JOURNAL OF CLINICAL MICROBIOLOGY, June 1980, p. 541-545
0095-1137/80/06-0547/05$02.00/0

Vol. 11, No. 6

Reproducibility of Serological Titers

ROSS J. WOOD* AND THENA M. DURHAM
Center for Disease Control, Atlanta, Georgia 30333

Serologists are making increasing use of the term "reproducibility" when referring to the reliability or repeatability of serological tests. The practical use of the concept, however, has been limited by the absence of an appropriate numerical scale on which differing reproducibilities can be quantified and objectively compared. This limitation can be overcome by adopting the proposed quantitative measure of reproducibility. The recommended measure is a natural extension of the common practice of considering a serological test to be acceptably reproducible so long as replicate titers remain within a twofold range. The measure can be readily used in the field of serology, and examples are given of how this can be done.

Measurement methods are continually studied in the laboratory sciences. As a minimum, clinical chemists and serologists must remain aware of the level of reliability or repeatability of their laboratory measurement methods. Clinical chemists have adopted the inverse of the standard deviation (precision) and the coefficient of variation (percent error) as handy indices of measurement method repeatability because the essence of repeatability is reflected in the variance or standard deviation of the measurements. In contrast, serologists seldom refer to repeatability in terms of precision or percent error; instead, they tend to refer to it as reproducibility. For serologists, "reproducibility" may connote the actual measurement process more closely than "precision" and "percent error"; the word itself emphasizes that interest is centered in producing measurements subject to minimal variation.

Through increasing use, chemists have developed and implemented precision and percent error as useful quantitative measures. In the field of serology, however, the analogous potential of the reproducibility concept has not been exploited. Perhaps the first positive step was taken a decade ago when Hall and Felker (1) suggested a working definition of reproducibility in the serological laboratory. This suggestion, however, has been largely ignored. In this paper, their suggestion serves as the starting point for a proposed quantitative measure of the reproducibility of serological titers.

Hall and Felker suggested that serum titer reproducibility be defined as "... the percentage of duplicate assays that yield the same titers or titers differing by only one dilution." In our modified version of this definition, serial dilutions used in assaying a serum are based on a dilution ratio of 2 (other dilution ratios will not be considered at this time), and serum titer reproducibility is defined as the probability that the maximum ratio of two distinct titers (obtained in the blind) on the same serum specimen will not exceed 2.

This definition permits reproducibility to be dependent on the individual specimen being assayed. It allows the possibility that the reproducibility of a test method may be different for each specimen. This flexible definition is consistent with the hemagglutination inhibition (HI) serum titer data presented later. These data show that reproducibility may decrease with increasing HI serum titers. This is not surprising because of two related points. First, the defined probability certainly diminishes as the standard deviation (of replicate titers on the same specimen) increases. Second, a rather common (but not universal) measurement phenomenon is that the standard deviation increases as the value being measured increases.

The purpose of this paper is to propose and advocate a particular quantitative measure of the repeatability of serological tests. This purpose could be served by using only artificial data in the examples. However, we decided to use real laboratory data from a designed study using the arbitrarily chosen FIAX test in order to provide a note of realism.

MATERIALS AND METHODS

Preselected, healthy human donors volunteered sera that were either HI positive or HI negative for rubella virus antibodies as determined by replicate testing in the Center for Disease Control rubella hemagglutination test (2). An automated fluoroimmunoassay system (FIAX, International Diagnostic Technology, Santa Clara, Calif.) was used by a single laboratorian, in accordance with the manufacturer's instructions, to assay 16 sera in three separate runs per day for 30 days within a 2-month period. The sera

were initially divided into 30 sets of 16 sera each, and a different set was used each day. All sera were stored at −70°C.

A 1:41 dilution of each sample (sera and FIAX calibrator) was made with an automatic pipetter-diluter, and the dilutions were incubated with dual-surface (rubella virus-coated/control) polymeric immunoabsorbent samplers. The samplers were then washed and stained with fluorescein isothiocyanate-labeled goat anti-human globulin (polyvalent). After a final wash, the fluorescence emission of each sampler was read in a fluorometer interconnected to a microcomputer that printed out the digital HI-analogous rubella titers. Because of the automated nature of this serological technique, blind coding of the specimens was not considered necessary. The same lot of rubella FIAX test kits was used throughout the study.

RESULTS

For each run with the FIAX system, a calibration curve is determined from results obtained with three calibrators. Antibody titers of the test specimens are obtained by interpolation from this curve. Consequently, titers do not come truncated (in this paper, truncated titer [TT] refers to a titer obtained by reducing an integer titer to the nearest lower integer power of 2) to the next lower integer power of 2 as do conventional HI serum dilution titers; they are free to take on any of a continuum of values.

The proposed method of estimating the reproducibility of a serological test is demonstrated with data from the 30 afternoon runs for 2 sera selected from the 16 included in the designed study. For the first two examples, the data used were obtained by truncating the test kit titers to the next lower integer power of 2 to simulate conventional HI serum dilution titers. In the third and final example, the data from the second example are used in the form of the actual untruncated test kit titers rounded to integers.

Since the titers for the two sera used in these examples were obtained on 30 different days, the results computed can only estimate among-day reproducibility in contrast to among-run-within-day or within-run reproducibility. Estimates of these latter two would require data having the proper among-run or within-run structure.

Example 1. The data in Table 1 are the TTs obtained for a single serum specimen in the 30 afternoon runs. The first set of 10 TTs was obtained on the first and each successive third day of the 30-day period. Approximately 5 days elapsed between each 2 adjacent days in this sequence. The second set of TTs was obtained on the second and each successive third day, and the third set was obtained on the third and each successive third day of the 30-day period. The data are split into these three independent subsamples of 10 to show variability among independent estimates of the unknown reproducibility parameter.

In computing the estimate of reproducibility, the sample relative frequencies are used as though they are the probabilities of observing the corresponding TTs on repeated assays of the same serum specimen. The estimate is computed as the probability of observing a pair of TTs whose maximum ratio does not exceed 2, that is, whose positive difference in $\log_2$ (TT) does not exceed 1. This can be done by listing the possible ways of selecting (with replacement) a pair of $\log_2$ (TT) from the observed sample frequency distribution. This list for set 1 from Table 1 and the associated probabilities are given in Table 2.

The nine possible ordered pairs are listed in the second column of Table 2. The seven pairs numbered 1, 2, 3, 6, 7, 8, and 9 each have a positive difference in $\log_2$ (TT) that does not exceed 1. The estimate of among-day reproducibility is the sum of the probabilities associated with these seven pairs. The estimate is $(25 + 15 + 15 + 9 + 6 + 6 + 4)/(10)^2 = 80/100 = 0.80$.

Because the sum of the probabilities in Table 2 is 1, the estimate of reproducibility can also be obtained from the probabilities associated with those pairs whose positive difference in $\log_2$ (TT) exceeds 1. In this example, only pairs 4 and 5 qualify, and the estimate of reproducibility is $1 - [(5 \times 2 \times 2)/(10)^2] = 0.80$. This is usually the

TABLE 1. *Frequency distribution of TTs obtained on 30 distinct days with a single serum specimen: example 1*

TT	$\log_2$ (TT)	Set 1	Set 2	Set 3	Total
64	6	0	0	1	1
128	7	5	5	4	14
256	8	3	4	3	10
512	9	2	1	2	5
Frequency total		10	10	10	30

TABLE 2. *All possible ordered pairs of $\log_2$ (TT) selected from the set 1 sample frequency distribution of Table 1*

Pair no.	$\log_2$ (TT) ordered pair	Probability of selecting the pair
1	7 and 7	$(5/10) \times (5/10) = 25/(10)^2$
2	7 and 8	$(5/10) \times (3/10) = 15/(10)^2$
3	8 and 7	$(3/10) \times (5/10) = 15/(10)^2$
4	7 and 9	$(5/10) \times (2/10) = 10/(10)^2$
5	9 and 7	$(2/10) \times (5/10) = 10/(10)^2$
6	8 and 8	$(3/10) \times (3/10) = 9/(10)^2$
7	8 and 9	$(3/10) \times (2/10) = 6/(10)^2$
8	9 and 8	$(2/10) \times (3/10) = 6/(10)^2$
9	9 and 9	$(2/10) \times (2/10) = 4/(10)^2$
		$100/(10)^2 = 1$

simplest way to compute the estimate of reproducibility.

The estimate is interpreted as follows. The sample of 10 replicate TTs, shown as set 1 in Table 1, contains information about the among-day reproducibility of the test system with this specimen. The information indicates that if the specimen were assayed on 2 distinct days, there would be approximately an 80% chance that the two TTs would not differ by more than a factor of 2; conversely, there would be approximately a 20% chance that the TTs would differ by more than a factor of 2.

The simplest way to estimate the reproducibility is to focus on the pairs whose $\log_2$ (TT) differences exceed 1. For set 2 in Table 1, the estimate is $1 - (5 \times 1 \times 2)/(10)^2 = 0.90$. Note that this second independent estimate is higher than the first estimate.

For set 3 in Table 1, 16 ordered pairs are possible (6-6, 6-7, 7-6, 6-8, 8-6, 6-9, 9-6, 7-7, 7-8, 8-7, 7-9, 9-7, 8-8, 8-9, 9-8, and 9-9). Six of these (6-8, 8-6, 6-9, 9-6, 7-9, and 9-7) have a difference in $\log_2$ (TT) that exceeds 1. Thus, the estimate of among-day reproducibility from these data is $1 - 2 \times [(1 \times 3) + (1 \times 2) + (4 \times 2)]/(10)^2 = 0.74$. This independent estimate is lower than the first two.

If the results from the three sets for all 30 days are combined, a fourth estimate, based on more data than either of the first three estimates, can be computed. This estimate can be expected to be somewhat closer to the real unknown reproducibility of the test system with this specimen. The combined data from the final column of Table 1 lead to the consolidated estimate $1 - 2 \times [(1 \times 10) + (1 \times 5) + (14 \times 5)]/(30)^2 = 0.811$.

Example 2. The example in Table 3 is intended to provide additional experience with the computation procedure for the proposed estimate of reproducibility. This example will be followed by a final, more general example which should fully acquaint the reader with this computation method. All assay results are from the study previously described. They differ only in having been obtained with different specimens. As in Table 1, the data in Table 3 have been subdivided into three 10-day sets.

With the previous estimating method, the reproducibility estimates for the three independent sets of data in Table 3 are 0.90, 0.96, and 0.92, respectively. The estimate using all 30 data points is $1 - 2 \times (3 \times 11)/(30)^2 = 0.926$.

The composite data in Table 1 have a geometric mean TT of approximately 199 and a reproducibility estimate of 0.811. The corresponding statistics for the second specimen from Table 3 are approximately 19 and 0.926. The difference in these two reproducibility estimates reflects, in addition to sampling error, a higher variability inherent in assay results for specimens with higher levels of antibody. Evidence of this difference in variability can be seen by contrasting the variation among the three subestimates from Table 1 with that from Table 3. These estimates are 0.80, 0.90, and 0.74 for the high-titered specimen in Table 1 and 0.90, 0.96, and 0.92 for the low-titered specimen in Table 3. The greater variation in the first three estimates is due to a greater inherent variability among repeated assay results for this specimen. Not only is this increased variability associated with a lower reproducibility parameter, it also leads to greater variability in sample estimates of that parameter. At the same time, this variability decreases as the number of replicate titers on the same specimen increases. This subject will be pursued more specifically in a future report.

TABLE 3. *Frequency distribution of TTs obtained on 30 distinct days with a single serum specimen: example 2*

TT	$\log_2$ (TT)	Set 1	Set 2	Set 3	Total
8	3	1	1	1	3
16	4	4	7	5	16
32	5	5	2	4	11
Frequency total		10	10	10	30

Example 3. Data for the final example are given in Table 4. They are the composite data from Table 3, except that they are the actual untruncated test kit titers rounded to integers. From Table 4 the number of pairs whose maximum ratio exceeds 2 can be obtained as follows. The number of such pairs with 10 as the lowest integer titer (LIT) is 1×26, the number with 14 as the LIT is 2×14, the number with 19 as the LIT is 1×4, etc. The first number in each product is the frequency observed for that LIT. The second number is the total of those observed frequencies of integer titers that exceed twice the subject LIT. These results for the example in Table 4 are given in Table 5.

From Table 5 the estimate of among-day reproducibility of the integer titers for this specimen is $1 - (2 \times 68)/(30)^2 = 0.848$, which is lower than the estimate of 0.926 from the corresponding 30 TTs in Table 3. This difference results from the fact that of the $(30)^2 = 900$ possible ordered pairings of the 30 titers, only 2×33, or 66, of them had a maximum ratio that exceeded 2 after truncation, whereas 2×68, or 136, met this condition before truncation. The difference consists of those 70 ordered pairs for which the two characteristics of $\log_2$ (integer titer) differed

TABLE 4. *Integer test kit titers obtained on 30 distinct days with a single specimen*

Integer titer	Frequency	Integer titer	Frequency
10	1	29	2
14	2	31	1
19	1	32	2
21	1	33	1
22	3	34	1
23	2	36	2
25	2	37	1
26	1	39	2
27	1	40	1
28	2	56	1

TABLE 5. *Numbers of pairs of integer titers (total of 68) from Table 4 having a maximum ratio greater than 2*

Lowest titer of the pair	No. of pairs	Lowest titer of the pair	No. of pairs
10	1 × 26	25	2 × 1
14	2 × 14	26	1 × 1
19	1 × 4	27	1 × 1
21	1 × 1	28	2 × 0
22	3 × 1		
23	2 × 1		

by exactly 1 and the mantissa for the larger member of the pair exceeded that of the smaller member. Because of this, for a given set of serum titers the estimate of reproducibility based on the truncated titers will, in general, be higher than the estimate based on the untruncated titers.

DISCUSSION

Since the time of Karl F. Gauss (1777–1855), the most commonly used quantitative index of measurement repeatability has been the standard deviation of repeated measurement results. This ubiquitous measure is, however, not the only such index that might be used, nor is it necessarily the most appropriate one for all applications.

Bypassing this most common measure of repeatability, it is proposed that serologists use an alternative measure which is to be called "reproducibility." For any distribution of measurements, this useful measure of the repeatability of serum dilution assays is related to the standard deviation. The specific nature of this relationship, together with implications of practical importance to serologists, will be discussed in a future report.

Among-day reproducibility is a quantitative index of a very important aspect of any serum assay method. It directly reflects the credibility to be attached to any lone titer result, that is, the degree to which the reported titer depends on the day it is obtained. There are situations, however, in which the among-day reproducibility is not of primary importance. One such situation arises when acute-phase and convalescent-phase specimens from the same patient are assayed together in the same run as an aid in deciding whether serum conversion has taken place. Here, it is the within-run reproducibility that is of concern because it reflects the degree of reliance that can be placed on the difference of two titers obtained in a single run.

The proposed estimate of reproducibility can be used to estimate among-day reproducibility, among-run-within-day reproducibility, and within-run reproducibility. One only needs a set of replicate serum dilution titers, collected (in the blind) in accordance with a design appropriate to the particular reproducibility to be estimated. When the reproducibility of serum dilution titers depends on the titer level of the specimen, an estimate of this reproducibility based on single pairs of titers from each of any number of different specimens would only be advisable if all of the specimens had essentially the same antibody level.

Given the appropriate replicate titers for a specimen, the estimate of serum dilution titer reproducibility is most efficiently computed as $1 - 2 \times W/N^2$, where W is the sum of the products of the observed frequencies associated with those titers that can be placed in pairs having a maximum ratio of titers exceeding 2 and N is the total number of replicate titers of the subject specimen. If the reproducibility parameter decreases as the level of antibodies increases, then (with a fixed number of replicate titers) the variability of the estimate of the parameter increases. For a given level of antibodies, the variability in the estimate decreases as the number of replicate titers increases. This algorithm for estimating serum dilution titer reproducibility can be readily automated, even on today's smaller computers.

The proposed reproducibility measure reflects the degree to which a test repeats itself on a given specimen. In contrast to this, sensitivity measures the ability of a test to correctly identify positive specimens, whereas specificity measures the ability to correctly identify negative specimens. Consequently, these measures depend upon reproducibility in addition to the degree to which the test results average near the correct target values. Thus, sensitivity and specificity are somewhat broader measures than reproducibility. Their assessment, however, requires

specimens whose state of positivity is known. This requirement is frequently difficult to satisfy.

LITERATURE CITED

1. **Hall, E. C., and M. B. Felker.** 1970. Reproducibility in the serological laboratory. Health Lab. Sci. **7**:63–68.
2. **Palmer, D. F., J. J. Cavallaro, and K. L. Herrmann.** 1977. A procedural guide to the performance of rubella hemagglutination inhibition tests. U. S. Department of Health, Education, and Welfare, Center for Disease Control, Atlanta, Ga.