This series of knowledge sharing articles is a project of the Standardized Biofilm Methods Laboratory in the CBE

KSA-SM-17

Assessing neutralization using ASTM E1054

Al Parker
11/15/2021

[Key Words: equivalence, neutralization]

Contents

Overview

Neutralization of a product's antimicrobial efficacy is a crucial yet understudied aspect of antimicrobial test methods. The effectiveness of a neutralizer is usually assessed up front and then not considered again when testing the efficacy of an antimicrobial in a variety of applications including on hard non-porous surfaces[1], hands or textiles. Without a neutralization step, it would be challenging to study the effect of an antimicrobial over a specific contact time. For example, on hard non-porous surfaces, contact times are typically 5-10 minutes, whereas for hand sanitizers the contact time is around 30 seconds. In this article we study the statistical methods associated with ASTM's standard method E1054 that describes how to assess whether a neutralizer really does neutralize an antimicrobial's effectiveness against the test microbe, and whether the neutralizer by itself is toxic to microbes. Historically, E1054 used statistical significance tests to answer these questions. In 2021, the method was updated to use statistical equivalence tests. Here, we discuss differences between equivalence and significance testing, elucidate examples, and show how to use the statistical software R and Excel to perform the calculations.

E1054 has been applied against a variety of bacteria and fungi, both planktonic and biofilms, and viruses. Briefly, E1054 considers four Test Groups (see Figure 1). Test A assesses neutralizer effectiveness by considering multiple replicate microbial samples subjected to both the antimicrobial product and the neutralizer. Test B assesses neutralizer toxicity by considering multiple replicate microbial samples subjected to just the neutralizer. The Test A and Test B samples are compared to a Test C group composed of multiple replicate samples that contain only microbes that serve as untreated controls. There is also a Test D that considers multiple replicate microbial samples subjected to the same concentration of the antimicrobial product that was used in Test A to verify that the antimicrobial product used in the neutralization test is efficacious against the microbes. There are other important details to E1054, such as repeating the neutralization evaluation **at least 3 times** (i.e., there should be at least 3 samples in each of the Test A, Test B, Test C and Test D groups) and then plating or filtering each sample at two different time points (immediately after sample preparation and after a hold period).

[1] One notable exception to this paradigm is the MBEC standard method E2799 that performs a neutralization test every time the method is performed.

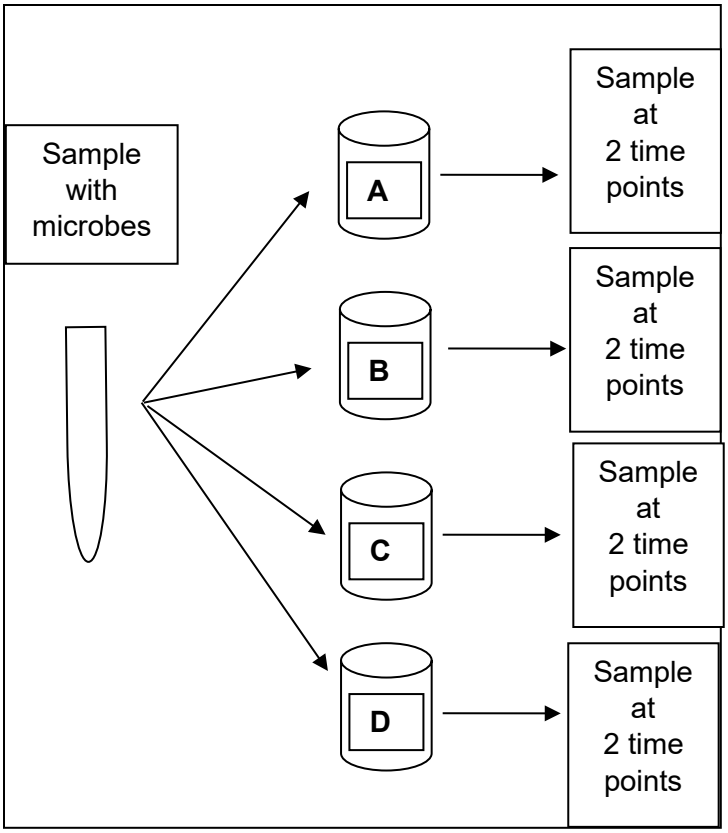

Figure 1. A single neutralization evaluation described by E1054. This should be repeated at least 3 times.

Historical use of statistical testing in E1054

Before 2021, E1054 suggested the use of statistical significance tests, such as a *t*-test, to compare the means of each of the Test A, Test B and Test D replicates to the mean of the Test C replicates. If the *p*-value for the statistical test applied to Test A was larger than 0.05, then E1054 concluded that the neutralizer was effective against the product being tested. If the *p*-value for the statistical test applied to Test B was larger than 0.05, then E1054 concluded that the neutralizer was not toxic to the microbes. And finally, if the *p*-value for the statistical test applied to Test D was less than 0.05, then E1054 concluded that the antimicrobial used in the test was effective at killing the microbes. Statisticians immediately understand the dubiousness of such emphatic conclusions based solely on a large *p*-value. To understand why, we will review the competing hypotheses that are weighed when performing a statistical test. In the context of neutralization tests, the null hypothesis states that there is no mean difference between the two groups (Test A and the Test C controls; or Test B and the Test C controls), while the alternative hypothesis states that there is a mean difference between the two groups. The statistical test is conducted assuming that the null hypothesis is true. If the data deviate sharply from this assumption of no mean difference, then the test yields a small p-value < 0.05 at which point one can reject the null hypothesis and conclude the alternative hypothesis. The conclusion is stated as if from a judge in a court of law: the evidence suggests that there is a mean difference between the two groups. When the test yields a large p-value > 0.05, then the conclusion is that the evidence fails to suggest that there is a mean difference between the two groups. For Test A, the null hypothesis is that the neutralizer is effective (there is no mean difference from the controls) and the alternative is that the neutralizer is not effective (there is a mean difference from the controls). So when there is a large *p*-value for Test A, the evidence fails to suggest that the neutralizer is not effective. This conclusion is not only confusing (a double negative!) but is also a much different conclusion than what we want to make, which is that the evidence **does** suggest that the neutralizer is effective. Absence of evidence

is not evidence of absence! For Test B, the null hypothesis is that the neutralizer is not toxic and the alternative hypothesis is that the neutralizer is toxic. So, when there is a large *p*-value for Test B, the evidence fails to suggest that the neutralizer is toxic.

Modern equivalence testing in E1054

The solution to this conundrum is to use a statistical equivalence test instead of a statistical significance test to compare Test A and Test B to the Test C controls. In an equivalence test, the traditional null and alternative hypotheses are flipped, so that the test is conducted assuming the null hypothesis that there is a large mean difference between the two groups, and the alternative hypothesis is that the two groups do not have a large mean difference. The clincher is that one must define what is meant by a "large difference" referred to as the *equivalency margin* and denoted by δ. One way to interpret an equivalency margin is that mean differences between groups less than the equivalency margin are negligible and not of practical importance. Conducting a statistical equivalence test is just as straightforward as conducting a significance test: one must calculate a 90% confidence interval (CI) for the mean difference to assess equivalence at 95% confidence[2] (Welleck 2010). If that 90% CI is contained within [-δ, δ], then one may conclude that the evidence suggests that the means of the two groups are equivalent. When Test A is equivalent to the Test C controls, then the evidence suggests that the neutralizer is effective. When Test B is equivalent to the Test C controls, then the evidence suggests that the neutralizer is not toxic. The CI can be constructed (just as *p*-values can be generated) using either a *t*-test, ANOVA, or a linear mixed effects model. In the examples below, we consider a *t*-test.

The 2021 revision to E1054 dispenses with the use of statistical significance testing and instead uses statistical equivalence tests to assess both neutralizer effectiveness and neutralizer toxicity. Based on precedent (Allkja et al 2021, Fritz et al 2015, Parker at al 2014), as well as guidance from EPA (Nelson et al 2013) and FDA (FDA 2020) regarding $\log_{10}$(CFU) data, E1054 specifies the equivalency margin δ=0.5. This means that differences as large as 0.5 between the mean $\log_{10}$(CFU)'s for Test A (or Test B) and the Test C controls are considered negligible and not of practical importance when assessing neutralizer efficacy. Put another way, using a 90% CI, the median CFU for the control can be as low as 32% of the Test median CFU and as high as 3.2 times larger than the Test median CFU and the conclusion is equivalence (see Appendix). Interestingly, a more restrictive equivalency margin of 0.1 on the $\log_{10}$-scale was proposed by FDA in 2001 for general equivalence testing. This corresponds to the median CFU for the controls being as low as 80% of the Test median CFU and as high as 1.25 times larger than the Test median CFU using a 90% CI and still conclude equivalence. In our experience, an equivalency margin of 0.1 would be too restrictive due to the large variability observed from microbes stressed by antimicrobial treatments (Parker et al 2018).

Software Tools

We suggest using the software R (R Core Team 2021) to perform the calculations needed for implementing an equivalence test. An Excel spreadsheet is also provided that can perform equivalence tests, although Excel uses an approach that is not as statistically powerful as the approach used by R (see Example #1 below). One advantage of using the Excel spreadsheet is that it provides the user with a convenient GUI interface for entering the data and seeing the outcome of the equivalency test. The spreadsheet needs inputs only in the yellow cells. To allow for complete transparency, the spreadsheet has not been locked down, which means that users may click on any cell and see the underlying formulas (if any) that drive the calculations for the equivalence test. To conduct a neutralization test, E1054 specifies 6 different tests:

1. Test A at Time 1 using an equivalency test
2. Test A at Time 2 (after a Hold) using an equivalency test
3. Test B at Time 1 using an equivalency test
4. Test B at Time 2 (after a Hold) using an equivalency test

[2] Yes, one conducts an equivalence test at 95% confidence using a 90% confidence interval.

5. Test D at Time 1 using a significance test
6. Test D at Time 2 (after a Hold) using a significance test

The spreadsheet implements the equivalence tests 1-4. To do this, the spreadsheet is composed of 7 separate sheets:

1. **Parameters.** Two inputs are needed to implement an equivalence test. The first input is the equivalency margin in cell E5. E1054 specifies 0.5, which is the default value in the spreadsheet. The second input is the equivalence test confidence level. E1054 specifies 95% confidence, which is the default value in the spreadsheet.
2. **Example.** This sheet shows an example Test A and control data as outlined in Example #1 below.
3. **Test A, Time 1.** Complete the following steps to test whether the Test A CFU data are equivalent to the control CFUs at the first time point:
    1) Go to the "Test A, Time 1" sheet
    2) Enter the sample size for the Test C control data into cell A5.
    3) You will see yellow cells appear in the "CFU" column in column B. The number of yellow cells that appear is the same as the sample size. Enter in the average Test C control CFUs for each replicate sample into the yellow cells in column B.
    4) Enter the sample size for the Test A data into cell G5.
    5) You will see yellow cells appear in the "CFU" column in column H. The number of yellow cells that appear is the same as the sample size. Enter in the average Test A CFUs for the replicate samples into the yellow in column H.
    6) Given these data, the spreadsheet will automatically calculate a CI for the difference in means $\log_{10}$(CFU)'s between the Test A and Test C control groups, compare the CI to [-0.5, 0.5], and declare the conclusion of either "Statistically Equivalent" or "Not Statistically Equivalent".
4. **Test A, Time 2.** The data for this sheet are entered the same as outlined for **Test A, Time 1**.
5. **Test B, Time 1.** The data for this sheet are entered the same as outlined for **Test A, Time 1**.
6. **Test B, Time 2.** The data for this sheet are entered the same as outlined for **Test A, Time 1**.
7. **Summary.** This sheet collects the results from the sheets **Test A, Time1; Test A, Time 2; Test B, Time 1;** and **Test B, Time 2** and provides a single concise statistical summary, including sample sizes, means, SDs, CIs, degrees of freedom and equivalence conclusions.

In addition to an empty Excel spreadsheet that users may populate with their own CFU data and then perform equivalence tests, this KSA also includes an example spreadsheet that has already been populated with CFU data for Tests A and B at both time points.

Examples

Four examples are considered below. The R code and output is provided for all 4 examples. The Excel spreadsheet is applied to Example #1 only. The examples convey a range of scenarios meant to show the benefit of equivalence testing versus the historical use of statistical significance testing to assess neutralization.

Example #1: Equivalence testing shows the neutralizer is efficacious

The Appendix in E1054 considers 3 replicate samples in Test A and 3 control replicate samples in group C as in Table 1. The average CFUs from Table 1 will be subjected to an equivalence test, where there is one average CFU per replicate sample.

Table 1. Example #1 CFU data generated from 2 plates for each replicate sample. A single mean CFU is calculated per replicate sample.

Group	Replicate Sample	Plate 1	Plate 2	Mean CFU
Control	1	20	21	20.5
	2	19	43	31
	3	26	10	18
Test A	1	41	23	32
	2	30	51	40.5
	3	11	27	19

R code and output

The R code and output when performing an equivalence test of Test A with the Test C controls is:

```
C = c(20.5, 31,18)
A = c(32, 40.5, 19)
t.test(log10(C),log10(A),conf.level=.9)
##
##  Welch Two Sample t-test
##
## data:  log10(C) and log10(A)
## t = -0.92181, df = 3.6675, p-value = 0.4132
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -0.3745113  0.1525309
## sample estimates:
## mean of x mean of y
##  1.352796  1.463786
```

Excel spreadsheet

The equivalence test is performed by the accompanying Excel spreadsheet by completing the following steps:

1) Go to the "Test A, Time 1" sheet
2) Enter the sample size of 3 for the Test C control data into cell A5.
3) You will see 3 yellow cells appear in the "CFU" column in column B. Enter in the average Test C control CFUs for each replicate sample from Table 1 into the yellow cells B5:B7.
4) Enter the sample size of 3 for the Test A data into cell G5.
5) You will see 3 yellow cells appear in the "CFU" column in column H. Enter in the average Test A CFUs for the replicate samples from Table 1 into the yellow cells H5:H7.
6) Given these data, the spreadsheet will automatically calculate a 90% CI for the difference in means, compare the CI to [-0.5, 0.5], and declare the conclusion of either "Statistically Equivalent" or "Not Statistically Equivalent".

The Example sheet in the spreadsheet shows how the populated sheet ought to look for this example.

Unfortunately, the 90% CI calculated by Excel is larger than the 90% CI calculated by R. This is because R can use any decimal valued degrees of freedom whereas Excel can only use integer valued degrees of freedom. Hence, Excel always rounds the degrees of freedom down to the next integer and then constructs the CI. The moral is to use R for equivalence testing instead of Excel. The spreadsheet is included for ease of use for those who do not wish to use R.

Discussion and Conclusion

The equivalence test proceeds by comparing the 90% CI [-0.375, 0.153] to [-0.5, 0.5]. Because [-0.375, 0.153] is contained within [-0.5, 0.5], then, at 95% confidence, Test A is statistically equivalent to the Test C controls. Hence, the evidence provided by implementation of E1054 suggests that the neutralizer is effective.

Compare this equivalence analysis to the historical use of statistical significance testing. The R output above shows that the *t*-test *p*-value = 0.413. Therefore, the evidence fails to suggest that there is a mean difference between Test A and the Test C controls. Put another way, the evidence fails to suggest that the neutralizer is not effective.

To see why equivalence and neutralizer efficacy were concluded in this example, see the means and SDs for these example data in Table 2.

Table 2. Means and SDs for Example #1 data set.

Group	*n*	mean $\log_{10}$(CFU)	SD
Control	3	1.35	0.123
Test A	3	1.46	0.168

Table 2 shows that the means for the two groups are close and also that the variability of the two groups is very tight. This is an important aspect of equivalence testing: the variability in the data must be small[3] AND the difference in means must be small in order to conclude equivalence.

Example #2: Equivalence testing fails to show that the neutralizer is efficacious

This example considers the data in Table 3. In this example, the CFUs for the control replicates are exactly the same as in Example #1 (Table 1), whereas the CFUs for the Test A replicates are much more variable than the Test C controls.

Table 3. Example #2 showing CFUs from 2 plates for each replicate sample. From these, a single average CFU is calculated per replicate sample.

Group	Replicate Sample	Plate 1	Plate 2	Mean CFU
Control	1	20	21	20.5
	2	19	43	31
	3	26	10	18
Test A	1	410	230	320
	2	3	5	4
	3	11	27	19

R code and output

The R code and output when performing an equivalence test of Test A with the Test C controls is:

```
C = c(20.5, 31,18)
A = c(320, 4, 19)
t.test(log10(C),log10(A),conf.level=.9)
```

[3] If the variability on the data is large, increasing the sample size *n* will yield a more narrow CI that may substantiate a conclusion of equivalence.

```
##
##  Welch Two Sample t-test
##
## data:  log10(C) and log10(A)
## t = -0.19447, df = 2.0653, p-value = 0.8633
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -1.713924  1.495540
## sample estimates:
## mean of x mean of y
##  1.352796  1.461988
```

Discussion and Conclusion

The equivalence test proceeds by comparing the 90% CI [-1.71, 1.50] to [-0.5, 0.5]. Because [-1.71, 1.50] is not contained within [-0.5, 0.5], then, at 95% confidence, the evidence fails to suggest that Test A is statistically equivalent to the Test C controls. Hence, the evidence provided by implementation of E1054 fails to suggest that the neutralizer is effective.

Compare this equivalence analysis to the historical use of statistical significance testing. The output above shows that the *p*-value for the *t*-test is $p = 0.863$. Therefore, the evidence fails to suggest that there is a mean difference between Test A and the Test C controls. Put another way, the evidence fails to suggest that the neutralizer is not effective.

To see why the evidence failed to suggest neutralizer efficacy, see the means and SDs for these data in Table 4.

Table 4. Means and SDs for Example #2 data set.

Group	***n***	**mean $\log_{10}$(CFU)**	**SD**
Control	3	1.35	0.123
Test A	3	1.46	0.965

Table 4 shows that the means for the two groups are just as close as they were for Example #1 (Table 2). However, the SD of Test A data is very large! This is an important aspect of equivalence testing: the variability in the data must be tight AND the means must be close in order to conclude equivalence. One way to overcome the large variability in the Test A group to assess equivalence with more statistical power would be to increase the number of replicates in the Test A group.

Example #3: Significance testing shows that the neutralizer is not efficacious

This example considers the data in Table 5. In this example, the CFUs for the control replicates are exactly the same as in Example #1 (Table 1), whereas the CFUs for the Test A replicates are much larger than the Test C controls.

Table 5. Example #3 showing CFUs from 2 plates for each replicate sample. From these, a single average CFU is calculated per replicate sample.

Group	Replicate Sample	Plate 1	Plate 2	Mean CFU
Control	1	20	21	20.5
	2	19	43	31
	3	26	10	18
Test A	1	410	230	320
	2	300	510	405
	3	110	270	190

R code and output

The R code and output when performing an equivalence test of Test A with the Test C controls is:

```
C = c(20.5, 31,18)
A = c(320, 405, 190)
t.test(log10(C),log10(A),conf.level=.9)
##
##  Welch Two Sample t-test
##
## data:  log10(C) and log10(A)
## t = -9.2271, df = 3.6675, p-value = 0.001144
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -1.3745113 -0.8474691
## sample estimates:
## mean of x mean of y
##  1.352796  2.463786
```

Discussion and Conclusion

The equivalence test proceeds by comparing the 90% CI [-1.37, -0.85] to [-0.5, 0.5]. Because [-1.37, -0.85] is not contained within [-0.5, 0.5], then, at 95% confidence, the evidence fails to suggest that Test A is statistically equivalent to the Test C controls. Hence, the evidence provided by implementation of E1054 fails to suggest that the neutralizer is effective.

Compare this equivalence analysis to the historical use of statistical significance testing. The output above shows that the *p*-value for the *t*-test is $p = 0.001$. Therefore, the evidence suggests that there is a mean difference between Test A and the Test C controls. Put another way, the evidence suggests that the neutralizer is not effective.

To see why the equivalence test failed to show the neutralizer was effective while the significance test showed that the neutralizer was not effective, see the means and SDs for these data in Table 6.

Table 6. Means and SDs for Example #3 data set.

Group	*n*	mean $\log_{10}$(CFU)	SD
Control	3	1.35	0.123
Test A	3	2.46	0.168

Table 6 shows that the means for the two groups are very different but that the variability of the two groups is very tight (the same as for Example #1 (Table 2)). This is an important aspect of equivalence testing: the variability in the data must be tight AND the means must be close in order to conclude equivalence.

Example #4: Conundrum? Equivalence testing shows the neutralizer is efficacious while significance testing shows that the neutralizer is not efficacious

This example considers the data in Table 7.

Table 7. Example #4 showing CFUs from 2 plates for each replicate sample. From these, a single average CFU is calculated per replicate sample.

Group	Replicate Sample	Plate 1	Plate 2	Mean CFU
Control	1	20	25	22.5
	2	19	29	24
	3	26	15	20.5
Test A	1	33	23	28
	2	10	51	30.5
	3	29	27	28

R code and output

The R code and output when performing an equivalence test of Test A with the Test C controls is:

```
C = c(22.5, 24, 20.5)
A = c(28, 30.5, 28)
t.test(log10(C),log10(A),conf.level=.9)
##
##  Welch Two Sample t-test
##
## data:  log10(C) and log10(A)
## t = -4.7622, df = 3.3495, p-value = 0.01363
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -0.16431780 -0.05866105
## sample estimates:
## mean of x mean of y
##  1.348049  1.459539
```

Discussion and Conclusion

The equivalence test proceeds by comparing the 90% CI [-0.164, -0.059] to [-0.5, 0.5]. Because [-0.164, -0.059] is contained within [-0.5, 0.5], then, at 95% confidence, the evidence suggests that Test A is statistically equivalent to the Test C controls. Hence, the evidence provided by implementation of E1054 suggests that the neutralizer is effective.

Compare this equivalence analysis to the historical use of statistical significance testing. The output above shows that the *p*-value for the *t*-test is $p = 0.014$. Therefore, the evidence suggests that there is a mean difference between Test A and the Test C controls. Put another way, the evidence suggests that the neutralizer is not effective.

These two conclusions at face value appear to contradict each other. However, when conducting the equivalence test, E1054 specifies that differences as large as 0.5 between the mean $\log_{10}$(CFU)'s for Test A and the Test C controls are negligible and not of practical importance. In other words, even though Test A and the Test C controls were found to be "statistically significantly" different on average ($p = 0.014 < 0.05$), the mean difference between the two groups is negligible are not of practical importance.

To see why the equivalence test showed the neutralizer was effective while the significance test showed that the neutralizer was not effective, see the means and SDs for these data in Table 8.

Table 8. Means and SDs for Example #3 data set.

Group	***n***	**mean $\log_{10}$(CFU)**	**SD**
Control	3	1.35	0.034
Test A	3	1.46	0.021

Table 8 shows that the means for the two groups are just as close as they were for Example #1 (Table 2). However, the variability of the two groups is even tighter! The final conclusion in this case follows from the equivalence test: the evidence suggests that the neutralizer is effective.

Conclusion

Table 9 summarizes the comparisons made between equivalence testing and significance testing for assessing neutralizer effectiveness in the previous 4 examples.

Table 9. When assessing neutralizer efficacy by comparing the neutralized disinfectant in Test A to the Test C controls, the possible conclusions in the following table occur depending on aspects of the CFU data in each of Test A and the Test C controls. The green cells indicate scenarios where the conclusion from equivalence testing is synergistic to the conclusion from significance testing.

	***p* > 0.05** The evidence fails to suggest the neutralizer is not efficacious	***p* < 0.05** The evidence suggests the neutralizer is not efficacious
90% CI is in [-0.5, 0.5] The evidence suggests the neutralizer is efficacious	Data has small variability, small difference in means (Example #1)	Data has very small variability, small difference in means (Example #4)
90% CI is not in [-0.5, 0.5] The evidence fails to suggest the neutralizer is efficacious	Data has large variability (Example #2)	Data has small variability, large difference in means (Example #3)

Table 10 gives similar conclusions as Table 9, this time for assessing neutralizer toxicity.

Table 10. When assessing neutralizer toxicity by comparing the neutralizer in Test B to the Test C controls, the possible conclusions in the following table occur depending on aspects of the CFU data in each of Test B and the Test C controls. The green cells indicate scenarios where the conclusion from equivalence testing is synergistic to the conclusion from significance testing.

	***p* > 0.05** The evidence fails to suggest the neutralizer is toxic	***p* < 0.05** The evidence suggests the neutralizer is toxic
90% CI is in [-0.5, 0.5] The evidence suggests the neutralizer is not toxic	Data has small variability, small difference in means	Data has very small variability, small difference in means
90% CI is not in [-0.5, 0.5] The evidence fails to suggest that the neutralizer is not toxic	Data has large variability	Data has small variability, large difference in means

The bottom line is that the historical use of significance testing led investigators to conclude that neutralization was successful based on a lack of evidence to the contrary. This lack of evidence can occur, not surprisingly, when the data exhibit large variability even when the mean difference between the Test and control groups is large (see Tables 9 and 10). Using equivalence testing requires that investigators generate data with small mean differences AND acceptably low variance. If low variance is not possible, then the investigator can increase the number of replicate samples in at least one group (Test A, Test B and/or the Test C controls, see Example #2) to overcome high levels of variance in the data.

References

Allkja, Juliana Porto de Abreu, Charante, Reigada, Guarch-Perez, Vasquez-Rodriguez, Cos, Coenye, Fallarero, Zaat, Felici, Ferrari, Azevedo, Parker, Goeres. Interlaboratory study for the evaluation of three microtiter plate-based biofilm quantification methods. *Scientific Reports* 11 13779, 2021.

Fritz, B., D. Walker, D. Goveia, A. Parker, and D. Goeres. Evaluation of Petrifilm Aerobic Count plates as an equivalent alternative to drop plating on R2A agar plates in a biofilm disinfectant efficacy test. *Current Microbiol*, 70(3): 450-456, 2015.

Nelson, M., LaBudde, R., Tomasino, S., Pines, R. Comparison of 3M Petrifilm TM aerobic plate counts to standard plating methodology for use with AOAC antimicrobial efficacy methods 955.14, 955.15, 964.02, and 966.04 as an alternative enumeration procedure: collaborative study. *JAOAC* 96 (4), 2013.

Parker, A., D. Walker, D. Goeres, N. Allan, M. Olson, and A. Omar. Ruggedness and reproducibility of the MBEC biofilm disinfectant efficacy test. *J Microbiol Methods*, 102: 55-64, 2014.

Parker, Hamilton, Goeres. Reproducibility of antimicrobial test methods. *Scientific Reports* 8:12531, 2018.

FDA, *Safety and Effectiveness of Health Care Antiseptics; Topical Antimicrobial Drug Products for Over-the-Counter Human Use, A Rule by the Food and Drug Administration*. Fed. Reg. 82(243): 60487, 21 CFR 310, Dec 17, 2020, URL:
https://www.federalregister.gov/documents/2017/12/20/2017-27317/safety-and-effectiveness-of-health-care-antiseptics-topical-antimicrobial-drug-products-for

FDA, Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER), January 2001.

R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria, 2021.

Welleck, S. *Testing Statistical Hypotheses of Equivalence and Noninferiority.* CRC Press, 2010.

Appendix

To see that conducting an equivalence test with an equivalency margin of 0.5 on the $\log_{10}$ scale is equivalent to allowing the control median CFU to be as low as 32% and as high as 3.2 times the Test median CFU, start with the equivalency requirement

$$-0.5 < \text{mean } \log_{10}(C) - \text{mean } \log_{10}(T) < 0.5$$

where C represents the CFUs (averaged over plates) for the Test C controls and T represents the CFUs (averaged over plates) for the Test group (either Test A or Test B). When the $\log_{10}(C)$ values for the controls and the $\log_{10}(T)$ values for the Test group are symmetric, as occurs when the data are normal[4], then the previous inequality becomes

$$-0.5 < \text{median } \log_{10}(C) - \text{median } \log_{10}(T) < 0.5.$$

Because the $\log_{10}$ transform is monotonic, then it preserves the ordering of the data and hence preserves medians so that the inequality becomes

$$-0.5 < \log_{10}(\text{median C}) - \log_{10}(\text{median T}) < 0.5.$$

Rules of logs now can be applied to rewrite this expression as

$$-0.5 < \log_{10}([\text{median C}]/[\text{median T}]) < 0.5$$
$$0.32 < [\text{median C}]/[\text{median T}] < 3.2.$$

Now consider an equivalency margin of 0.1 on the $\log_{10}$ scale. The following expressions show that this is the same as allowing the median control CFU be as low as 80% and as high as 1.25 times the median Test group CFU,

$$-0.1 < \text{mean } \log_{10}(C) - \text{mean } \log_{10}(T) < 0.1$$
$$-0.1 < \log_{10}(\text{median C}) - \text{mean } \log_{10}(\text{median T}) < 0.1$$
$$-0.1 < \log_{10}([\text{median C}]/[\text{median T}]) < 0.1$$
$$0.8 < [\text{median C}]/[\text{median T}] < 1.25.$$

[4] Normality of the data in each group is the assumption of ANOVA, *t*-tests and linear mixed effects models for small sample sizes such as those considered by neutralization tests.