MONTANA STATE UNIVERSITY College of ENGINEERING
Center for Biofilm Engineering
a National Science Foundation Engineering Research Center

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

KSA-SM-011

How to decide whether the reproducibility standard deviation is small enough

[*Key Words*: maximum acceptable discrepancy, collaborative study, β-expectation tolerance interval, prediction interval, S_R]

Nomenclature Highlights

log reduction, LR

Reproducibility SD, S_R

Goal of article

Using historical precedent for deciding whether S_R is sufficiently small

Criticism of the historical precedent approach

Published: September 15, 2011
Revised: October 30, 2011

Background

A disinfectant test is a laboratory method for measuring the efficacy of a disinfectant treatment. In order for a disinfectant test method to be accepted for standard use, it must exhibit good reproducibility as determined by a collaborative study in which the same disinfectant treatment is tested by different laboratories. In the typical situation where efficacy is measured by the log reduction (LR), good reproducibility is demonstrated when the LR values observed in the collaborative study are judged to be similar enough. To help with that judgment, it is conventional to quantify the variability among LR values by calculating the reproducibility standard deviation, S_R, of LR (KSA-SM-03: Testing surface disinfectants: desirable attributes of a standardized method). The acceptability of the test method amounts to deciding whether the observed S_R is small enough. However, that judgment is problematic because decision criteria have not been established for the disinfectant test S_R. The goal of this KSA is to provide a quantitative framework for determining whether the collaborative study S_R is acceptably small.

Decisions about the acceptability of an S_R often have relied on historical precedent; i.e., on the S_R values for disinfectant test methods that have been accepted already by the community of experts. The rationale is that, if a method is considered to be an acceptable disinfectant test, perhaps because it was accepted by a reputable standards setting organization, then the S_R for that method necessarily must be sufficiently small. In our work at the CBE, we have relied on the literature review by Tilt and Hamilton (1999) which found that S_R ranged from 0.31 to 1.54 across a variety of accepted suspension tests and dried surface tests when commonly-used disinfectant formulations were tested against the usual laboratory microbes. Based on that information, we have recommended the acceptability threshold, $S_R \leq 1.5$ (KSA-SM-10: Assessing Resemblance, Repeatability, and Reproducibility for Quantitative Methods).

There are some important weaknesses in reliance on historical precedent. One is that the Tilt and Hamilton results were based only on those tests for which comparable collaborative study S_R values were available to the authors, not on a representative sample of all accepted disinfectant test methods. The review included many

© 2011 MSU Center for Biofilm Engineering

standardized European dried surface test methods and suspension test methods that were validated before 1995. Because of advances in laboratory practice, technology, and equipment since 1995, one would expect that current methods are more reproducible. Another disadvantage is a lack of specificity because it was necessary to aggregate results from a spectrum of disinfectant tests (applied using different microbes to test different disinfectant treatments), with minimal consideration of the real-world application that each test represented. Finally, the historical approach relies on subjective, possibly flawed, decisions about the acceptability of disinfectant test methods.

Need a new framework for S_R acceptability decisions

There is a need for an acceptability decision framework that is statistically sound, flexible, relatively easy to understand, and not dependent on historical data. To meet that need, we have borrowed some ideas from another scientific field. Pharmaceutical statisticians have recently developed a framework for judging the acceptability of a chemical assay method (e.g., Hubert et al. 2004; Hoffman and Kringle 2007; Rozet et al. 2007; Feinberg et al. 2010). The details of their approach are not directly applicable to disinfectant tests; however, we have adapted their main concepts to create a decision framework suitable for evaluating a disinfectant test S_R. In this article we describe the decision framework and illustrate its application using published collaborative study data for a quantitative, dried surface test method.

This article describes a promising new framework for evaluating S_R

Acceptable discrepancy framework for evaluating S_R

Instead of focusing directly on the standard deviation S_R, we will start by discussing a desirable characteristic of a disinfectant test; namely, that it produces an LR result *likely* to be *near* the true log reduction. This characteristic can be expressed quantitatively by the subject area expert (e.g., a stakeholder such as a manufacturer or a regulatory authority) via numerical values for *likely* and *near*. For example, the expert could require that, with probability 0.9, the LR from the next test will be within 1.0 of the true log reduction; here, 0.9 quantitatively expresses *likely* and 1.0 quantitatively describes *near*. Now we need to convert these quantitative specifications for *likely* and *near* into a mathematically equivalent acceptability criterion for S_R. The pharmaceutical statisticians showed that such a conversion can be accomplished using conventional statistical techniques such as tolerance interval or prediction interval calculations (Vardeman 1992).

true, unknown LR, λ

discrepancy, D

max acceptable D, δ
probability, β

Mathematical notation is helpful for describing this approach. For a specified disinfectant treatment (formulation, use contact time, use concentration, etc.), let λ denote the true, unknown LR. For the next test of the disinfectant formulation, a test that hasn't been conducted yet, the symbol **LR** holds the place for the actual numerical log reduction that will be observed. Let D denote the "discrepancy" between the test outcome and the truth; that is, $D = \mathbf{LR} - \lambda$. The expert chooses a numerical "maximum acceptable discrepancy," denoted by δ (defining near), and a probability, denoted by β (likely). The quantity δ is a positive value on the same logarithmic scale as log reduction values and β lies between 0.5 and 1. Those values are the key components in the acceptability specification. For the numerical example in the preceding paragraph, the expert chose $\delta = 1.0$ for the maximum acceptable discrepancy and $\beta = 0.9$ for the probability. The chosen δ and β could be different for different disinfectant test methods depending on the real-world conditions where the tested disinfectants will be applied.

© 2011 MSU Center for Biofilm Engineering

For some application situations, the stakeholder might require that the observed **LR** value is neither too small (negative discrepancy) or too large (positive discrepancy). In that case, the nearness criterion is that D should be between $-\delta$ and δ or $-\delta < D < \delta$. For other application situations, the stakeholder might decide that only one of the discrepancy directions is important. For example, a regulatory agency may be concerned only with the possibility that the observed **LR** is too high by more than δ. In that case, the criterion is $D < \delta$. Therefore, there are two cases to consider, the two-sided acceptability specification and the one-sided acceptability specification.

Pr{Event}

Specifications

2-sided:

$\Pr\{|D| \le \delta\} \ge \beta$

1-sided:

$\Pr\{D \le \delta\} \ge \beta$

or $\Pr\{-\delta \le D\} \ge \beta$

Let $\Pr\{Event\}$ denote the probability of the *Event*. For the two-sided case, the mathematical version of the acceptability specification is $\Pr\{|D| \le \delta\} \ge \beta$ or in words, "for the next disinfectant test result, a discrepancy (sign ignored) less than δ has probability of β at least." For the one-sided case, the acceptability specification is $\Pr\{D \le \delta\} \ge \beta$ when an upper limit is specified. When a lower limit is specified, the acceptability specification is $\Pr\{-\delta \le D\} \ge \beta$.

β-expectation tolerance interval

tolerance factor, *T*

After specific numerical values are chosen for β and δ, the statistician can convert the specification for the maximum acceptable discrepancy into a logically equivalent acceptability specification for S_R. Here is the way it works for the two-sided specification. In the field of statistics, an interval running from $-\delta$ to δ that satisfies the equation $\Pr\{|D| \le \delta\} \ge \beta$ is called a "β-probability prediction interval for D" and it is also called a "β-expectation tolerance interval for D." For convenience, we will use only the latter term. Using established statistical methods for forming a β-expectation two-sided tolerance interval, calculate a numerical quantity T such that $\Pr\{|D| \le (T \times S_R)\} \ge \beta$. The T denotes the numerical value of a β-expectation tolerance factor that can be calculated using special statistical techniques (e.g., Mee 1984); calculation details will not be presented here. Now the acceptability specification will be met if $T \times S_R \le \delta$ or equivalently, if $S_R \le \delta / T$. In other words, the β probability acceptability specification, $|D| \le \delta$, has been converted into an expression in terms of S_R, viz., $S_R \le \delta / T$.

Acceptability specification for S_R:

$S_R \le \delta / T$

For a one-sided specification, statistical methods for forming a β-expectation one-sided tolerance interval are used to calculate T. The numerical value for the 1-sided T differs from the 2-sided T. Using the 1-sided T, the one-sided acceptable discrepancy specification will be met if $(T \times S_R) \le \delta$ or equivalently, if $S_R \le \delta / T$.

Throughout this presentation, the discrepancy D is the random variable in fundamental probability statements. Although discrepancy is an important conceptual quantity for use in setting the β and δ specifications, the actual discrepancy value for any test is not observable (because the true LR, λ, is unknowable). However, the actual discrepancy is not important because the goal is to evaluate S_R and the critical calculations for S_R can be completed.

The stakeholder can use all available information, perhaps including historical data, to choose β and δ. Because those values depend on the intended use for the test method, different stakeholders may choose different values or some stakeholders may choose a one-sided specification while other stakeholders choose a two-sided

© 2011 MSU Center for Biofilm Engineering

specification. In fact, two different stakeholders could provide one-sided specifications that are in opposite directions. In the end, the S_R may be acceptable to one stakeholder, but not to the other. In some circumstances, stakeholders who hold conflicting views will negotiate mutually acceptable specifications.

Presenting collaborative study results

δ_{min}

Because the results of a collaborative study potentially will be used by different stakeholders and each stakeholder may choose unique acceptability criteria, there is merit in presenting results that are not directed at just one (β, δ) specification. After the study has been conducted and the realized numerical value of S_R is available, one can calculate the numerical value of $T \times S_R$ for any specified β. Let δ_{min} denote $T \times S_R$, the minimum value of δ for which the observed S_R is acceptable. The stakeholder will conclude that S_R is acceptable if the specified δ is as large as δ_{min} and if the stakeholder is satisfied with the β value used when calculating T. The stakeholder does not need to specify a single numerical value for δ, but needs only to decide whether the appropriate δ is as large as the realized δ_{min}; if so, the S_R is acceptable.

Most stakeholders probably will be interested in β values that are 0.80 or larger (Feinberg et al. 2010). We recommend that the collaborative study report includes a table of δ_{min} values for β = 0.80, 0.90, and 0.95 and for both one-sided and two-sided acceptability specifications (see the numerical examples below).

Numerical example using published collaborative study data

Numerical example

This example is based on the collaborative study results for a dried surface sporicide test of a disinfectant treatment that was a presumably high efficacy level of glutaraldehyde. The study comprised 3 replicate tests in each of 8 laboratories (Tomasino et al. 2008). The observed reproducibility SD was $S_R = 0.65$ which is below the historical acceptability threshold of 1.5 (see above) and the authors concluded that the test method exhibited acceptable reproducibility. For this study design, the pharmaceutical statisticians (e.g., Section 5 of Rozet et al. 2007; Feinbert et al 2010) suggest the β-expectation tolerance interval calculations of Mee (1984) for finding T. For these examples, we calculated T by programming Mee's technique in the R statistical computing language (http://www.r-project.org/), details not provided.

2-sided

Two-sided acceptability specification

Suppose β = 0.90 and δ = 1.0. This two-sided specification might be used by a standards setting organization. Using techniques from Mee (1984), the tolerance factor is $T = 1.7742$. Thus the acceptability specification for S_R is $S_R \leq 1.0/1.7742 = 0.56$. The observed S_R of 0.65 is too large to be acceptable to the (fictitious) standards setting organization.

1-sided

Upper one-sided acceptability specification

Consider an upper one-sided specification with δ = 1.0 and β = 0.90 (i.e., $\Pr\{D \leq 1.0\} \geq 0.90$). This specification might be used by a regulatory agency. The Mee (1984) one-sided calculations produce $T = 1.3636$. Thus the specification is that S_R must be no greater than 1.0/1.3636 = 0.73. For this one-sided specification, the observed S_R of 0.65 is sufficiently small to be acceptable to the (fictitious) regulatory agency.

© 2011 MSU Center for Biofilm Engineering

Table of δ_{min} values

Table of δ_{min} values

The Table below provides the T and δ_{min} values for one-sided and two-sided discrepancy specifications for each of three β probabilities. As in the preceding, T is calculated using the Mee (1984) method. For the one-sided specifications, the δ_{min} value is independent of the direction of the discrepancy specification, lower or upper. For a one-sided discrepancy specification and β = 0.90, the Table shows that T = 1.36 and the smallest δ for which S_R = 0.65 is acceptable is δ_{min} = 0.89 (= 1.36×0.65). For a 2-sided discrepancy specification and β = 0.90, the Table shows that T = 1.77 and δ_{min} = 1.15 (= 1.77×0.65).

Table. One-sided and two-sided T and δ_{min} values for each specified β when S_R = 0.65

	β		
	0.80	0.90	0.95
	one-sided specification		
T	0.88	1.36	1.77
δ_{min}	0.58	0.89	1.15
	two-sided specification		
T	1.36	1.77	2.15
δ_{min}	0.89	1.15	1.40

Conduct multiple tests and report the mean LR

Extension to multiple test protocols

M, N

Suppose that the collaborative study was conducted and the S_R was not sufficiently small to attain the specified β and δ. If the test is otherwise acceptable, a multiple test protocol can be devised that will shrink the S_R to meet specifications. A multiple test protocol requires that a disinfectant treatment must be tested multiple times; e.g., M separate tests in each of N laboratories. The mean of the M×N observed LR values is used as the multiple tests LR for the disinfectant treatment. That mean will have a smaller reproducibility SD than the S_R for a single test (M = 1 and N = 1). The collaborative study provides the information required to calculate that smaller SD. Let MTS_R denote the reproducibility SD for the multiple tests LR.

MTS_R

Numerical example showing how multiple testing improves reproducibility

For example, use the results of the collaborative study discussed above (Tomasino et al. 2008) and consider a multiple testing protocol in which L = 2 and M = 3 (3 tests in each of 2 laboratories). Conventional calculations show that MTS_R = 0.32 for the mean LR of the 6 tests (calculations not presented here; multiple testing will be the topic of a future *Knowledge Sharing Article*). For the two-sided acceptability specification of δ = 1.0 and β = 0.90, the acceptability limit for the reproducibility SD was 0.56 (see example above); the 6 test MTS_R of 0.32 is acceptably small. In fact, for β = 0.90, MTS_R = 0.32 would be acceptable even if δ was as small as 0.57 (= 1.77×0.32).

An acceptable reproducibility SD always can be accomplished by multiple testing. Statistical tools exist for finding the most efficient multiple testing protocol among protocols that produce an acceptable MTS_R. The most efficient protocol might require too many tests for routine use. Nevertheless, it usually will be informative to derive the optimum protocol for the stakeholder to consider.

© 2011 MSU Center for Biofilm Engineering

Conclusion: Recommend the acceptable discrepancy framework for assessing the reproducibility of a disinfectant test method

Conclusion

The acceptable discrepancy framework is logical, practical, statistically sound, and flexible. It does not dependent on historical decisions. We recommend the acceptable discrepancy framework for determining whether a disinfectant test method produced an acceptable collaborative study S_R.

References

Feinberg, M., Granier, G., and Mermet, J-M. (2010) Interpretation of interlaboratory trials based on accuracy profiles. *J. AOAC International* 93:725-733.

Hoffman, D. and Kringle, R. (2007) A Total Error Approach for the Validation of Quantitative Analytical Methods. *Pharmaceutical Research* 24(6):1157-1164.

Hubert, Ph., Nguyen-Huu, J.-J., Boulanger, B., Chapuzet, E., Chiap, P., Cohen, N., Compagnon, P.-A., Dewé, W., Feinberg, M., Lallier, M., Laurentie, M., Mercier, N., Muzard, G., Nivet, C., Valat, L. (2004) Harmonization of strategies for the validation of quantitative analytical procedures: A SFSTP proposal - part I. *J. Pharmaceutical and Biomedical Analysis* 36:579–586.

Mee, R. W. (1984) β-expectation and β-content tolerance limits for balanced one-way ANOVA random model. Technometrics 26:251-254.

Rozet, E., Ceccato, A., Hubert, C., Ziemons, E., Oprean, R., Rudaz, S., Boulanger, B., and Hubert, P. (2007) Analysis of recent pharaceutical regulatory documents on analytical method validation. J. Chromatography A 1158:111-125.

Tilt, N. and Hamilton, M.A. (1999) Repeatability & reproducibility of germicide tests: a literature review. *J. AOAC International* 82:384-389.

Tomasino, S. F., Pines, R. M., Cottrill, M. P., and Hamilton, M. A. (2008) Determining the efficacy of liquid sporicides against spores of *Bacillus subtilis* on a hard nonporous surface using the quantitative three step method: collaborative study. *J. AOAC International* 91:833-852.

Vardeman, S. B. (1992) What about the other intervals? *American Statistician* 46:193-197.

Revised: October 30, 2011
Original publication date: September 15, 2011

Primary author: Martin A. Hamilton, mhamilton@biofilm.montana.edu

© 2011 MSU Center for Biofilm Engineering