The logic of equivalence testing and its use in laboratory medicine

Cristiano Ialongo*[1,2]

[1]Department of Human Physiology and Pharmacology, University of Rome Sapienza, Rome, Italy
[2]Laboratory Medicine Department, "Tor Vergata" University Hospital, Rome, Italy

Corresponding author: cristiano.ialongo@gmail.com

Abstract

Hypothesis testing is a methodological paradigm widely popularized outside the field of pure statistics, and nowadays more or less familiar to the largest part of biomedical researchers. Conversely, the equivalence testing is still somehow obscure and misunderstood, although it represents a conceptual mainstay for some biomedical fields like pharmacology. In order to appreciate the way it could suit laboratory medicine, it is necessary to understand the philosophy behind it, and in turn how it stemmed and differentiated along the history of classical hypothesis testing. Here we present the framework of equivalence testing, the various tests used to assess equivalence and discuss their applicability to laboratory medicine research and issues.

Key words: biostatistics; statistical data analysis; methodological studies

Received: May 20, 2016 *Accepted: October 12, 2016*

Introduction – historical genesis of difference testing

Chance is part of nature, and whatever we did there is always a probabilistic aspect we necessarily should deal with. This was utterly clear to researchers at the beginning of 20th century, which were involved in the positivistic approach to natural sciences, especially agriculture (1). In this field, they had the opportunity to manipulate nature to produce changes, but at the same time they had the necessity to recognize whether what they observed was the consequence of an intervention or just a mere fluke (2). Indeed, they already knew that nature was variable, and Galton had shown that biological traits followed the laws of probability (3,4). Therefore, researchers had the necessity to turn observations into evidences, and the laws of probability gave them means to assess uncertainty in findings.

It's no surprise that Ronald Fisher, at the time he set the hypothesis testing, was employed in an experimental agricultural station and developed a pragmatic view of statistics aimed to rule out chance from empirical evidences (5,6). In his framework, Fisher conceived the experimental research as a process attempting to induce, through the application of a factor, a certain difference in the observations taken with respect to an untreated condition*. (*NOTE: We would like to remark that this narrative description of Fisher's achievements was intentionally oversimplified. For instance, we omitted to mention that "observed differences" were instead "average difference between observations", through which it was possible to apply the Student's method of statistical comparison using probability distributions). However, what he cleverly did was approaching the proof of such an experimental factor starting from the opposite perspective of no experimental effect. Although it might seem paradoxical, placing an hypothesis of null effect gave means for representing observed differences as erratic fluctuations produced by

chance around an hypothetic value of no-difference equating zero (5). Of course, this explicitly recalled the probabilistic description of measurement errors that Gauss had given about half a century before and that was already familiar to researchers: the larger the difference from the expected outcome, the lower the probability of its random realization (7). Thus, the observation of large differences with an associated low probability was unlikely due to a random fluctuation, disproving in turn the null hypothesis. For practical reasons of experimental reproducibility, Fisher set the probability threshold for disproving randomness as low as < 0.05 (or less than 1 out of 20 trials), aiming to assure enough confidence when stating the alternative hypothesis of an experimental factor (5).

Contemporarily to the efforts of Fisher, Egon Pearson and mostly Jerzy Neyman refined the concepts of hypothesis testing (5). Neyman, which was concerned with mathematics and logic more than with experiments, formally showed that the rejection of a null hypothesis could be achieved only at the expense of a certain uncertainty regarding its truth. Such an uncertainty corresponded to the probability of erroneously rejecting a null hypothesis when there was no effect (he called Type I error), and equated the Fisher's threshold for achieving significance (or α) (Table 1). Noteworthy, Neyman also showed that the gain of confidence of correctly rejecting a null hypothesis always happened at the expense of sensitivity of detecting an actual effect, a concept he termed statistical power. Indeed, he showed that when sensitivity increases, the probability of accepting a null hypothesis when there is an effect (he called Type II error or β, so that sensitivity was 1 - β) decreased (Table 1).

The "decision making" approach of Neyman is usually seen as an alternative to the Fisher's framework, although the two should be considered rather complementary (8). Indeed, Neyman's approach helps to demonstrate that the hypothesis testing is indeed a difference testing. In fact, uncertainty is used to show whether a factor is strong enough to prove itself against chance. Thus, in the "classical framework" the burden of proof is on difference.

A trick of logic

So far, we have seen that the hypothesis testing was devised to prove an experimental hypothesis concerning a treatment or factor. Now, let us imagine that a researcher was concerned with the simple issue of replacing an old instrument with a new one. He would measure the same set of items using the old device and the new one, and then he would statistically compare the data. Could the investigator conclude the instruments were equal if he found no statistically significant difference in the compared items? Formally, he could not state this at all. In fact, the researcher observed some differences between the two sets of measures, but they were random to cause no significant changes or bias. In other words, any researcher could not recognize an instrument by the set of measures it produced, and vice versa. Therefore, he could just say that the two devices "agreed" in measuring the same objects, although he could not conclude they were literally "equal".

However, in some situations it is necessary to state explicitly that two compared items not simply agree, but are equivalent. The reader should notice at this point that we used the term "equivalent" instead of "equal", as the latter represents something that is practically unachievable due to the randomness that arises in any natural process. This is a fundamental passage, and must be carefully considered, in that "equivalent" means something that, although not strictly the same, is formally alike. Let us imagine that we compared several individuals using anthropometric measures. Supposing that they were all related, for instance brothers and cousins, reasonably there would be no statistically significant difference between them. However, an individual would be more like his brother than to his cousin, so that the issue would be showing which is the amount of difference in measures that makes any two individuals looking substantially alike. Hence, to show "equivalence" between individuals, their differences should be less than a certain amount considered a mark of significant dissimilarities. Thus, the issue would be twofold: first, we should accept the fact that missing to show a difference does not imply

Table 1. Hypothesis testing and decision-making

		DECISION	
		Accept H_0	**Reject H_0**
TRUTH	**H_0**	correct decision (1 - α or specificity)	Type I error (α)
	H_1	Type II error (β)	correct decision (1 - β or sensitivity)
		Statement	
DIFFERENCE TESTING	H_0: there is no difference	no difference and there was none	a difference but there was none
	H_1: there is a difference	no difference but there was actually one	a difference and there was one
EQUIVALENCE TESTING	H_0: there is a certain difference	a certain difference and there was one	no certain difference but there was one
	H_1: there is no certain difference	a certain difference but there was none	no certain difference and there was none

The hypothesis testing framework is shown by the decision-making standpoint. By carefully reading the statements in the lower part of the table, it is possible to understand why in difference testing the sensitivity corresponds to 1 - α (showing a difference when there is one) and not to 1 - β. The term "equivalence" means "within a certain difference", so the statement of "no certain difference" is indeed synonym of "equivalence".

equivalence, and second, we should answer the question on how close is close enough to be considered (practically) equivalent. It is evident that for answering both it is necessary a trick of logic to bend the Fisher and Neyman's framework to the needs of equivalence testing.

The rise of equivalence testing

The theoretic discussion on a possible testing procedure for equivalence is not new in statistics and was implicitly addressed by Eric Lehmann yet in 1959 (9). However, it was just around 1970s that the work of Wilfred Westlake on the statistical assessment of equivalent formulations of drugs made of it a true practical concern (10-13). He firstly recognized that whatever sufficiently large cohort of subjects would have shown as statistically significant even a negligible difference, an issue of excessive study sensitivity known as statistical overpowering. To control for β level (Table 1), Westlake devised a method relying on an interval of maximum acceptable difference between average responses of compared drugs (defined as $-\Delta$; $+\Delta$), within which the actual difference (indicated as δ) of equivalent drugs could stay. In other words, the method controlled the study sensitivity setting the largest effect size at which the actual difference was considered negligible. Hence, although not producing strictly the same response, the two drugs were considered practically interchangeable, turning out to be equivalent. The Westlake's method concluded equivalence if the confidence interval around δ (a probabilistic measure of the true location of δ) rested entirely within $-\Delta$; $+\Delta$. Noteworthy, such an approach suffered for inflation of the actual probability level (the P-value) by which the equivalence was ruled out, and flawed the application of Westlake's method and its analogues in real decision making on bioequivalence (14,15).

David Rocke was among the ones who openly recognized that the Westlake's method forced the classical framework of difference testing to control for β level, while it was devised to control for α instead (16). Hence, in 1983, he proposed to shift the burden of proof from difference to non-difference, turning upside down the perspective on the experimental hypothesis. In classical framework, equivalence was proven through non-difference as follows:

- null hypothesis (H_0) → non-difference: $|\delta| = 0$ or $|\delta| \leq |\Delta|$
- alternative hypothesis (H_1) → difference: $|\delta| > |\Delta|$.

Thus, he proposed the procedure below:

- null hypothesis (H_0) → non-equivalence: $|\delta| \geq |\Delta|$
- alternative hypothesis (H_1) → equivalence: $|\delta| < |\Delta|$.

Therefore, what was Type II error in difference testing became Type I error in equivalence, allowing to easily control for the probability of erroneously accepting an inexistent difference (now corresponding to α), overcoming the limitations of Westlake's method (Table 1).

Around the same years, Walter Hauck and Sharon Anderson advanced their procedure, which relied on the concept of "interval hypothesis" for testing equivalence as the experimental one placed under the alternative hypothesis (15). Let m_S and m_E represent the average response to a standard and experimental formulation of a drug, respectively, and A being the lower and B the upper (with B > A) boundary of the equivalence interval thereof. Then, *per* the Anderson and Hauck's procedure, it is possible to state the hypothesis testing as follows:

- null hypothesis (H_0) → non-equivalence: $m_E - m_S \leq A$ or $m_E - m_S \geq B$
- alternative hypothesis (H_1) → equivalence: $A < m_E - m_S < B$.

In fact, two sets of observations are said to be not equivalent if their average difference $m_E - m_S = \delta$ encroaches the equivalence limits, or $|\delta| \geq (B - A)$. Now, for a parallel design (two independent groups without crossing effect), it is possible to build a statistical test using the frame of a Student's two-sided t-test (Figure 1A):

$$\textbf{Eq.1.1} \quad T = \frac{\delta - 0{,}5 \times (A + B)}{S \times \sqrt{(N_E + N_S)^{-1}}},$$

in which the numerator represents the distance of the average differences form the center of the quivalence interval, N_E and N_S the size of experimental and standard group respectively, and S the pooled sample variance that in this case can be estimated as follows:

$$\textbf{Eq.1.2} \quad S_{pooled} = \sqrt{\frac{S_E^2 \times (N_E - 1) + S_S^2 \times (N_S - 1)}{N_E + N_S - 2}},$$

with S_E and S_S being the standard deviation of the experimental and standard group respectively. Then, significance for T can be found using a Student's t distribution with $\nu = N_E + N_S - 2$ degrees of freedom. Noteworthy, this method allows to properly size the study applying in the appropriate way the rules of power analysis. In fact, in the difference testing, the power refers to the uncertainty in rejecting a significant differ-

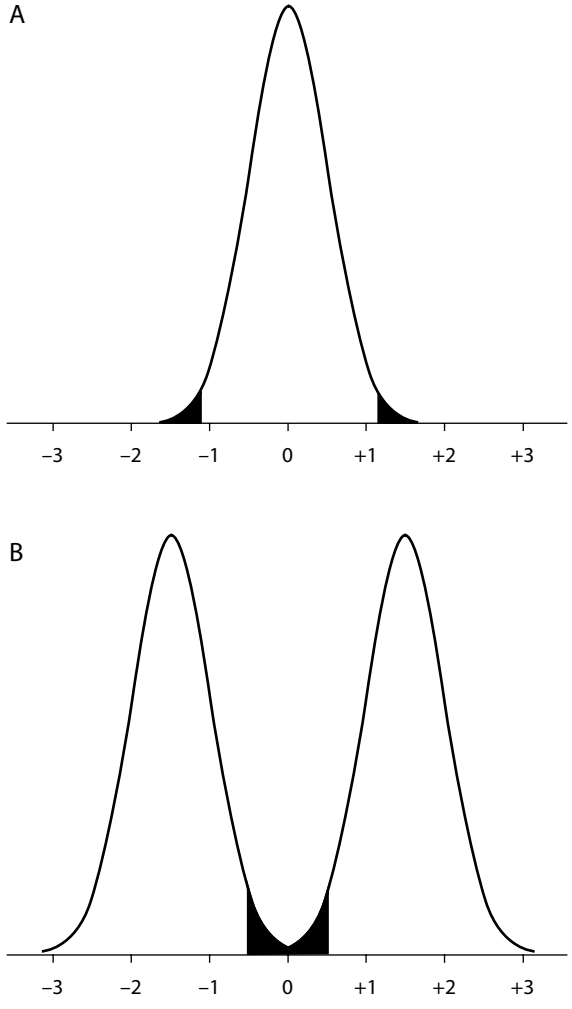

Figure 1. Rejection region (solid black area under the curve) for different procedures to test equivalence. (A) The Anderson and Hauck's two-sided test - each area accounts for α / 2, so that the confidence is 1 − α. (B) The Schuirmann's two one-sided test - each area accounts for α, so that the confidence is 1 - 2α.

ence when there is actually one, while in equivalence it is the uncertainty in rejecting equivalence when there is no difference. Anderson and Hauck also proved their procedure to be the most powerful test for equivalence comparing to any confidence interval approach previously advanced (15). In other words, they showed that given a certain interval of acceptability for equivalence, their method produced the lowest rate of false negatives in the sense of erroneously rejected equivalent items (in their case drugs formulations).

The two one-sided tests (TOST)

In 1987, Donald Schuirmann discussed the power of an alternative procedure to test equivalence, which was based on an adaptation of the Westlake's method to the null hypothesis of non-equivalence formulated in 1981 (17,18). Let us consider the equivalence interval hypothesis as presented so far:

- null hypothesis (H_0) → non-equivalence: $m_E - m_S \leq A$ or $m_E - m_S \geq B$
- alternative hypothesis (H_1) → equivalence: $A < m_E - m_S < B$.

Then, it can be rewritten "decomposing" the interval in two single hypotheses:

- null hypothesis (H_{01}) → inferiority: $m_E - m_S \leq A$
- alternative hypothesis (H_{11}) → non-inferiority: $m_E - m_S > A$,

and

- null hypothesis (H_{02}) → superiority: $m_E - m_S \geq B$
- alternative hypothesis (H_{12}) → non-superiority: $m_E - m_S < B$.

Thereby, it is possible to test both null hypotheses H_{01} and H_{02} applying to each of them a single sided test at the nominal level of significance α. Thus, to prove equivalence, it is necessary that both the hypothesis of inferiority and superiority be disproved simultaneously. Formally, the two one-sided tests (TOST) can be written as shown below for a two parallel groups design (Figure 1B and Appendix A for an example):

$$\textit{Eq.2.1}\quad T_{inferiority} = \frac{\delta - A}{S \times \sqrt{N^{-1}}} \text{ and } T_{superiority} = \frac{B - \delta}{S \times \sqrt{N^{-1}}},$$

with N being the total sample size (sum of two groups). For each test, the significance level is found on a Student's t distribution with ν = N - 2 degrees of freedom and S can be estimated *per* Eq.1.2. It must be noticed that the overall significance level of the procedure is 1-2α, in that each directional one-sided test (of inferiority and superiority) has an individual significance level of α (see Appendix B for example). The sample size is necessary for providing each single one-sided test with the adequate sensitivity. The method of Schuirmann, as a refinement of Westlake's procedure, can be also seen under the point of view of the confidence interval (CI). In that we set α = 0.05, thus we can build a (1 - 2α) = 0.9 CI around δ choosing the appropriate value of t:

$$\textit{Eq.2.2}\quad \textbf{90\% CI} = \delta \pm t \times S\sqrt{N^{-1}}$$

If an interval –Δ to +Δ is placed, then the 90% CI approach offers the possibility to show how well the alternative formulation fits the bioequivalence requirement (see Figure 2 and Appendix B for an example). Nowadays the TOST is considered the standard test for bioequivalence assessment (19).

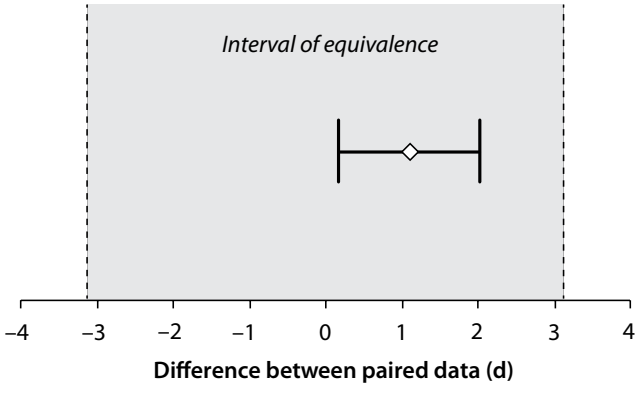

Figure 2. The confidence interval approach (Westlake's method) for TOST-P. The diamond represents the average difference (d = 1.1), while the whiskers are the 90% CI (0.18; 2.20); the grey shaded area is the interval of equivalence with the dashed lines marking its boundaries (-3.12; 3.12).

Equivalence or agreement for laboratory medicine?

So far, we have seen how the equivalence testing stemmed from classical hypothesis testing and became the reference approach for bioequivalence problems. In 1995, Hartmann and co-authors cleverly addressed the limitations of difference testing in statistical procedures for method validation, invoking the adoption of principles of bioequivalence assessment (20). Noteworthy, in 2001 Kondratovich and co-authors recognized the suitability of equivalence testing for comparative studies in laboratory medicine (21). Notably, they also showed that for whatever regression model used to measure the agreement between paired observations, it was possible to reformulate its testing framework using a composite hypothesis of equivalence:

- null hypothesis (H_0) → non-equivalence: slope $\leq 1 - \delta$ or slope $\geq 1 + \delta$
- alternative hypothesis (H_1) → equivalence: $1 - \delta <$ slope $< 1 + \delta$.

In a broader discussion, provided by Mascha and Sessler, it was shown that testing hypothesis of equivalence could be easily achieved through regression analysis (see Appendix C) (22). Thus, despite such evidences we might wonder why we still ignore equivalence testing in laboratory medicine, if we are often concerned with comparing devices and procedures. Do we really need it?

To answer, we should take into consideration two main aspects. First, one is of cultural kind, and concerns the way we have been raised in our probabilistic approach to scientific research. Of course, in laboratory medicine we have inherited the idea of experimental science with the burden of proof resting on difference. Hence, although we might set comparative studies to answer whether a new device or procedure could replace an old one, we still approach it by a classic perspective ignoring equivalence. Thus, we could figure out to easily fill the gap just by popularizing equivalence among biomedical researchers, expecting to see all the new studies approached through such an alternative way within the next ten years, as it just happened in pharmacology.

However, and this is the second aspect, equivalence could be uneasily handled in laboratory medicine. Equivalence cannot stand by itself, but demands the external support provided through the so-called equivalence interval to gain a meaning. Indeed, it is an aprioristic and conservative approach that demands to set an interval of allowance before proving the actual existence of any bias. On the contrary, agreement is a more liberal approach, in that it considers any difference just as an erratic and uninfluential factor unless it is proven otherwise. Thereby, it is concerned with bias only afterwards, giving room to more pragmatic considerations. Thus, apart from our statistical heritage, we should set about stating how much different is equivalent when comparing devices and procedures, and this could generate some confusion. An interesting example of this scenario was offered by Lung and co-authors which discussed the suitability of equivalence testing for assessing automated test procedures (23). In their manuscript, when presenting the analysis of data using TOST, authors advanced two distinct equivalence intervals, ± 2% for assay result and ± 3% for content uniformity. Translated into real laboratory medicine, we should invoke different equivalence intervals for different applications, like therapeutic drug monitoring and hormone testing, and different domains, like comparison between alternative analytical methods and comparison between standard and non-standard pre-analytical procedures. Therefore, any study on equivalence would have its own truth on equivalent methods depending on the criterion adopted, with considerable practical consequences. Let us imagine we set an equivalence interval of ± 5% for analytical results in therapeutic drug monitoring, basing on some considerations regarding the allowable uncertainty at the medical decision limit of the therapeutic window. How many methods would result equivalent within such a narrow range? Almost none if we consider immuno-enzymatic methods, very few if we consider high performance liquid chromatography (HPLC) methods and some more in case of liquid chromatography with tandem mass spectrometry (LC-MS/MS) methods. Thus, we should figure out a scenario in which immuno-en-

zymatic methods were banned from laboratories, but how could we perform urgent testing in the night shift when the LC-MS/MS facility is not operating? In this regard, Feng and co-authors adopted a criterion based on the 15% uncertainty, usually accepted for analytical methods validation above the lower limit of quantitation, to relax the acceptance criterion (24). However, as stated above, it should be assessed earlier, which one is more appropriate with respect to a given scenario. Hence, authoritative organizations in the field of laboratory medicine, like the International Federation of Clinical Chemistry, should discuss the topic before thinking the equivalence replacing the agreement in comparative studies. Otherwise, researchers should be concerned with showing the suitability of a certain criterion of equivalence, before assessing the equivalence itself.

Potential conflict of interest

None declared.

Appendix A – TOST for comparisons on single sample

In the following example, the application of the TOST for comparing hypothetical results of a laboratory assay produced by a standard and an alternative procedure of sampling, considering a ± 5 % of equivalence interval on results with respect to the standard one is presented. In this case, we will use a simple within-subjects repeated-measures single group design without crossing-over, and thus we will carry out the procedure using a two one-sided paired t-test. This is also known as TOST-P (25).

Let us consider first the set of N = 20 paired data:

- standard: 23, 28, 33, 36, 41, 44, 44, 48, 52, 56, 66, 68, 72, 79, 84, 88, 91, 93, 99, 102
- alternative: 25, 28, 39, 38, 38, 43, 46, 49, 53, 52, 61, 73, 77, 82, 86, 86, 95, 96, 97, 105.

The linear correlation is r = 0.99, the average difference d = 1.1 with variance = 9.57 and standard deviation of paired difference S_D = 3.09. The average of standard group is m_s = 62.35, so a ± 5 % difference corresponds to an interval A = - 3.12 and B = 3.12. Recalling the formula of a paired t-test, we can build a TOST-P as follows:

$$\mathbf{TOST-P}_{inferiority} = \frac{d - A}{S_D \times \sqrt{N^{-1}}} \quad \mathbf{TOST-P}_{superiority} = \frac{B - d}{S_D \times \sqrt{N^{-1}}} .$$

Then, we can calculate the T statistics using the data above:

- hypothesis of inferiority: T = (1.1- (- 3.12)) / (9.57 / 20)$^{0.5}$ = 6.10
- hypothesis of superiority: T = (3.12 – 1.1) / (9.57 / 20)$^{0.5}$ = 2.92.

The critical value of T corresponding to a t distribution with N – 1 = 19 degrees of freedom at α = 0.05 is 1.73. Thus we can write:

- hypothesis of inferiority: T observed > t critical → reject → conclude non-inferiority
- hypothesisof superiority: T observed > t critical → reject → conclude non-superiority.

Thus, as data support both non-inferiority and non-superiority, we can conclude the two procedures being equivalent within a margin of ± 5%. We remark that the normality of d distribution must be checked before carrying out the TOST-P, and mostly outliers should be checked and eventually removed to reduce the skew. As an alternative, the same procedure can be adapted to non-parametric Wilcoxon test (26). Calculations were performed using electronic spreadsheet and printed tables of critical t values.

APPENDIX B – TOST-P and the confidence interval approach

Equivalence of procedures shown in Appendix A can be proven also using the Westlake-Schurimann's method of confidence interval. Let's consider the average difference d = 1.1 and its standard deviation S_D = 3.09, the confidence interval can be written as:

$$\textbf{90\% CI} = d \pm t \times S_D \sqrt{N^{-1}}.$$

In this case, the confidence is at 1 - 2α level, so that t = 1.33 and not 1.73 as previously. Therefore we can write:

- upper boundary = 1.1 + 1.33 x (3.09 x 0.22) = 2.02
- lower boundary = 1.1 - 1.33 x (3.09 x 0.22) = 0.18.

Recalling that the ± 5% equivalence interval around the average result of the standard procedure was ± 3.12, we can conclude that the actual 90 CI% is in within and thus the evidence of equivalence is supported (Figure 2). Figure 3 represents all the possible conclusions that can be drawn observing the overlapping between 90% CI and the equivalence interval. It should be remarked that Berger and Hsu in 1999 strongly criticized the 1 - 2α confidence level, proposing alterative procedures to build ordinary 1 - α confidence intervals (27).

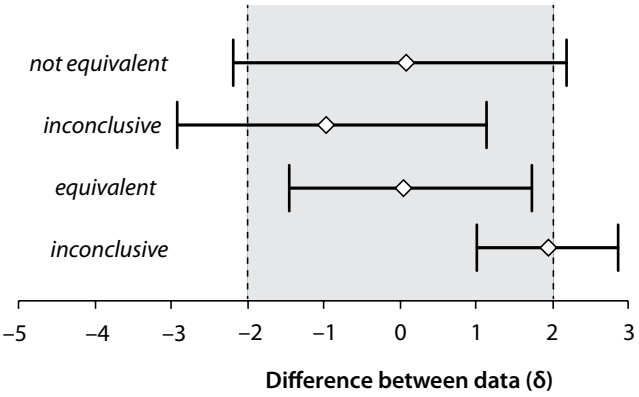

Figure 3. The interpretation of 90% CI with respect to the interval of equivalence. Equivalence can be stated only when the whole confidence interval (whiskers) about the average difference (diamond) rests within the equivalence boundaries (dashed lines), otherwise equivalence can be ruled out; however, if the confidence interval encroaches the equivalence boundaries just on one side the test is inconclusive (i.e. neither it can be stated nor ruled out the equivalence).

APPENDIX C – interval of equivalence for Passing-Bablok regression

Let us imagine that we decided to compare two analytical methods and we obtained a set of two paired results as follows (we reduced the size to 10 pairs for simplicity):

- Reference: 24, 25, 33, 40, 45, 49, 56, 63, 71, 79
- Alternative: 22, 26, 38, 42, 51, 53, 58, 69, 73, 80.

The estimation of regression parameters using the Passing and Bablok model gave a slope of 1.02 and the 95% CI estimated with the jack knife procedure was 0.83 to 1.12, showing no proportional bias (28). However, we were interested in finding whether the two methods resulted equivalent within a ± 15% bias, that is equivalent to the interval:

- lower equivalence boundary: (1 - 0.15) x no bias slope = 0.85 x 1 = 0.85
- upper equivalence boundary: (1 + 0.15) x no bias slope = 1.15 x 1 = 1.15

Thus, we applied the Westlake-Schuirmann's method as described in Appendix B and calculated the 90% CI using the same jack knife procedure as before. With an average pseudo median slope equal to 0.97 and a standard error 0.08, the 90% CI for a critical t value of 1.38 with 9 degrees of freedom at α=0.1 it was:

- lower 90% CI boundary: 0.97 - (1.38 x 0.08) = 0.87
- upper 90% CI boundary: 0.97 + (1.38 x 0.08) = 1.08

Thus, we found the 90% CI resting within the ± 15% bias, letting us to conclude the method being equivalent at the given bias. All calculations were performed using Microsoft Excel spreadsheet and full details are available through supplementary material of this manuscript.

In should be noticed that we used the jack knife method instead of the nested bootstrapping, so that CI value herein might differ from what usually returned by statistical packages. The use of jack knife is purely didactical, in that, although less precise with small sized samples, it is easier to carry out with a simple electronic spreadsheet without any particular adjunctive function and better intelligible (full calculations are available through the MS Excel file which can be provided on request).

References

1. *Salsburg D, ed. The lady tasting tea: how statistics revolutionized science in the twentieth century. 1st ed. New York: W.H. Freeman; 2001. p. 340.*
2. *Yates F. Sir Ronald Fisher and the Design of Experiments. Biometrics 1964;20:307-21. https://doi.org/10.2307/2528399.*
3. *Forrest DW, ed. Francis Galton: the life and work of a Victorian genius. New York: Taplinger; 1974.*
4. *Stigler SM. Darwin, Galton and the Statistical Enlightenment. J R Statist Soc 2010;173:469-82. https://doi.org/10.1111/j.1467-985X.2010.00643.x.*
5. *Lehmann EL. Fisher, Neyman, and the creation of classical statistics. New York: Springer science+Bussines Media; 2011. p. 115.*
6. *Box JF. R. A. Fisher, the life of a scientist. New York: Wiley; 1978. p. 512.*
7. *Stigler SM. The history of statistics: the measurement of uncertainty before 1900. Cambridge, Mass: Belknap Press of Harvard University Press; 1986. p. 410.*
8. *Lehmann EL. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Amer Statist Assoc 1993;88:1242-9. https://doi.org/10.1080/01621459.1993.10476404.*
9. *Lehmann EL. Testing statistical hypotheses. New York: Wiley; 1959. p. 369.*
10. *Westlake WE, Ittig M, Gunther FA. Determination of m-sec-butylphenyl N-methyl-N-thiophenylcarbamate (RE-11775) in water, soil and vegetation. Bull Environ Contam Toxicol 1972;8:109-12. https://doi.org/10.1007/BF01684516.*
11. *Westlake WJ. Statistical aspects of comparative bioavailability trials. Biometrics 1979;35:273-80. https://doi.org/10.2307/2529949.*
12. *Kirkwood TBL. Bioequivalence testing: a need to rethink (reader reaction). Biometrics 1981;37:589-91. https://doi.org/10.2307/2530573.*
13. *Westlake WJ. Response to bioequivalence testing: a need to rethink (reader reaction response). Biometrics 1981;37:591-3.*
14. *O'Quigley J, Baudoin C. General approaches to the problem of bioequivalence. Statistician 1988;37:51-8. https://doi.org/10.2307/2348378.*
15. *Hauck WW, Anderson S. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. J Pharmacokinet Biopharm 1984;12:83-91. https://doi.org/10.1007/BF01063612.*
16. *Rocke DM. On testing for bioequivalence. Biometrics 1984;40:225-30. https://doi.org/10.2307/2530763.*
17. *Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 1987;15:657-80. https://doi.org/10.1007/BF01068419.*
18. *Schuirmann DJ. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics 1981;37:617.*
19. *U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. 2001.*
20. *Hartmann C, Smeyers-Verbeke J, Penninckx W, Vander Heyden Y, Vankeerberghen P, Massart DL. Reappraisal of hypothesis testing for method validation: detection of systematic error by comparing the means of two methods or of two laboratories. Anal Chem 1995;67:4491-9. https://doi.org/10.1021/ac00120a011.*
21. *Kondratovich MV, Yue LQ, Dawson JM. Power approach versus two one-sided tests procedure in equivalence evaluation with diagnostic medical devices. Proceedings of the Annual Meeting of the American Statistical Association; 2001; Atlanta, Georgia.*
22. *Mascha EJ, Sessler DI. Equivalence and noninferiority testing in regression models and repeated-measures designs. Anesth Analg 2011;112:678-87. https://doi.org/10.1213/ANE.0b013e318206f872.*
23. *Lung KR, Gorko MA, Llewelyn J, Wiggins N. Statistical method for the determination of equivalence of automated test procedures. J Autom Methods Manag Che 2003;25:123-7. https://doi.org/10.1155/S146392460300021X.*
24. *Feng S, Liang Q, Kinser RD, Newland K, Guilbaud R. Testing equivalence between two laboratories or two methods using paired-sample analysis and interval hypothesis testing. Anal Bioanal Che 2006;385:975-81. https://doi.org/10.1007/s00216-006-0417-2.*
25. *Mara CA, Cribbie RA. Paired-samples test of equivalence. Commun Stat Simulat 2012;41:1928-43. https://doi.org/10.1080/03610918.2011.626545.*
26. *Meier U. Nonparametric equivalence testing with respect to the median difference. Pharm Stat 2010;9:142-50.*
27. *Berger RL, Hsu JC. Bioequivalence trials, intersection-unions tests and equivalence confidence sets. Stat Sci 1999;11:283-319.*
28. *Abdi H, Williams LJ. Jackknife. In: Salkind N, ed. Encyclopedia of research design. Thousand Oaks, CA: Sage; 2010.*