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-16

Multi-laboratory study design for assessing the reproducibility, repeatability and responsiveness of an antimicrobial test method

Al Parker
3/31/2021

[Key Words: antimicrobial agent, log reduction, performance standard]

Overview. For method validation purposes, an antimicrobial test method (ATM) should be assessed for *reproducibility, repeatability and responsiveness*. When testing an antimicrobial agent using an ATM, the *log reduction* (LR) is the quantitative outcome that measures efficacy (KSA-SM-7). For each test, the LR is typically calculated from recovered colony forming units (CFU) as

$$LR = [\text{control mean}(\log_{10}(CFU/carrier)] - [\text{treated mean } \log_{10}(CFU/carrier)].$$

The value for $\log_{10}$(CFU/carrier) is referred to the *log density* of microbes. The control mean log density is the *TestLD*. The LR can also be calculated for *semi-quantitative* and *qualitative* ATMs that produce positive/negative outcomes (KSA-SM-2, Hamilton et al. 2013). When using an ATM to test the same antimicrobial agent across different labs, the *TestLD*s and LRs will vary across the labs. This variability is quantified by a reproducibility SD. Even when testing the same agent in the same lab over independent test days the *TestLDs* and LRs will vary, quantified by a repeatability SD. If a multi-laboratory study of the ATM is performed, then the reproducibility and repeatability SDs can be estimated for the *TestLD* and LR (KSA-SM-10). These components of variance are critical to the implementation of the ATM for research and regulatory purposes: analysis of the control *TestLDs* quantifies consistency of the bio-challenge used by the ATM; analysis of the LRs quantifies consistency of the efficacy outcome when testing the same antimicrobial agent. The *responsiveness* of the ATM is quantified by comparing LRs from different efficacy levels of the same agent.

Multi-laboratory study design. A minimum of 3 participating labs is recommended, with at least 2 independent replicate test days at each lab. ASTM International recommends a minimum of 6 labs (ASTM E 691)[1] and 3 test days at each. Preferably, at least 3 different antimicrobial agents, with 2 different efficacy levels per agent, should be tested[2]. The 2 efficacy levels for each agent should be chosen so that: (1) a difference is expected in the resulting LR; (2) the mean LRs for all 6 agent/level combinations span a wide range, including a very low LR (not much kill) and a very high LR (almost complete kill). The 2 efficacy levels for each agent should be tested on the same test day at each lab with the order of testing randomized.

[1] resulting in a 67% increase in precision when estimating the mean LR at 95% confidence (i.e., using 5 degrees of freedom instead of 2).
[2] Including 3 agents addresses the potential criticism that an ATM is biased against a particular active ingredient(s); otherwise a single agent with 6 efficacy levels could be used.

Analysis of multi-laboratory data. To generate the reproducibility and repeatability SDs for the LR, a linear mixed effects model (LMM) is fit to the LRs separately for each agent and efficacy level combination, with a random effect for laboratory (KSA-SM-13, Hamilton et al. 2013). Across all the agents and efficacy levels, the reproducibility (and repeatability) SDs of an ATM is a non-linear, frown-shaped function of the mean LR (Parker et al. 2018). This means that highly ineffective and highly effective agents are expected to generate LRs with lower reproducibly SDs while moderately efficacious agents are much more variable with higher SDs. The multi-laboratory study design above allows for accurate estimation of the reproducibility (and repeatability) frown-shaped curve of SDs by fitting a quadratic regression to the 6 reproducibility variances versus the 6 mean log reductions[3] then square rooting. Repeatability and reproducibility SDs are generated for the control *TestLDs* by fitting an LMM to the control data with nested random effects for test day and laboratory. To quantify responsiveness when using the multi-lab design above, the LRs for the efficacy levels on each test day for each agent can be differenced then analyzed with an LMM with a random effect for lab. Residual plots are used to assess each LMM's fit to the data by: investigating potential outliers, confirming the constant variance assumption, and checking that the residuals approximately follow a normal distribution (Hamilton et al. 2013). This analysis is consistent with guidelines published by ASTM E691 and AOAC (2016). Explicit code using the software R (R Core Team 2020) for analyzing multi-laboratory data is available in KSA-SM-13.

Assessing performance standards. For registration purposes and depending on the specific claim, EPA requires that an ATM be used to test the agent against specified microbes in multiple independent tests. In each test, the agent must achieve a required LR. A performance standard (PS) specifies which microbes, the number of labs (usually 1), number of tests (usually 3 per microbe), and the required LR in each test. Any proposed PS can be assessed by two metrics (Parker et al 2014): (1) pass-error percentage[4] which is the percentage of ineffective antimicrobial agents that incorrectly pass the PS; (2) fail-error percentage[5] which is the percentage of highly effective antimicrobial agents that incorrectly fail the PS. Ideally, these two error percentages are minimized (e.g., less than 5%). The pass-error and fail-error percentages can be calculated for any proposed PS given the following inputs: (1) reproducibility and repeatability frown-shaped curves as a function of the mean LR from a multi-lab study; (2) LR specification defining ineffective antimicrobial agents; and (3) LR specification defining highly effective antimicrobial agents. Given the LR specification for ineffective and highly effective antimicrobial agents, the reproducibility and repeatability curves will predict the SDs, respectively, for these two types of antimicrobial agents[6]. Given these SDs, a multivariate *t*-distribution is applied within a conventional hypothesis testing approach to estimate the pass-error and fail-error percentages (with the null hypothesis defined by the LR specification for ineffective antimicrobial agents, and the alternative hypothesis defined by the LR specification for highly effective antimicrobial agents). Parker et al 2018 describe estimating the SDs from the reproducibility curves. Parker et al. 2014 provide a detailed description of the use of the multivariate *t* and provide R code for calculating the pass-error and fail-error percentages. In practice, PSs for a range of numbers of tests and required LRs are assessed, and candidate PSs are highlighted that achieve low error rates. This approach was applied by Tomasino et al. 2014 to assess and update the performance standards for the use dilution method.

References.

[3] There will be 6 means and SDs if the study is conducted with 3 agents, each at 2 efficacy levels. This will result in 3 degrees of freedom for error after estimating the regression coefficients for the parabola.

[4] Or "Type I" error; 100% – pass error % = confidence level = specificity = the percentage of ineffective antimicrobial agents that correctly fail the PS.

[5] Or "Type II" error; 100% – fail error % = statistical power = sensitivity = the percentage of highly effective antimicrobial agents that correctly pass the PS.

[6] Due to the frown shape of the curves, highly effective antimicrobial agents will tend to have lower variability.

AOAC, *Guidelines for Collaborative Study Procedures to Validate Characteristics of a Method of Analysis*, in *Official Methods of Analysis of AOAC International*. 2016.

ASTM standard E691, 2013. Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method. ASTM International, West Conshohocken, PA.

M. Hamilton, G. Hamilton, D. Goeres, and A. Parker. Guidelines for the Statistical Analysis of a Collaborative Study of a Laboratory Disinfectant Product Performance Test Method. JAOAC International 96(5):1138-1151, 2013.

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

A. Parker, M. Hamilton, and S. Tomasino. A Statistical Model for Assessing Performance Standards for Quantitative and Semi-quantitative Disinfectant Test Methods. *JAOAC International*, 97(1):58-67, 2014.

R Core Team, *R: A language and environment for statistical computing*. 2020, R Foundation for Statistical Computing: Vienna, Austria.

S. Tomasino, A. Parker, and M. Hamilton. Use of Statistical Modeling to Reassess the Performance Standard for the AOAC Use-dilution Methods (955.15 and 964.02) *JAOAC International*, 97(1): 68-77, 2014.