DISINFECTANTS

Should the AOAC Use-Dilution Method Be Continued for Regulatory Purposes?

NAVID OMIDBAKHSH
Virox Technologies, Research and Development, 2770 Coventry Rd, Mississauga, Ontario L6H 6R1, Canada

Despite its very poor reproducibility, AOAC INTERNATIONAL's use-dilution method (UDM) for bactericidal activity (AOAC Methods 964.02, 955.14, and 955.15) has been required by the U.S. Environmental Protection Agency (EPA) since 1953 for regulatory purposes, while methods with better reproducibility have been adopted in Canada and Australia. This study reviews UDM from a statistical perspective. Additionally, the test's expected results were compared to those obtained from actual evaluation of several formulations. Significant gaps have been identified in the reproducibility of the test data as predicted by statistical analysis and those presented to the EPA for product registration. UDM's poor reproducibility, along with its qualitative nature, requires the concentration of the active ingredient to be high enough to ensure all or most carriers to be free of any viable organisms. This is not in accord with the current trends towards sustainability, human safety, and environmental protection. It is recommended that the use of the method for regulatory purposes be phased out as soon as possible, and methods with better design and reproducibility be adopted instead.

AOAC INTERNATIONAL's Use-Dilution Method (UDM) was first introduced in the United States in 1953 (1, 2) for testing and registration of liquid chemicals as bactericides. Subsequently, the method was changed somewhat to address concerns with its basic design and the high variability in the data it yielded (3–8). However, many of those improvements have failed to address the fundamental flaws and shortcomings of the method; therefore, the product formulators have yet to deal with the poor reproducibility of these test methods while developing new products. Further, the Antimicrobial Testing Program (ATP), the U.S. Environmental Protection Agency's (EPA) post-registration regulatory system, evaluates the bactericidal effectiveness of EPA-registered disinfectants (9) using this method. ATP has found that nearly 30% of the registered disinfectants fail the test (10). This shows that either the registered products aren't robust enough, or the results based on these methods are not reliable.

This review critically examines UDM's basic design and those factors contributing to the lack of reproducibility and reliability of the test data.

Received February 2, 2011. Accepted by AH August 3, 2011.
Corresponding author's e-mail: navid@virox.com
DOI: 10.5740/jaoacint.11-015

Basic Design and Performance of UDM

According to the latest version of the method (11–13), stainless steel penicylinders with a relatively smooth surface are first dipped in a 48 h-old broth culture of the test bacterium. The inoculum is then dried, and each cylinder placed in a tube with 10 mL of the use-dilution of the test substance for the required contact time. The cylinders are deposited one by one in a neutralizer (primary subculture), placed in an incubator for 30 min, then transferred to a secondary tube of sterile medium; both tubes are incubated for 48 h at 36 ± 1°C. To generate data for product registration, 60 cylinders must be included for each of the three species of test bacteria, namely, *Pseudomonas aeruginosa, Staphylococcus aureus,* and *Salmonella enterica* (formerly known as *S. choleraesuis*). Since three manufacturer's lots of a given formulation are to be tested, the total number of cylinders for each such assessment for data submission is 540. With some exceptions, as noted below, each tube of the recovery broth must remain free of turbidity for a pass.

Certain aspects of the method are highly operator-sensitive. For example, the placement of the cylinders in the tubes of recovery broth requires much skill and steady hands. Slight variations in the growth conditions and processing, of *P. aeruginosa* in particular, can substantially affect the outcome of the test. Dipping the cylinders in the bacterial culture and their subsequent placement for drying have the potential to vary the bacterial loading, and thus, the degree of kill required on each cylinder.

Design Flaws

Because the surface of the penicylinders is relatively smooth, it does not allow for the sequestering of the bacterial cells, as would be expected in most field situations. Placing contaminated cylinders in tubes with 10 mL of the test substance results in a very high ratio of disinfectant volume to surface area of the carrier (295.3 mm^2), thus favoring the test substance further.

The apparent statistical power is built into the test through the use of a relatively large number (60 for each bacterial species) of carriers, but with the outcome based on a yes or no answer for bacterial growth, it is a frequent source of variability. Not long ago, carrier counts to ensure a minimum level of bacterial challenge became a requirement (7).

The UDM allows for a maximum of one positive out of 60 carriers as the passing criterion. Even though the AOAC performance criterion requires killing at least 59 of 60, and EPA mentions that this inactivation level provides a 95% confidence level, the question is why conducting post-registration tests results in a 30% failure rate (10). This study examines UDM from a statistical point of view and estimates its pass/fail probability.

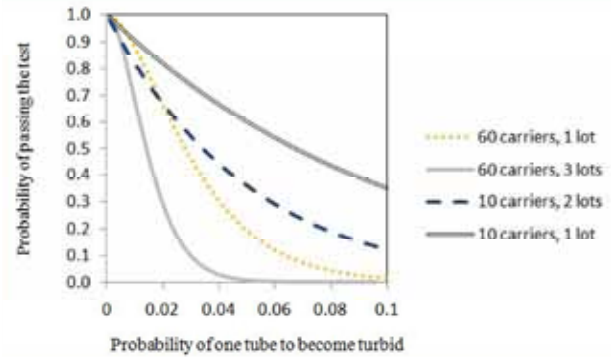

Figure 1. Probability of passing the test versus the probability of a single tube to fail.

The probability of failing a test passing the registration criteria was compared to the ATP failure rate. The expected failure probability does not conform to the ATP results, reinforcing the concerns enumerated here.

Statistical Evaluation of UDM

In the UDM, each tube of recovery broth is a Bernoulli trial (14) since it only has two outcomes (pass or fail), and the experiment with n test tubes follows a binomial distribution (15). EPA requires testing of a substance with claims against the three required types of test bacteria with three lots and 60 carriers each (16). Therefore, for the registration process, the probability of passing the test for a single type of the test bacterium will be:

$$P_a = [(1 - p_a)^{60} + 60p_a(1-p_a)^{59}]^3 \quad (1)$$

Let's imagine that we have a product that is tested against only one type of bacteria, with the probability of each test tube to fail, p. Figure 1 shows the probability of passing the test for different number of carriers and different lots.

For the bacteria other than these three, EPA only requires two lots, 10 carriers each. Therefore, to register a product, the probability of passing the test for a product with claims against three main types of bacteria and three others can be written as:

$$P_a = [(1 - p_a)^{60} + 60p_a(1-p_a)^{59}]^3[(1 - p_b)^{60} + 60p_b(1-p_b)^{59}]^3 [(1 - p_c)^{60} + 60p_c(1-p_c)^{59}]^3[(1 - p_d)^{20}(1-p_e)^{20}(1-p_f)^{20} \quad (2)$$

where p_a to p_f are the probability of failure of each tube against bacteria a to f, and a, b and, c are the three main bacteria, as mentioned above.

P_a can be shown versus p if all p values are known. In practice, p is not known and can be only guessed from the UDM results. For a hospital grade disinfectant, at least the three main types of bacteria, as mentioned above, must be tested; therefore, we will have:

$$P_a = [(1 - p_a)^{60} + 60p_a(1-p_a)^{59}]^3[(1 - p_b)^{60} + 60p_b(1-p_b)^{59}]^3 [(1 - p_c)^{60} + 60p_c(1-p_c)^{59}]^3 \quad (3)$$

p_a, p_b, and p_c values are chemistry dependent (17, 18) and cannot be assigned fixed numbers. To estimate the probability of a product to pass the test against the three main bacteria, we assign certain values for p_a, p_b, and p_c based on published literature. Quantitative efficacy evaluations (19) show less than one $\log_{10}$ difference in reduction in the viability of the three main test bacterial types against chlorine, quaternary ammonium compounds, and alcohols. Therefore, based on the similarities between the sensitivities of these organisms to disinfectants, the p_a, p_b, and p_c values are assumed to be equal here. If we consider a nonrobust product (with a probability of 50% to pass all its tests), by solving Equation 3, p is about 0.007.

A product is considered robust if it passes the test with high confidence, for example, 95% confidence. Solving Equation 3 results in a p of 0.002, meaning that if a product has a probability of failure of 0.002 or smaller for each tube, it will pass the test for the three main bacteria in at least 95% of the cases. If the product is tested against more bacterial species, then a p of smaller than 0.002 will be needed to result in 95% confidence. For example, if the three main test bacterial species and three others (*Escherichia coli,* Vancomycin-resistant *Enterococcus,* and methicillin-resistant *Staphylococcus aureus*) are tested, in order to pass all six tests with 95% confidence, the probability of a single tube to fail is about 0.00065, which is much smaller than that of the product with claims against the three main species. Table 1 shows the probability of having 0 to 7 positive tubes for these three cases. This shows that, even for such a nonrobust formulation (p = 0.007), which is an extreme case, there is 93 and 87% probability of passing a 60-carrier post-registration test for one and two bacterial species, respectively. (ATP testing is usually performed against two types of bacteria, one lot, 60 carriers each.)

$$p = 0.007\text{: } P_a = [(1 - p)^{60} + 60p(1-p)^{59}]^2 = 0.87$$

Most hospital grade disinfectants have label claims against at least six or seven types of bacteria, which further reduces the p value, and therefore, in theory, increases the probability of passing the product if tested again. Figure 2 shows the probability of a product passing against three main bacteria (three lots, 60 carriers for each bacterium), three main and three additional bacteria (three lots, 60 carriers for each main bacterium, and two lots, 10 carriers each for the three additional ones), and two bacteria (one lot, 60 carriers for each) similar to

Table 1. Probability of having different numbers of positive tubes in a single test with 60 carriers

No. of positive tubes (out of 60)	P = 0.00065	P = 0.002	P = 0.007
0	0.9617	0.8868	0.6561
1	0.0375	0.1066	0.2775
2	0.0007	0.0063	0.0577
3	0.0000	0.0002	0.0079
4	0.0000	0.0000	0.0008
5	0.0000	0.0000	0.0001
6	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000

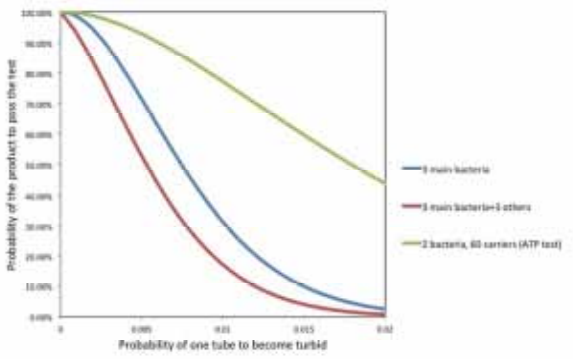

Figure 2. Probability of a product to pass the test versus the probability of one tube to fail, for different number of bacteria/carriers tested.

the ATP tests. As can be seen, the larger the number of bacteria and total carrier number, the lower the probability of passing for a product.

ATP results show that nearly 30% of the products fail (10), while the statistical analysis even for the very unlikely and worst-case scenario predict a failure rate of 13%. A robust product (95% confidence) with claims against the three main bacterial species (p = 0.002) is expected to fail the post-registration test (60-carrier test, two bacteria) with a probability of 1.3%; for a robust product with claims against six bacteria (p = 0.00065), there is a 0.16% probability to fail the same test.

Test Reproducibility

The concentration of the formulation needs to be fine-tuned in the product development phase. Therefore, a few screening tests are required before the product formulation is finalized.

Table 2 shows the probability of a test to result in 0 to 7 positive tubes versus p, the probability of one test tube to fail. This table shows that for a formulation with a p of 0.005, the probabilities of having 0/60, 1/60, 2/60, and >2/60 are 0.74, 0.22, 0.033, and <0.003, respectively. That is, it is very unlikely to have two independently run trials with results more than three or four apart. The data in the literature (1) show that identical samples (already registered disinfectants) tested by different analysts and different labs have very different numbers of positive tubes. For a sample tested with undiluted initial titer, one test has less than 10 positive tubes, while another shows more than 25, which is statistically very unlikely. Even though this discrepancy seems to be addressed by diluting the initial titers (1), current EPA regulations do not allow for any titer dilution. Therefore, such a modification to the method cannot be implemented.

Variability in the Initial Bacterial Counts

In the above calculations, it was assumed that the probability of each single carrier to pass/fail the test is constant; however, in practice this is not the case, because the initial microbial count varies for different carriers (1, 4, 7). The importance of this comes from the kinetics of disinfection models, where the final count is linearly dependent on the initial count given a constant contact time and disinfectant concentration. For example, in the Chick model (20):

$$N = N_0 e^{-kt} \tag{4}$$

where N and N_0 are the final and initial bacterial counts respectively, k is the disinfection rate constant, and t is the disinfection contact time. This equation shows that the final count is linearly dependent on the initial count; therefore, p is a function of initial count inverse.

If we include this variable in our calculations, we will have:

$$P_a = P_{0/60} + P_{1/60} \tag{5}$$

where P_a is the probability of a product to pass a test against one organism on 60 carriers, $P_{0/60}$ is the probability of having zero positives, and $P_{1/60}$ is the probability of having one positive tube.

$$P_{0/60} = (1-p_1)(1-p_2)\ldots(1-p_{60}) = \prod_{i=1}^{60}(1-p_i) \tag{6}$$

$$P_{1/60} = p_1(1-p_2)\ldots(1-p_{60}) + p_2(1-p_1)\ldots(1-p_{60}) + \ldots + p_{60}(1-p_1)\ldots(1-p_{59}) \tag{7}$$

This can be simplified to:

$$P_{1/60} = \prod_{i=1}^{60}(1-p_i)\left(\sum_{i=1}^{60}\frac{p_i}{1-p_i}\right) \tag{8}$$

and

$$P_a = \prod_{i=1}^{60}(1-p_i)\left(1+\sum_{i=1}^{60}\frac{p_i}{1-p_i}\right) \tag{9}$$

Table 2. Probability of positive carriers versus the probability of one tube to fail

	Probability of having 0 to 7 positive tubes in a test with 60 carriers							
p	0	1	2	3	4	5	6	7
0.001	0.9417	0.0566	0.0017	0	0	0	0	0
0.005	0.7403	0.2232	0.0331	0.0032	0.0002	0	0	0
0.01	0.5472	0.3316	0.0988	0.0193	0.0028	0.0003	0	0
0.025	0.2189	0.3368	0.2548	0.1263	0.0461	0.0133	0.0031	0.0006
0.05	0.0461	0.1455	0.2259	0.2298	0.1724	0.1016	0.049	0.0199

Table 3. Probability of passing a 60-carrier use-dilution test (given various *p* and σ)

Probability of observing a positive carrier, *p*	Probability of passing the test				
	σ = 0	σ = 0.29	σ = 0.5	σ = 0.6	σ = 0.7
0.001	0.9983	0.9974	0.9933	0.9891	0.9796
0.002	0.9934	0.9898	0.9756	0.9582	0.9261
0.003	0.9858	0.9787	0.9503	0.9199	0.8552
0.004	0.9757	0.9635	0.9200	0.8685	0.7853
0.005	0.9635	0.9458	0.8813	0.8156	0.7097
0.006	0.9493	0.9250	0.8442	0.7600	0.6146
0.007	0.9336	0.9030	0.7973	0.7020	0.5399

where p_1, p_2,..., p_{60} are the probability of a tube to become positive.

Here, we assume a normal distribution for the $\log_{10}$ count of the bacterial load on different carriers. Based on the literature (7), we assume an initial average $\log_{10}$ count of $N_0 = 6.7$, normally distributed with SD of 0.29. To see what difference this variability can make in the results of one or three bacteria tested in a confirmatory test, a Monte-Carlo simulation with 500 realizations is used. Several values for sigma are examined here to find the SD value at which significant difference in the results are observed. Table 3 shows the probability of passing a 60-carrier use dilution test, given various *p* and σ values. Based on the Chick model (Equation 4), the actual *p* for each tube will depend on its initial bacterial count. The second column (σ = 0) is based on the assumption that each of the 60 tubes will have the same number of initial counts and therefore, the probability distribution follows a binomial distribution, as illustrated above. As can be seen, for the actual reported cases (σ = 0.29, i.e., the variability of the titers of the counts in a laboratory has a SD value of 0.29 or smaller), the probability of passing the test is very close to the case where σ = 0, meaning that the counts' variability in use-dilution tests (σ ≤ 0.29) is not significant in the test results.

Table 3 shows that the probability of passing the test is diminished by increasing σ. For $p = 0.007$ (a nonrobust product), if σ = 0.6, the probability of passing the test will be about 20% less than that of σ = 0.29. For a $p = 0.002$, however, even at a high σ value (σ = 0.6), there is only 3.2% less chance to pass the test. Given the reported average SD of σ ≤ 0.29, it is expected that counts' variability has an insignificant impact on test results.

Conclusions

The UDM has been criticized in the past few decades for its high variability (1, 3, 5, 6). Here, we have reviewed this method from a statistical perspective. The results of the analyses show a significant disconnect between the probability of pass/fail estimated by statistical analysis and that of the post-registration ATP results. This can be due to several important factors, which have not been taken into account in the initial method design or its further improvements. It is known that there is a significant variability in numbers of test bacteria that adhere to the carriers (2), but it was shown here that this factor, although effective, does not account for most of the gap between ATP failure percentage and the statistical analysis failure prediction based on the test methodology. This emphasizes the already-known operator error factor even more. Slight changes in the procedure by the operator (1) can potentially introduce wide variations in the results. For example, how gently or vigorously the carriers are shaken to release the bacteria on them can affect the number of tubes of the recovery medium showing growth. Even a slight contact with the inside upper portions of the tubes during the placement of bacteria-loaded carriers can influence the results. These are some of the causes of operator error in performance of the method. Furthermore, it has commonly been seen (1, 6) that the number of positive tubes (for a 60-carrier test) can significantly vary for the same product under the same conditions. It was shown here that for a binomial distribution, such a high variability in the number of positives is almost impossible, and it is less likely to have more than a three or four positive tube difference in the product performance.

Besides high variability in the test results, the UDM has other disadvantages: (*1*) It cannot be a good representative of real life, especially for products with a high active evaporation rate, such as ethanol or isopropanol. In practice, the concentration of the active ingredient in the solution decreases rapidly when the product is exposed to the surface, due to the high surface-to-volume ratio, whereas in UDM, 10 mL of the disinfectant solution has much higher ratio of volume to surface; therefore, the evaporation ratio is totally different from that on actual surface disinfection, leading to unrealistic results. (*2*) The ratio of disinfectant to inoculum is very different from that in practice. (*3*) There is limited ability to include other environmental surfaces.

In light of these significant deficiencies, most importantly its unpredictable results, it seems neither scientific nor fair that the EPA continue to evaluate disinfectant formulations using the UDM method, despite the modifications made to it in the past several years. Besides frustration for product registrants, this approach has also been unfriendly to the environment. The current trend in the whole industry and around the globe is sustainability. In the chemical industry, sustainability translates into using less toxic and more environmentally friendly products. EPA has also introduced a program called "Design for the Environment" to encourage the industry to be more sustainable (21). Given the high variability in the test results, and very expensive and long EPA registration process, formulators may attempt to develop products using concentrations of active ingredients much higher than required in the formulations to address this high variability. This results in consumption of

more-than-required chemicals (higher carbon footprint in production phase) and greater release of chemicals to the environment, which is totally against the sustainability concept.

To address these deficiencies, EPA, AOAC, and their stakeholders are currently working on the possibility of adopting alternative Organization for Economic Co-operation and Development methods based on quantitative carrier tests. However, it is expected that regulatory agencies around the globe, and specifically EPA, would further accelerate the adoption of more accurate test methods such as quantitative methods (22, 23), which have already been extensively scrutinized and found to be much more reproducible than the current UDM.

Finally, it is well recognized that marginally effective formulations tend to show wider variations in the results, regardless of the type of test method used. However, such variability has a much greater impact on repeat testing using a method such as UDM, which has a higher risk of failure of a registered product under ATP. On the other hand, a quantitative test protocol, although more stringent, may provide a greater level of confidence in the data with a set product performance criterion in $\log_{10}$ reductions rather than a simple pass/fail measure.

References

(1) Arlea, C., King, S., Bennie, B., Kemp, K., Mertz, E., & Staub, R. (2008) *J. AOAC Int.* **91**, 152–158

(2) Tomasino, S.F., Fiumara, R.M., & Cottrill, M.P. (2006) *J. AOAC Int.* **89**, 1629–1634

(3) Alfano, E.M., Cole, E.C., & Rutala, W.A. (1988) *J. Assoc. Off. Anal. Chem.* **71**, 868–871

(4) Cole, E.C., Rutala, W.A., & Carson, J.L. (1987) *J. Assoc. Off. Anal. Chem.* **70**, 903–906

(5) Cole, E.C., Rutala, W.A., & Samsa, G.P. (1987) *J. Assoc. Off. Anal. Chem.* **70**, 635–637

(6) Cole, E.C., Rutala, W.A., & Samsa, G.P. (1988) *J. Assoc. Off. Anal. Chem.* **71**, 1187–1194

(7) Tomasino, S.F., Pines, R.M., & Hamilton, M.A. (2009) *J. AOAC Int.* **92**, 1531–1540

(8) Cole, E.C., & Rutala, W.A. (1988) *J. Assoc. Off. Anal. Chem.* **71**, 9–11

(9) *Antimicrobial Testing Program, Pesticides: Regulating Pesticides* (2009) U.S. Environmental Protection Agency, http://www.epa.gov/oppad001/antimicrobial-testing-program.html, accessed in May 2011

(10) Bond, S., Curley, G., Dorsey, J., Harris, J., & Joseph, L. (2009) *Results of Hotline Complaint Review of EPA's Antimicrobial Testing Program*, Report No. 09-P-0152, U.S. EPA, Washington, DC

(11) *Testing Disinfectants Against Pseudomonas aeruginosa, Use-Dilution Method* (2009) AOAC INTERNATIONAL, Gaithersburg, MD, Method **964.02**

(12) *Testing Disinfectants Against Staphylococcus aureus, Use-Dilution Method* (2009) AOAC INTERNATIONAL, Gaithersburg, MD, Method **955.15**

(13) *Testing Disinfectants Against Salmonella choleraesuis, Use-Dilution Method* (2006) AOAC INTERNATIONAL, Gaithersburg, MD, Method **955.14**

(14) Trout, J.R. (1985) *J. Assoc. Off. Anal. Chem.* **68**, 763–765

(15) Montgomery, D.C., & Runger, G.C. (2003) *Applied Statistics and Probability for Engineers,* 3rd Ed., John Wiley & Sons, Inc., New York, NY

(16) *Efficacy Data Requirements, Disinfectants for Use on Hard Surfaces* (1982) http://www.epa.gov/oppad001/dis_tss_docs/dis-01.html, accessed in April 2011

(17) Merianos, J.J. (2001) in *Disinfection, Sterilization and Preservation*, 5th Ed., Lippincott Williams & Wilkins, Philadelphia, PA, Chapter 14, p. 306

(18) Block, S.S. (2001) in *Disinfection, Sterilization and Preservation*, 5th Ed., Lippincott Williams & Wilkins, Philadelphia, PA, Chapter 9, p. 187

(19) Rutala, W.A., Barbee, S.L., Aguiar, N., Sobsey, M.D., & Weber, D.J. (2000) *Infect. Control Hosp. Epidemiol.* **21**, 33–38. http://dx.doi.org/10.1086/501694

(20) Chick, H. (1908) *J. Hyg.* **8**, 92–157. http://dx.doi.org/10.1017/S0022172400006987

(21) *Design for the Environment*, U.S. EPA Partnership Program (2011) http://www.epa.gov/dfe, accessed in May 2011

(22) *Standard Quantitative Carrier Test Method to Evaluate the Bactericidal, Fungicidal, Mycobactericidal, and Sporicidal Potencies of Liquid Chemical Germicides* (2005) ASTM E2111-05, ASTM International, West Conshohocken, PA

(23) *Quantitative Disk Carrier Test Method for Determining the Bactericidal, Virucidal, Fungicidal, Mycobactericidal, and Sporicidal Activities of Liquid Chemical Germicides* (2011) Document E-2197, ASTM International, West Conshohocken, PA