	95.7	97.0	3.0	5.8
	200	94.2	95.1	96.0	4.0	6.2
	250	93.8	94.3	96.3	3.7	4.4

Table III. Simulated coverage probabilities and type I error rates related to the ratio of detection proportions for the experiment with two dilutions (blank and $\lambda = 2$).

				Type I error rates (%)		
			Coverages (%) of CIs	Based on CI	LRTs	
η_1, η_2	θ_1, θ_2	n	θ_1/θ_2	θ_1/θ_2	LRT_θ	LRT_2
0.005, 0.005	1, 1	150	94.6	5.4	5.6	5.6
		200	94.8	5.2	5.4	5.1
		250	95.4	4.6	4.6	4.6
0.03, 0.005	1, 1	150	95.3	4.7	4.8	4.4
		200	93.5	6.5	6.5	6.4
		250	93.6	6.4	6.5	6.3
0.03, 0.03	1, 1	150	96.3	3.7	3.8	3.5
		200	95.3	4.7	4.7	5.3
		250	95.5	4.5	4.5	4.4
0.005, 0.005	0.7, 0.7	150	96.6	3.4	3.4	3.2
		200	94.0	6.0	6.1	6.3
		250	95.6	4.4	4.4	4.4
0.03, 0.005	0.7, 0.7	150	95.5	4.5	4.5	4.3
		200	94.3	5.7	5.7	6.3
		250	95.7	4.3	4.3	5.0
0.03, 0.03	0.7, 0.7	150	95.5	4.5	4.5	4.0
		200	94.3	5.7	5.8	5.8
		250	94.5	5.5	5.6	5.1

Approximate confidence limits for ξ_h are derived with the delta method on the estimate $\log \hat{\xi}_h$. The variance τ_h^2 of $\hat{\xi}_h$ is approximated by

$$\tau_h^2 = \frac{\mu_{h1}}{(1-\mu_{h1})n_{h1}} + \frac{\mu_{h2}}{(1-\mu_{h2})n_{h2}}, \qquad (12)$$

where the covariance term vanishes due to independence of the test samples from the different dilutions. An estimate $\hat{\tau}_h$ for this variance is straightforward by substituting the estimates $\hat{\mu}_{h1}$ and $\hat{\mu}_{h2}$ into (12). The approximate variance of $\log \hat{\xi}_h$, using the delta method, is τ_h^2/ξ_h^2 and it can be estimated by $\hat{\tau}_h^2/\hat{\xi}_h^2$. Hence, the confidence limits for ξ_h can now be approximated by $\exp\left(\log \hat{\xi}_h \pm z_{1-\alpha/2}\hat{\tau}_h/\hat{\xi}_h\right)$, with $z_1 - \alpha/2$ the 100(1-alpha/2)th percentile of the standard normal distribution. Using the asymptotic normality of $\log \hat{\xi}_h$ and then transforming it back to the original scale has the advantage that it avoids lower confidence limits, which are negative [20]. Finally, a confidence interval for the ratio θ_1/θ_2 of the detection proportions is derived from the estimator $\log\left(\hat{\xi}_1/\hat{\xi}_2\right)$ and the delta method. The confidence limits are given by

$$\exp\left(\log\left(\hat{\xi}_1/\hat{\xi}_2\right) \pm z_{1-\alpha/2}\sqrt{\hat{\tau}_1^2/\hat{\xi}_1^2 + \hat{\tau}_2^2/\hat{\xi}_2^2}\right). \qquad (13)$$

Note that the estimator $\log\left(\hat{\xi}_1/\hat{\xi}_2\right)$ is related to the estimator of the log relative potency in bioassays, when the false positive rates η_1 and η_2 are identical. Indeed, if $\eta_1 = \eta_2$, then the detection proportions for the alternative and compendial methods in (6) satisfy the assumption of *similarity* in bioassays [21], that is, $\mu_1(\lambda) = \mu_2(\lambda\theta_1/\theta_2)$, with $\mu_h(x) = 1 - (1-\eta_h)\exp(-\theta_h x)$, λ any concentration, and θ_1/θ_2 the true relative potency. If the false positive rates are unequal, similarity does not hold anymore and the potency estimate must depend on the false positive rates (as it does in our case). In bioassays, similarity is an important aspect [22], because it would indicate that the tested product behaves as a dilution of the reference standard to which it is compared. However, in microbiological method validation, we do not just test for similarity, because this would focus only on equality of the false positive rates. For validation, we are much more interested in the detection proportions, even if the false positive rates might be different, because this part would truly demonstrate whether organisms in a test sample are better detected with the alternative method than with the pharmacopoeial method. The confidence interval for the ratio of detection proportions in (13) can then be used to compare the performances of the two methods, whether similarity ($\eta_1 = \eta_2$) would hold or not.

The confidence intervals for the difference between false positive rates and for the ratio of detection proportions can of course also be used to test null hypotheses $\eta_1 = \eta_2$ and $\theta_1 = \theta_2$. Alternatively, the likelihood ratio tests can be applied too. For the null hypothesis on the false positive rates, the likelihood ratio test LRT_1 in (8) can be applied, whereas LRT_2 in (8) would test the combined null hypothesis $H_0 : \theta_1 = \theta_2 \wedge \eta_1 = \eta_2$ and does not specify whether rejection of this null hypothesis is caused by a difference in detection proportions, in false positives, or in both. A specific likelihood ratio test, say LRT_θ, for the null hypothesis $H_0 : \theta_1 = \theta_2$, which would be the most important part of the validation study, can easily be formulated too, but it does not have a closed-form expression.

4. SIMULATIONS

Simulations have been carried out in order to evaluate the performance of the confidence intervals and the likelihood ratio tests for the proposed experiment with a blank dilution and a dilution with a bacterial density of about two. Coverage probabilities as well as type I errors and powers were evaluated. A variety of parameter settings was considered, but different settings did not provide additional insights. We provide only the results for detection proportions of 0.7 and 1, and false positive rates equal to 0.005 and 0.03. Per parameter setting we performed 1000 simulations.

Table IV. Simulated coverage probabilities and powers related to the ratio of detection proportions for the experiment with two dilutions (blank and $\lambda = 2$).

				Powers (%)		
			Coverages (%) of CIs	Based on CI	LRTs	
η_1, η_2	θ_1, θ_2	n	θ_1/θ_2	θ_1/θ_2	LRT_θ	LRT_2
0.005, 0.005	0.7, 1	150	96.0	68.7	68.7	69.6
		200	95.0	82.7	82.8	82.4
		250	95.6	87.6	87.6	87.5
0.03, 0.005	0.7, 1	150	95.3	68.9	69.1	65.4
		200	95.2	79.9	80.0	75.9
		250	95.6	87.0	87.0	83.3
0.005, 0.03	0.7, 1	150	94.8	69.1	69.1	72.2
		200	94.7	81.6	81.5	83.9
		250	95.8	88.3	88.3	90.4
0.03, 0.03	0.7, 1	150	95.9	67.6	67.6	68.7
		200	95.7	80.1	80.2	79.7
		250	93.6	88.7	88.7	88.5

Table II presents the results for the false positive rates. The coverage probabilities of the Wilson confidence intervals without continuity correction [18] for η_1 and η_2 are close to the nominal level of 95%. The coverage probabilities of the Wilson confidence interval (without continuity correction) for $\eta_1 - \eta_2$ are generally higher than 95%. This is in line with the results in [19] for differences in proportions below 0.05. Along with this, the power for testing equality of false positive rates based on the confidence interval for the difference, is generally lower than the power of LRT_1. The conservativeness of the confidence interval-based test is also reflected in the type I errors that are below 5%. The relatively high power of LRT_1 is accompanied by a type I error that is sometimes slightly higher than the nominal value of $\alpha = 0.05$, but it is still acceptable.

For the ratio of the detection proportions θ_1/θ_2, the coverage probabilities of the approximate 95% confidence intervals are well around their nominal value, see Tables III and IV. Correspondingly, in Table III the type I error rates for the confidence interval-based test are around 5%. Also the type I error rates of the likelihood ratio tests are close to 5%. Apparently, the difference in false positive rates between the alternative and compendial method hardly increases the type I error rate with LRT_2. This is not surprising, since the expected proportions of positive test results in (6) are not much affected by the low false positive rates when $\lambda = 2$ and the detection proportions are between 0.7 and 1.

Comparing the power of LRT_θ with the power of LRT_2, Table IV demonstrates that in cases where a higher false positive rate (partly) compensates for a lower detection proportion, LRT_2 has more difficulties in detecting the difference between the two methods. This is due to the fact that LRT_2 tests the null hypothesis $H_0 : \mu_{12} = \mu_{22}$, while LRT_θ tests the null hypothesis $H_0 : \theta_1 = \theta_2$. Conversely, in cases where a higher false positive rate accompanies a higher detection proportion, LRT_2 has a higher power than LRT_θ. In case of equal false positive rates, the powers of LRT_2 and LRT_θ are similar but not identical, since they still test different null hypotheses. Furthermore, Table IV demonstrates that the power of LRT_θ is similar to the power of the confidence interval-based test for $H_0 : \theta_1 = \theta_2$, which is simpler to calculate and more informative.

5. CONCLUSIONS AND DISCUSSION

Based on a stochastic formulation of the detection mechanism $\pi_h(y)$ of qualitative test method h, the relevant validation parameters are specificity and accuracy. Specificity is clearly related to the false positive rate η_h of method h, because extraneous matter in the test system should not interfere with the test method. The accuracy is related to the ratio θ_1/θ_2 of detection proportions θ_1 and θ_2 for the alternative and compendial method, respectively. The detection proportions indicate the probability of correctly detecting one micro-organism. To validate the alternative method, a comparison with the compendial method is needed using an experiment with two dilutions. One dilution should contain approximately two CFUs per test unit and the other dilution is a blank. To test whether the specificity differs between the two methods, a likelihood ratio test is recommended (which is not difficult to calculate). To give insight in the size of the false positive rates for the two methods, the Wilson confidence intervals can be reported. On the other hand, for the accuracy, we would recommend a confidence interval-based test. It has a similar power to the likelihood ratio test for testing the null hypothesis $H_0 : \theta_1 = \theta_2$, but it is simpler to calculate and it is more informative. For a significance level of $\alpha = 0.05$ and a power of 80%, approximately 200 test samples per dilution per test method are required to find an accuracy of at most 70% for the alternative method relative to the compendial method.

Our proposed strategy deviates from the current regulatory guidances at several points. First of all, the guidances request more validation parameters to be validated. Assuming our stochastic detection mechanisms, other validation parameters will not give additional insight in differences between the test methods. Although we assumed a (zero-deflated) binomial detection mechanism, we believe that our strategy remains valid whenever $\pi_1(y)$ is systematically larger or smaller than $\pi_2(y)$ for $y \geqslant 1$, irrespective of the number of parameters involved or the shape of the detection mechanisms. Our dilution experiment may not necessarily be optimal anymore, but it will always provide information on differences between the detection mechanisms. Secondly, the suggested validation experiments in the guidances with five ten-fold dilutions will most likely not result in an optimal experiment for testing differences between detection mechanisms. When $\pi_1(y)$ is ordered with respect to $\pi_2(y)$ for $y \geqslant 1$, there will be a single dilution that maximizes the difference between the detection mechanisms. Consequently, the suggested approach to compare the MPN between the alternative and the compendial method using routine experiments will also lose power compared with our two dilution approach. Moreover, our stochastic formulation of the detection mechanisms has changed the interpretation of the MPN and demonstrates its lack of suitability for validation purposes, particularly in the presence of false positives. Finally, the guidances underestimate the number of test samples needed to detect relevant differences in specificity and accuracy. A low number of test samples introduces the risk of approving an alternative test method with a detection proportion that is substantially less than 70% of the detection proportion of the compendial method.

The USP <1223> is currently under revision, but it seems that their view on the choice of validation parameters did not change yet. However, they did improve the statistical parts, eliminated a few unnecessary comparisons (like the one on agreement between the alternate and compendial method), which is in line with our view, and increased sample sizes for validation. The most important change in our opinion is a shift in scope towards equivalence testing. In terms of our models for detection of micro-organisms, their approach would only be suitable when the alternate and compendial method do not provide or have hardly any false positives. Our focus here has been on traditional hypothesis testing, but our proposed methods for construction of confidence intervals on the ratio of detection proportions and on the difference in false positive rates make it possible also to perform equivalence testing on accuracy and specificity separately.

Our strategy is also different from the approach by [12] and [7]. They imposed the deterministic detection mechanism and assumed no false positives. Under their assumptions, they only required an estimate for the limit of detection of the alternative method without any comparison with the compendial method. However, the limit of detection quantifies the detection performance after all preparation steps, thus they needed also an estimate of the recovery of the alternative method with respect to the compendial enumeration test to investigate if organisms are lost during testing. This approach would also be acceptable when the detection mechanism is of the binomial form, because the recovery estimate would be related to the accuracy θ_1/θ_2 (see our discussion in Section 2.2). However, when false positives

would occur, our approach for validation with the zero-deflated binomial mechanism would be more appropriate and a direct comparison with the compendial qualitative test method should be conducted.

For the validation of analytical limit tests, both specificity and limit of detection should be determined (ICH Q2(R1), [4]). These estimates can be obtained without making any comparisons to an existing method and are therefore different from our approach. However, we do believe that our detection mechanisms may also be valuable for limit tests, because they provide better insight in how the limit tests would detect low levels of quantities. The advantage of validation of limit tests is that spiking low levels of quantities or concentrations is not a real problem, contrary to microbiological method validation. This means that the detection proportion can be directly estimated.

The binomial mechanism was suggested and applied by [8], who used it for enumeration tests. Their goodness-of-fit test on a real case study did not demonstrate the need for an alternative detection mechanism, indicating that their assumptions were reasonable. Their binomial detection mechanism for enumeration tests would lead to our binomial detection mechanism for qualitative tests. Thus, our choice of a zero-deflated binomial detection mechanism may also be considered realistic, but more research is needed on possible other shapes for the detection mechanisms of qualitative test methods. One particular question is what shape of detection mechanism will lead to a logistic curve for the expected proportion of positive test results. This is important, because logistic regression is frequently applied for validation, but not suitable under our assumed detection mechanisms. Additionally, more research should be conducted on goodness-of-fit tests, in particular on the selection of optimal designs, because goodness-of-fit tests require more than our proposed two dilutions to be able to test the validity of our zero-deflated detection mechanism. Introducing additional dilutions for goodness-of-fit would then diminish the performance of the proposed hypothesis tests on specificity and accuracy.

REFERENCES

[1] European Pharmacopoeia 8.0. Alternative methods for control of microbiological quality. EDQM: Strasbourg, 2008; Chapter 5.1.6, pp. 560–570.

[2] United States Pharmacopoeia 36. <1223> Validation of alternative microbiological methods. U.S. Pharmacopoeial Convention: Rockville, MD, 2008, pp. 979–982.

[3] Cochran WG. Estimation of bacterial densities by means of the 'Most Probable Number'. *Biometrics* 1950; **6**(2):105–116.

[4] ICH Harmonised Tripartite Guideline. Validation of analytical procedures: text and methodology Q2(R1), 2005.

[5] Abbott WS. A method of computing the effectiveness of an insecticide. *Journal of Economic Entomology* 1925; **18**:265–267.

[6] Ridout MS, Fenlon JS, Hughes PR. A generalized one-hit model for bioassays of insect viruses. *Biometrics* 1993; **49**:1136–1141.

[7] Van den Heuvel E. Estimation of the limit of detection for quantal bioassays. *Pharmaceutical Statistics* 2011; **10** (3):203–212. **DOI:** 10.1002/pst.435

[8] Van den Heuvel ER, IJzerman-Boon PC. A comparison of test statistics for the recovery of rapid growth-based enumeration tests. *Pharmaceutical Statistics* 2013; **12**(5):291–299. **DOI:** 10.1002/pst.1581

[9] Wyshak G, Detre K. Estimating the number of organisms in quantal assays. *Applied Microbiology* 1972; **23**(4):784–790.

[10] Morgan CA, Bigeni P, Herman N, Gauci M, White PA, Vesey G. Production of precise microbiology standards using cytometry and freeze drying. *Cytometry, Part A* 2004; **62A**:162–168.

[11] Van den Heuvel ER, Verdonk GPHT, IJzerman-Boon PC. Statistical methods for detection of organisms with sterility tests. In *Rapid sterility tests*. Moldenhauer, J (ed.). Parenteral Drug Association, Davis Healthcare International Publications: USA, River Grove, 2011; pp. 201–243.

[12] Verdonk GPHT, Willemse MJ, Hoefs SGG, Cremers G, Van den Heuvel ER. The Most Probable Limit of Detection (MPL) for rapid microbiology. *Journal of Microbiological Methods* 2010; **82**:193–197. **DOI:** 10.1016/j.mimet.2010.04.012

[13] Chrzanowski TH, Smith RL. Validation of an amplified-ATP bioluminescence method for the rapid detection of contamination in a betamethasone suspension. *American Pharmaceutical Review* 2012; **15**(7):1–4.

[14] Hosmer DW, Lemeshow S. *Applied logistic regression* (2nd edn). Wiley: New York, 2000.

[15] Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 1938; **9**(1):60–62.

[16] Strijbosch LWG, Does RJMM, Albers W. Multiple-dose design and bias-reducing methods for limiting dilution assays. *Statistica Neerlandica* 1990; **44**(4):241–261.

[17] Fisher RA. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* 1922; **222**:309–368. **DOI:** 10.1098/rsta.1922.0009

[18] Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 1998a; **17**:857–872.

[19] Newcombe RG. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* 1998b; **17**:873–890.

[20] Ridout MS. A comparison of confidence interval methods for dilution series experiments. *Biometrics* 1994; **50**:289–296.

[21] Finney DJ. *Statistical method in biological assay* (3rd edn). Charles Griffin & Company Ltd.: London, 1978.

[22] Liao JJ, Tian Y, Capen RC. Assessing similarity in bioanalytical methods. *PDA Journal of Pharmaceutical Science and Technology* 2011; **65**(1):55–62.