CODEX ALIMENTARIUS COMMISSION

E

Food and Agriculture Organization of the United Nations

World Health Organization

Viale delle Terme di Caracalla, 00153 Rome, Italy - Tel: (+39) 06 57051 - Fax: (+39) 06 5705 4593 - E-mail: codex@fao.org - www.codexalimentarius.org

Agenda Item 5 **CX/MAS 15/36/5**

JOINT FAO/WHO FOOD STANDARDS PROGRAMME

CODEX COMMITTEE ON METHODS OF ANALYSIS AND SAMPLING

Thirty-sixth Session

Budapest, Hungary, 23 – 27 February 2015

DISCUSSION PAPER ON DEVELOPMENT OF PROCEDURES/GUIDELINES FOR DETERMINING EQUIVALENCY TO TYPE I METHODS

(prepared by the United States of America)

BACKGROUND

1. At its 35th session, the Codex Committee on Methods of Analysis and Sampling (CCMAS) considered the "*Discussion Paper on Considering Procedures for Establishing Criteria for Multi-analyte and Type I Methods*".

2. As part of those discussions CCMAS agreed that numerical criteria for Type I methods should not be established, however that is might be useful to consider and discuss procedures for establishing equivalency to Type I methods (Para 59 of REP14/MAS).

3. Based on that decision an electronic working group, chaired by the United States of America and working in English was established to develop a discussion paper which would consider different procedures/guidelines for determining equivalency to Type I methods (Para 61 of REP14/MAS).

4. There was considerable interest in the eWG with 20 participants. However, due to other duties, the United States of America was not able to complete a draft discussion paper in time to allow for discussion among the participants of the eWG. Therefore this discussion paper is presented as a first draft, with request for comment and further discussion to proceed during the 36th Session of CCMAS.

INTRODUCTION

5. The Procedural Manual defines a Type I method as "*A method which determines a value that can only be arrived at in terms of the method per se and serves by definition as the only method for establishing the accepted value of the item measured."* (21st Ed. 2013, English Version, p 63). Based on this definition, a Type I method that has been endorsed by Codex is expected to produce the true value for the applicable measurand and would serve as the benchmark for determining the trueness of any alternative to the Type I method.

6. The purpose of this paper is to discuss possible procedures for establishing equivalence to an existing Type I method.

7. General criteria for Codex methods of analysis have been specified and include the requirement for an inter-laboratory or single laboratory validation to confirm that the method is 'fit for purpose.' This means that a rigorous process has been applied to evaluate the performance characteristics of the method using appropriate samples that define the scope of the method, preferably in multiple laboratories. These same principles should apply to any method that is being considered as equivalent to a Type I method. However, in addition to meeting all of the general criteria as specified in the Procedural Manual, this "alternative" method should also produce results for each sample that are equivalent to the existing Type I method. The question of equivalence is not only if the alternative method is fit for purpose; but also, does the proposed alternative method produce results that are equivalent to the endorsed Type I method?

8. In order to evaluate the equivalence of a proposed alternative to a Type I method, it will be necessary to run one or more sets of samples using both methods and then to compare the results. For this reason, a prerequisite is that the two methods produce results for the same measurand in the same units. Samples run by each method should be homogeneous and cover each representative

matrix and concentration range defined in the scope of the alternative method. One sample set should be run for each matrix and concentration level that is necessary to demonstrate equivalency for the new method scope. It is necessary that each and every sample set pass an appropriate test for equivalence in order for the alternative method to be considered truly equivalent.

9. Unfortunately, there is little guidance from regulatory agencies or scientific associations such as AOAC or ISO on the exact procedures for establishing the equivalence of analytical methods. In recent years, the pharmaceutical industry has been required to establish procedures for evaluating method equivalence based on regulations for the bioequivalence of orally administered drug products. As a result, the issue of method equivalency has been thoroughly discussed in this context and the discussion of procedural options that follows is taken largely from a review of these papers.

STATISTICAL APPROACHES

Option 1: Two-sample *t*-Test

10. One option evaluating method equivalency is to compare the mean values of the data sets for each method by doing a test for statistical significance using the two-sample *t*-test.[1, 11] This is a statistical hypothesis test in which the null hypothesis or initial assumption is that the means for the methods are equal. The test attempts to disprove this hypothesis, and if successful, proves to a specified confidence level (usually 95%) that the means are significantly different. If the test is unsuccessful, showing that there is not a statistical difference in the means between the data sets, the default is to assume that the means are equal. The test assumes that both data sets contain results that are normally distributed, that the standard deviations are similar and that the Student's t distribution describes the random variation for sample sizes less than $n = 30$. To conduct this test, the value of the *t* statistic is calculated by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad (1)$$

where $\bar{x}_1$and $\bar{x}_2$are the mean values from each sample set for each method, s_p is the estimate for the pooled standard deviation for each the sample sets, and $n_1$and n_2 are the number of samples run in each set. s_p for the replicate sets of measurements is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \qquad (2)$$

where $s_1$and s_2 are the standard deviation values for each sample set. The *t*-value from equation (1) is calculated and compared to the critical *t*-value for a specific confidence level (usually 95%). This value can be found in statistical tables or by using the TINV function in Microsoft Excel. If the absolute value of the calculated *t*-value is greater than or equal to the critical t-value, the two sets of data are declared statistically different.

11. A second method for accomplishing the same test is to calculate the confidence interval (CI) for the difference between the two mean values using

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{95,(n_1+n_2-2)} \cdot \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad (3)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the mean values from each sample set for each method, $t_{95,(n_1+n_2-2)}$ is the t-value at 95% confidence with n_1+n_2-2 degrees of freedom, s_p is the estimate for the pooled standard deviation for the sample sets as given in equation (2), and $n_1$and n_2 are the number of samples run in each set. The value for $t_{95,(n_1+n_2-2)}$ can be found in statistical tables or by using the TINV function in Excel. If CI is does not include 0, then the means are declared significantly different.

12. It has been pointed out that the two-sample *t*-test may not be appropriate to demonstrate equivalency between two methods, since the test is designed to prove that two sets of data have different means, not that the two sets have equivalent means. [6, 8, 9, 10, 11] Since the null or default hypothesis is that the means are equal, the method provides supporting evidence only in the case

that the two means are unequal. When there is insufficient evidence to prove a significant difference in the means the default is to accept the null hypothesis, but this does not prove that the means are actually equal. Two specific problems have been identified with the application of the two-sample *t*-test for evaluating method equivalence. First, when the standard deviations of the two data sets are relatively high and the number of measurements is relatively low, the *t*-test can result in no significant difference when the inclusion of additional measurements would have demonstrated that the means were actually statistically different. Second, when the standard deviations for the two sets are relatively small, the test can result in the finding of a statistical difference when the absolute difference is not practically important.

Option 2: Two one-sided *t*-Test (TOST)

13. A statistical test called the two one-side t-test (TOST) has been proposed as the recommended alternative to the two sample *t*-test and begins with the opposite null hypothesis; that the two mean values for the methods are not equivalent.[6, 7, 11] A positive test for significance then results in demonstrating, at a specified confidence level, that the two data sets are equivalent. The TOST test requires the specification of a parameter called the acceptance criterion, $\pm\theta$, which represents the smallest difference in mean values for the two methods that is deemed as practically important. The null hypothesis, H_0, and the alternate, H_a, are described in terms of the difference in means, $\mu_1 - \mu_2$, and θ by the following

$$H_0: \mu_1 - \mu_2 \leq \theta_L \; or \; \mu_1 - \mu_2 \geq \theta_U$$

$$H_a: \theta_L < \mu_1 - \mu_2 < \theta_U$$

14. The alternative hypothesis is proven at a specified level of confidence when the true difference in means between the methods is within the boundaries specified by $\pm\theta$.

15. TOST is carried out by first specifying θ and then the confidence interval (CI) for the difference in means at a specific level of confidence (usually 95%) is calculated by

$$CI = \bar{x}_1 - \bar{x}_2 \pm t_{90,(n_1+n_2-2)} \cdot \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad (4)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the mean values from each sample set for each method, $t_{90,(n_1+n_2-2)}$ is the t-value at 90% confidence with $n_1 + n_2 - 2$ degrees of freedom, s_p is the estimate for the pooled standard deviation for the sample sets as given in equation (2) and $n_1$and n_2 are the number of samples run in each set. The value for $t_{90,(n_1+n_2-2)}$ can be found in statistical tables or by using the TINV function in Excel. (Note that the 95% one-sided confidence limit is equivalent to the 90% two-sided confidence limit). If CI is completely contained within the range defined by $\pm\theta$, the methods as defined by this data set are deemed equivalent.

16. Both of the problems related to the two-sample *t*-test are solved by using TOST due to the specification of the acceptance criterion, θ. The situation with the identification of a statistical difference in means that is not practically important is solved since the specification of the range ($\pm\theta$) allows for a statistically significant bias as long as it is sufficiently small to be contained within the range. The situation where the method s values are relatively large is also accounted for, since method equivalence will not be established unless the calculated CI for the two methods is contained completely within the range defined by $\pm\theta$.

17. This also leads to what might be perceived as a disadvantage of TOST; the need to specify the acceptance criterion. It may be difficult to establish a single acceptance criterion for all methods, since it requires the identification of the smallest practically significant difference in the mean values and this may require an individual assessment for each method.

18. Obtaining the adequate power or confidence in the conclusion for a statistical significance test requires running a sufficient number of samples. The number of samples needed is dependent on the ratio of θ to s_p with larger ratios requiring a smaller number of samples. When θ and s_p are similar in value, 18-27 samples are required to obtain sufficient power for the equivalence test.[9]

Option 3: Limits of Agreement

19. A third option for evaluating equivalence is an approach based on the calculation of the limits of agreement (LOA) around the mean of the endorsed Type I method.[2, 3, 4, 5] For each data set that is run, the calculated LOA represents the interval around the mean in which 95% of the results would be expected to lie. As with the other procedures, this assumes a normal distribution and a sufficient number of samples to give reasonably good estimates for the mean and standard deviation values for each method.

$$LOA = u \pm z_{95} \cdot \sigma = 1.96 \cdot \sigma$$

where u is the mean, z_{95} is the 95% percentile of the normal distribution, and σ is the standard deviation for the data set for the endorsed Type I method.

20. This LOA would then become the range to which the individual values for the alternative method would be compared. If 95% of these values lie within the limits of agreement, the alternative method would be deemed equivalent for that particular data set. To provide sufficient confidence for this particular test, it is recommended that a minimum of 20 samples be run for each data set since 19/20 within this range would constitute 95% of the results. The *LOA* method would allow for bias between the methods, but only if it is small enough so that 95% of the individual results lie within the LOA defined by the endorsed Type I method.

SUMMARY

21. The advantage of the two-sample *t*-test is that it does not require any additional parameters such as an acceptance criterion. However, the *t*-test is not the appropriate statistical test to use when the desire is to demonstrate equivalence. In particular, the use of the *t*-test could lead to the conclusion that the methods are statistically different when the absolute difference is not practically important. The test can also lead to the erroneous conclusion that two methods are equivalent when there were insufficient samples run to demonstrate the actual statistical difference.

22. The TOST test for statistical equivalence is probably the most rigorous approach, but it does require the specification of a practical acceptance limit that represents the smallest bias between the methods that can be accepted. In practice, it may be difficult to establish one set of acceptance criteria for all Codex methods, since each has a specific application. One option could be to use the recovery factors specified in the Codex Procedural Manual under the section on "guidelines for establishing numeric values for criteria" using the appropriate concentration. For example, the specified recovery of 95-105% for concentrations of ≥0.1 suggests an acceptance criterion of ±5%. However, many methods will not have an applicable recovery factor. This would require the Codex commodity committee that is recommending the method to consider the appropriate minimum acceptable bias on a case-by-case basis.

23. The LOA approach that involves calculating a 95% probability interval around the mean of the existing Type I method appears to be a possible option. This option allows for bias only if the alternative method also yields a smaller standard deviation so that 95% of the individual results still fall within the range defined by the LOA. The advantage of this approach would be that it is very simple to calculate and it does not require acceptance criteria to be established. A disadvantage would be that it would not allow for any bias in the alternative method if the standard deviation of the alternative method is equal to that of the endorsed method. The comparison of multiple data sets representing different matrices and concentrations may make it difficult for any alternative method to demonstrate equivalence using this procedure.

24. Each of these possible procedures would require the analysis of multiple sample sets that include the appropriate matrix and concentration levels to cover the scope of the method. It is suggested that at least two concentrations be run for each relevant matrix. One should be at the ML (minimum or maximum level) and the other at the limit of quantitation. If the method does not have a specified ML, a low and high concentration should be chosen that covers the concentration range of the method. For the TOST procedure, the number of samples to be run in order to demonstrate equivalence varies with the standard deviation and the acceptance criterion, but a conservative estimate of 18-27 samples for each data set would cover situations when θ is similar in magnitude to s_p. At least 20 samples are recommended for the LOA approach in order to obtain a reasonably good estimate of the standard deviation. The equivalence studies should be performed under reproducibility

conditions as much as possible. The studies could be performed in a single laboratory, but it would be desirable to utilize different analysts, materials, and equipment as much as possible.

QUESTIONS FOR DISCUSSION

25. Although the paper generally describes a number of statistical approaches for establishing equivalency between methods it clearly does not address specific details in implementing any of these methods. If a general approach can be recommended, there are still outstanding questions as to the application of that approach to specific methods. For instance:

i. Based on the discussion and the range of methods is it practical to establish one set of equivalence criteria for all Codex Methods?

ii. If such criteria or even general procedures for evaluating equivalence were established where would they reside in Codex, as part of the Procedural Manual or in a Guidance document?

References

1. Freund, J. E.; *Modern Elementary Statistics*, 7th ed.; Prentice-Hall: Englewood Cliffs, New Jersey, 1988; p 313.
2. Zhong, J.; Lee, K.; Tsong, Y.; *J. Biopharm. Stat*. **2008**, *18*, 1005-1012.
3. Bland, D. G.; Altman, J. M. *Lancet* **1986,** *8,* 255.
4. Bland, D. G.; Altman, J. M.; *Stat. Methods Med. Res.* **1999,** *8,* 137-160.
5. Dewe, W.; *J. Chrom. B* **2009**, *877*, 2208-2213.
6. Hartmann, C.; Smeyers-Verbeke, J.; Penninckx, W.; Vander Heyden, Y.; Vankeerberghen, P.; Massart, D. L. *Anal. Chem*. **1995**, *67*, 4491-4499.
7. Schuirmann, D.; *J. Pharmaco. Biopharm*. **1987**, *15*, 657-680.
8. Chatfield, M. J.; Borman, P. J. *Anal. Chem.* **2009**, *81*, 9841-9848.
9. Borman, P. J.; Chatfield, M. J.; Damjanov, I.; Jackson, P. *Anal. Chem.* **2009**, *81*, 9849-9857.
10. Chambers, D.; Kelly, G.; Limentani, G.; Lister, A.; Lung, K. R.; Warner, E. *Pharm. Tech*. **2005**, *29*, 64-80.
11. Limentani, G. B.; Ringo; M. C., Ye, F.; Bergquist, M. L.; McSorley, E. O. *Anal. Chem.* **2005**, *77*, 221A-226A.