ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
23 (2000) 671–686

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Structured use of the median in the analytical measurement process

Jeffrey D. Hofer *, John R. Murphy

Eli Lilly and Company, Lilly Corporate Center, Drop Code 3834, Indianapolis, IN 46285, USA

Accepted 18 April 2000

Abstract

In the pharmaceutical industry, the process of measuring a product's attributes can be very complicated and the potential for an analytical mistake can be quite high. Often, an unexpected result leads to an investigation to assess the possibility that a mistake was made in the laboratory. Traditionally, the data generated in these investigations has been used, along with various outlier tests, to attempt to negate the original data. Sometimes, historical estimates of the S.D. of the analytical method are not available for use in outlier testing and the power of the outlier tests to detect true mistakes without such historical estimates is often very low due to the small amount of data available. This leads to a great deal of inconsistency in the amount of data that is further generated and how the data is ultimately handled in making a decision. Recently, FDA demands for consistent and objective laboratory investigations have raised concerns about these practices. An alternative approach, involving a systematic investigation strategy and data handling via the structured use of the median, is proposed in this paper. The operating characteristics of the traditional and proposed approaches are compared to show their similarity and the advantages of the proposed approach. It is strongly believed by the authors that the structured use of the median will lead to more consistent investigations and data handling, which will benefit industry, the FDA and ultimately, the consumer, by allowing more accurate decisions to be made more efficiently.

Keywords: Median; Out of specification; Retest; Outlier; Barr decision

1. Introduction

In the pharmaceutical industry, the process of measuring a product's attributes can be very complicated and time-consuming and there is always the potential for an analytical mistake. The US Pharmacopeia (USP) has recognized this and suggests outlier tests for the purpose of identifying outliers in bioassays and microbiological assays [1]. However, procedures for investigating whether an analytical mistake has occurred are often inconsistent and subjective. For example, one approach is the practice of generating additional test results that, if they provide 'normal-looking' results, will allow the use of a statistical outlier test to eliminate the aberrant result [2–7]. This approach has the potential of being applied subjectively, with little or no consistency across

* Corresponding author.

0731-7085/00/$ - see front matter
PII: S0731-7085(00)00349-6

time and/or location. In this paper, we propose a structured use of the median to investigate the possibility of an analytical mistake and to handle all of the data generated. We will demonstrate that this approach is simple, consistent and objective.

2. Investigational process

How does one determine that an analytical mistake may have occurred? Consider a situation such as that illustrated in Fig. 1, showing a plot of the potency results for product *X* for the last 31 assay runs. The statistical process control limits for the product are 97.0–103.0. Excluding the last result, the process appears to be capable of producing potency results well within those limits.

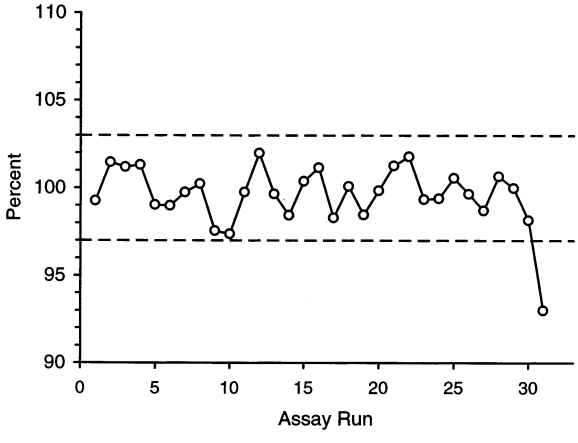

Fig. 1. Plot of last 31 Lot results.

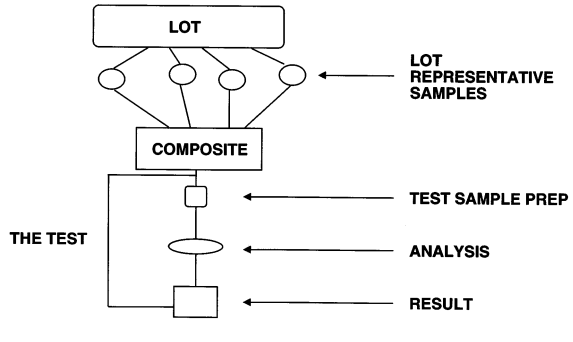

Fig. 2. Schematic of the test.

The last result, 93.0, was unusually low. The question is 'has an analytical mistake occurred, or does this sample truly have a potency of 93.0?' A consistent, simple and objective procedure is needed to address this question.

The first step of any investigative procedure is to check for an assignable cause, an analytical mistake which would explain the unusual initial result. Some examples are:

1. Use of an incorrect analytical procedure
2. Failure to calibrate the instrument
3. Use of an out-of-date standard or the wrong standard
4. Use of an incorrect standard potency
5. A calculation error

If an assignable cause can be verified and documented, the initial result would be invalidated. If the calculation can be corrected, as in 4 or 5 above, the result would be recalculated. If the mistake cannot be corrected, the testing process must be repeated.

When no assignable cause for the unusual result can be determined, further investigation is warranted. With our proposed methodology, the next step is to determine whether it is probable that an analytical mistake has occurred. The investigative procedure requires a thorough understanding of the testing process. As used in this document, the term test will refer to all steps in obtaining a sample result. In Fig. 2 below, a test typical of many HPLC content assays is illustrated. Multiple samples (tablets, capsules or vials) are randomly obtained from the lot and are combined to form an homogeneous composite. The analytical part of the test consists of two levels:

1. The preparation level (e.g. weighing, dilution, etc. of samples and standards)
2. The analysis level (e.g. injection of solutions on the HPLC, analysis of solutions by Autoanalyzer, etc.)

In order to determine whether an analytical mistake is probable and to isolate where it may have occurred, the investigation begins with the analysis level and follows with the preparation level if necessary. A reanalysis is defined as a reevaluation of the original prepared samples or

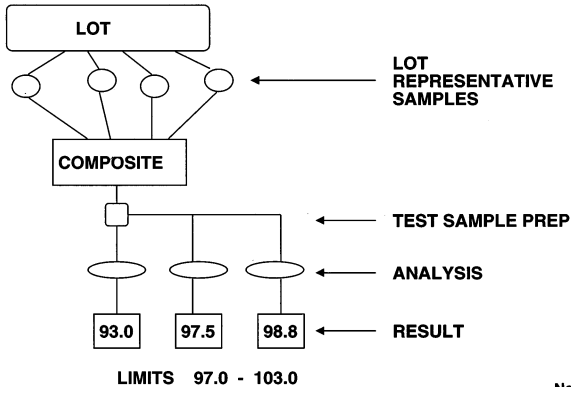

Fig. 3. Reanalyses schematic.

sample solutions (e.g. reinjection of standard and sample solutions on HPLC to investigate the possibility of an instrumentation mistake or malfunction). A retest consists of performing the entire test again (including a new standard curve) on new preparations from the same homogeneous composite sample.

The general investigation philosophy can be quite flexible. However, in specific applications, some questions and issues need to be resolved beforehand in order to promote consistency and objectivity. For example,

1. Should the sample and standard solutions be reanalyzed? If so, how many reanalyses should be performed?
2. If the additional results contradict the initial result, how should the data be combined?
3. Should new standards be prepared and/or should new sample preparations be prepared from the original sample(s)? If so, how many sample preparations should be made?
4. If these results also contradict the original result, how should the data be combined?

In discussing the recommended approach, we consider first the steps applicable at the analysis level. We then turn to a consideration of issues at the preparation level.

If sample and standard solutions are stable and still available, a reanalysis may be performed. Here, one is looking for mistakes such as a defective injection. We propose that the sample and standard solutions be reanalyzed twice to provide two results in addition to the original result, as illustrated in Fig. 3.

In this illustration, the two reanalysis results appear to contradict the original result. In a situation such as this, an outlier test — such as Dixon's test or the extreme studentized deviate (ESD) test (see Refs. [4–6] and Appendix A) — is often applied in an attempt to identify the original result as an outlier and to set it aside from further calculations. There are several drawbacks to the use of outlier tests in this setting. For one, unless a historical estimate of the relevant S.D. is available, the power of outlier tests is low for small sample sizes, which means that obvious blunders or grossly spurious results will sometimes fail to be tagged as outliers. These historical S.D. estimates are not always available. In addition, the choice of outlier test and significance level can lead to ad hoc and subjective testing procedures. Further, when the initial outlier test fails to invalidate a result, there is the potential of continuing the analytical and statistical testing processes until there is sufficient data (adequate power) to reject the original result. Finally, there is always the issue of whether a statistical outlier is necessarily erroneous.

The disadvantages and drawbacks of outlier testing demonstrate the need for an alternative that is more objective and consistent. We propose a procedure that computes the median of all results. This approach offers several advantages over outlier testing. First, the calculations are simpler. The median of a set of ranked values is the middle one if there are an odd number of values, and is midway between the middle two if there are an even number of values. More importantly, the use of the median provides a consistent and objective approach, since the need for ad hoc and/or subjective application of statistical outlier testing is eliminated because of the superior ruggedness of the median as a statistic. Finally, all data are used to compute a result except those invalidated due to assignable causes.

Returning to the example shown in Fig. 3, the original result was 93.0 and the two reanalysis results were 97.5 and 98.8, respectively. Given these results and the previous product history, one could infer that the unusually low value of 93.0

was caused by an unidentifiable analytical mistake. The real problem is not that the mistake occurred but, rather, what to do about it in order to come to a logical and rational decision on the analytical result. Unless an historical estimate of the S.D. is available and utilized, the unusual result is not identified as an outlier by the extreme studentized deviate test at the 5% level of significance. Since nothing was invalidated or eliminated by outlier testing, a combined result would normally be obtained by averaging the three results, giving 96.4 which falls below the limit of 97.0. This illustrates a dilemma often encountered with outlier testing: either to assign a result influenced by a likely analytical mistake; or to perform additional testing to attempt to overcome the suspected analytical mistake. In contrast, the median result would be 97.5, which is above the limit and provides a result which is more representative of the likely true sample potency based on past product history. By using the median, the need to perform further testing on this sample has been eliminated.

Now, imagine a scenario where it is not possible to perform a reanalysis, or where no apparent discrepancies are found by reanalysis. One should then test for the possibility of mistakes in, say, weighing or dilution at the preparation level by performing a retest. A retest involves independent standard preparations and new sample preparations from the original sample. The original sample material is often, but not always, a composite. We propose that two independent retests be performed. Fig. 4 depicts an example where a reanalysis was conducted and provided no evidence that an analysis error had occurred. In this case, the two reanalysis results were 93.2 and 93.6, which confirm that the original value of 93.0 was not the result of a problem with analysis of the original prepared solutions. Suppose, now, that the two retest results were 97.5 and 98.8. This indicates a likelihood of some unnoticed mistake such as a weighing or dilution of the original sample or standard.

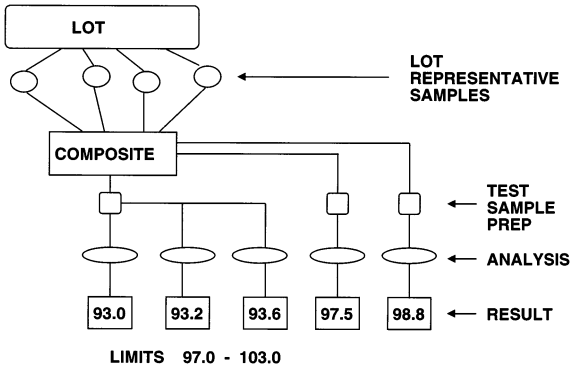

Fig. 4. Retest with reanalysis schematic.

Again, the real problem is not that the error occurred but, rather, what to do about it. We now have five data items to utilize, but these should not all be weighted equally because of the way they were generated. A simple and straightforward way to achieve a reasonable weighting is to calculate a single result from the three data points from the first sample preparation and then to combine this result with the two values obtained from two independent retests of additional preparations. This stepwise use of the median is the 'structure' referred to in the name of the methodology.

Consider how an outlier procedure would operate and where it would lead us in this situation. Since all three results from the initial sample preparation were consistent with each other, no outlier test would be performed at that point. The average for those three results is 93.3, compared to the two other results of 97.5 and 98.8. Again assuming no historical estimate of the S.D. is available, the extreme studentized deviate test using the sample S.D. estimate does not identify 93.3 as an outlier at the 5% level of significance. The arithmetic average of all three results is 96.5 and, once again, we are left with the prospect of generating additional results in order to overcome what, in all probability, was a simple mistake in weighing or dilution.

On the other hand, the proposed median procedure would give 93.2 as the combined result (median) from the initial sample preparation and would then yield 97.5 as the combined result (median) from the three values 93.2, 97.5 and 98.8. With the median procedure, no further testing is required to assign a result to the sample.

3. General performance of the median and outlier procedures

The examples given in the preceding section illustrate only a few of the many possible situations and outcomes when using the different procedures. In particular, the examples illustrated cases where the use of the median appeared to give a result more representative of the true sample potency than the result obtained through use of the outlier tests. Other examples could be constructed where this would not occur or not be as evident. In order to compare the performance of the outlier approach versus the median approach over a wider range of possible scenarios, a simulation study was carried out as described in detail in Appendix B.

A simulation study, also called a Monte Carlo experiment, is a way to characterize the behavior of a method or technique by performing multiple repetitions using a computer to generate the results. Simulation is often used when it is difficult to obtain closed-form mathematical expressions.

For the present case, we simulate a two-stage procedure that calls for additional retest or reanalysis results whenever an initial result falls below some predetermined limit. This setup is a simplified version of the actual situation, but it retains most of the essential elements and permits a straightforward comparison. Factors that may affect performance include: the number of additional results to be generated, the magnitude of an analytical mistake relative to the true sample result, the likelihood of such an analytical mistake and the location of the true sample result relative to the limit. This simulation study varied these parameters in a controlled pattern according to a set protocol.

Three potential procedures for handling results were studied:

1. Report the median of all the results.
2. Report the arithmetic average after removing outliers via the ESD outlier test using only the current sample results to estimate the S.D.
3. Report the arithmetic average after removing outliers via the ESD outlier test using a historical estimate of the S.D.

The first issue examined was a comparison of the procedures with respect to how often they produced passing results after an initial result fell below the limit. In order to pass

1. all of the additional results generated must be greater than the limit and
2. the final computed result must be greater than the limit.

In this situation, a good procedure is one that gives a high proportion of passing results when the true sample value is above the limit and a low proportion of passing results when the true sample value is below the limit. The second criterion of interest was how close the reported values were to the true value for each sample. The best procedure in this case would have the smallest root mean square error, which is the square root of the sum of the squared bias plus variance. The third and final issue was the likelihood of getting a passing result with the two-stage procedures, as compared with a single-stage procedure that used only the single initial result. Again, a desirable procedure would be one with more passing results for samples above the limit, but not for samples below the limit.

The simulation found that no one procedure is best over all factors or in all respects, but the following trends were observed:

The probability of producing a passing result after an initial result fell below the limit was similar for all the procedures.

The root mean square error was similar for all the procedures, but when the true mean was 2-σ above the limit, the median was a better estimator than the ESD outlier test using only the sample S.D.

The root mean square error from the ESD outlier test, using an historical S.D., was the lowest of all of the procedures in all situations, although the median procedure was very comparable.

As estimators, the median procedure and the ESD outlier test using an historical S.D. performed equally well regardless of the size of the potential analytical mistake, while the ESD outlier test using the sample S.D. was affected by this factor.

The increased probability of producing a passing result afforded by a second stage was similar for all three procedures studied. Thus, given that reanalysis or retest will be done, use of the median is neither more nor less likely to produce passing results than the other forms of averaging.

Little difference in the probability of passing was observed as a function of the number of retests performed. Thus, it is recommended that two retests are usually sufficient to assess whether the test sample passes the limit.

The interested reader may wish to examine Appendix B to obtain more detail regarding the basis for these conclusions.

4. Conclusions and recommendations

In this paper, we have outlined a procedure for investigating for evidence of probable analytical error based on the structured use of the median and we have compared the performance of the median to outlier testing in a simplified simulation study.

The paper has presented the case for using the median as an alternative to traditional outlier testing. The median provides a consistent approach that is both simple and objective and it provides a methodology that can be followed easily by all individuals in an organization. Since the use of the median does not involve a statistical test for outliers and does not require that aberrant data be disregarded, it overcomes some of the deficiencies inherent in outlier testing.

The statistical properties of the median are also favorable. The simulation study demonstrated that the performance of the median was not strongly affected by changes in several of the important parameters studied and that the performance of the median was comparable or even superior to the outlier procedures.

The performance of the outlier test procedures is highly dependent upon having a prior estimate of the S.D. to use in the evaluation. When an historical S.D. estimate is available, the outlier test performs well. However, when only the sample S.D. is available, the outlier test does not perform very well. This could be an issue early in drug development when there is limited historical information on the variability of the analytical method.

Based on the analysis in this report, we recommend that the structured use of the median be given strong consideration for adoption in appropriate analytical laboratory procedures. For situations where there is a good deal of historical information available, the ESD outlier test procedure using an historical estimate of the S.D. is also a viable alternative.

Acknowledgements

The authors wish to thank Dr Wendell Smith for his suggestions concerning the initial problem, his helpful comments on the simulations and his review of the manuscript. The authors are also indebted to Dr Doris Weisman and Dr Robert Obenchain for their helpful review of earlier drafts of the paper. The authors would also like to acknowledge the helpful comments of the statistical reviewer.

Appendix A. Extreme studentized deviate outlier procedures

A.1. Extreme studentized deviate test (S.D. calculated from the sample)

The initial result observed below the limit is $X1$. Generate additional results $X2$, $X3, \ldots, Xn$. Compute the sample average $\bar{X}$ and S.D. s. Then compute T using the following formulas:

$$T = [\bar{X} - X1]/s$$

$X1$ is the suspected outlier.

The suspect observation is deemed an outlier if T exceeds a tabled critical value. For sample sizes 3, 4 or 5, the one-sided $\alpha = 0.05$ critical values are:

Sample size	3	4	5
Critical value	1.15	1.46	1.67

A.1.1. Example

Consider the sample values 94.0, 99.3 and 100.2. The ESD test for 94.0 being an outlier gives

$$T = (97.833 - 94.0)/3.350 = 1.144$$

Since T is less than the tabulated value 1.15, the low value 94.0 is not deemed an outlier by the ESD procedure.

A.2. Extreme studentized deviate test (historical S.D.)

The initial result observed below the limit is $X1$. Generate additional results $X2$, $X3, \ldots, Xn$. Compute the sample average $\bar{X}$. An estimate of the historical S.D. *sh*, is available. Compute T using the following formulas:

$$T = [\bar{X} - X1]/sh$$

$X1$ is the suspected outlier

The suspect observation is deemed an outlier if T exceeds a tabled critical value.

A.2.1. Example

Consider the sample values 94.0, 99.3 and 100.2. Assume the historical S.D. with 10 dof to be 1.12. The ESD test for 94.0 being an outlier gives

$$T = (97.833 - 94.0)/1.12 = 3.42$$

For sample sizes 3, 4 or 5, the one-sided $\alpha = 0.05$ critical values for 10 dof are:

Sample size	3	4	5
Critical value	2.01	2.27	2.46

Since T is greater than the tabulated value 2.01, the low value 94.0 is deemed an outlier by the ESD procedure using the historical S.D. estimate.

Appendix B. Simulation comparing the performance of the median and outlier procedures

A.1. Introduction

In order to understand and compare the performance of candidate procedures, a simulation study was conducted. For the interested reader, a copy of the SAS[1] software code used to generate the results is included in Appendix C.

In setting up the simulation, we note that none of the procedures comes into play unless an initial result seems 'unusual', (such as falling outside 3σ control limits or falling outside of an action limit). To simplify, let us assume that there is a one-sided lower limit of zero. A result ≥ 0 passes, while a result < 0 does not pass. This simulation examines simple two-stage procedures that operate in the following manner:

1. A single initial result is obtained.
2. If the initial result is $\geqq 0$, then the test requirement is met.
3. If the initial result is < 0, then two or more additional results are obtained.
4. Perform appropriate procedure to evaluate additional results
   - 4.1. *ESD outlier procedure using the sample estimate of the S.D.*: extreme studentized deviate test is performed with the S.D. calculated from the sample results only. The test requirement is met if all of the additional results are $\geqq 0$ and the average of the results, excluding any detected outliers, $\geqq 0$.
   - 4.2. *ESD outlier procedure with an historical S.D. estimate with 10 dof*: Extreme studentized deviate test is performed with the S.D. randomly generated from a true distribution with S.D. $= 1$ and 10 dof. The choice of 10 dof was made since the critical values for the ESD are tabulated only down to 10 dof in the literature. By using the fewest degrees of freedom, the extreme studentized deviate test using the historical S.D. would have the least discriminatory power. It was expected that even with as few as 10 dof this approach would perform very well, and this expec-

[1] SAS Institute, Inc., SAS Circle P.O. Box 8686, Cary, N.C. 27512-8000, USA.

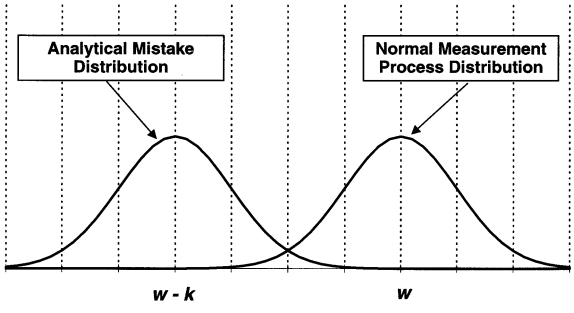

Fig. 5. Plot showing simulation distributions. w, true test sample mean; k, size of analytical mistake.

tation was confirmed in the simulations. The performance of this test improves as the degrees of freedom for estimating the historical S.D. increase. As above, the test requirement is met if all of the additional results are greater than or equal to zero and the average of the results, excluding any detected outliers, is $\geqq 0$.

4.3. *Median procedure*: The test is met if the all of the additional results are $\geqq 0$ and the median of all the results is $\geqq 0$.

We are interested in comparing how the above procedures perform in the conditional subset (second stage) where the initial result was below the limit.

A.2. Parameter space

It is known that on each assay there is a chance of an analytical mistake. For a validated and in-control assay procedure, the probability (δ) of such a mistake is small. In the simulation, $\delta = 0.01$ was used.

This paper recommends that only two additional results be obtained for reanalysis or retest, but it is of interest how the procedures compare when more, say, three or four reanalyses or retests are done, and the simulation included these cases, as well. We use n to denote the total number of results, making $n = 1$ plus the number of reanalyses or retests.

The results were generated by drawing them from one of two simulated unit ($\sigma = 1$) normal distributions as follows:

1. With probability $100(1 - \delta)$, the result came from a true sample distribution with mean w
2. With probability δ, it came from an analytical mistake distribution with mean w-k

Fig. 5 provides a graphical illustration.

The distribution with mean w represents the population of results that come from the normal measurement process for a sample, while the distribution with mean w-k denotes the population of aberrant or outlying results that arise from an analytical mistake of magnitude k. The ideal procedure would always fail a sample for which $w < 0$ and to always pass any sample that has $w > 0$.

The simulation study was conducted over the following parameter space: $\delta = 0.01$; $n = 3$, 4 and 5; $k = 2$, 4 and 6; $w = -3.0$–4.0×0.5.

A.3. Methodology

For each parameter combination above, a minimum of 5686 sets of observations were obtained as described earlier. At each iteration, an initial result was randomly generated from the true sample distribution with probability 1-δ or the mistake distribution with probability δ until an observation less than the limit was obtained. Once a result < 0 was obtained, then an additional 2, 3 or 4 results were generated each with a probability 1-δ of coming from the true sample distribution and probability δ of coming from the analytical mistake distribution.

For each set of observations, the set of results was scored as a 'pass' or a 'fail'. For each of the outlier procedures, the average of the results excluded any outliers detected by that procedure. If that average was $\geqq 0$ and all of the additional results generated were $\geqq 0$, the corresponding outlier procedure was scored as a 'pass' and scored a 'fail' otherwise. For the median procedure, if the median of all the results was $\geqq 0$ and all of the additional results generated were $\geqq 0$, the median procedure was scored as a 'pass' and scored a 'fail' otherwise.

A.4. Conditional operating characteristics

The first question examined was the following, 'Are the procedures similar in the proportion of samples passed?' For example, does one procedure pass more samples with true means below the limit ($w < 0$) or pass fewer samples with true means above the limit ($w > 0$)? To answer these questions, the procedures were compared on the basis of the conditional probability of passing the test, given that the initial result was < 0 (conditional operating characteristics).

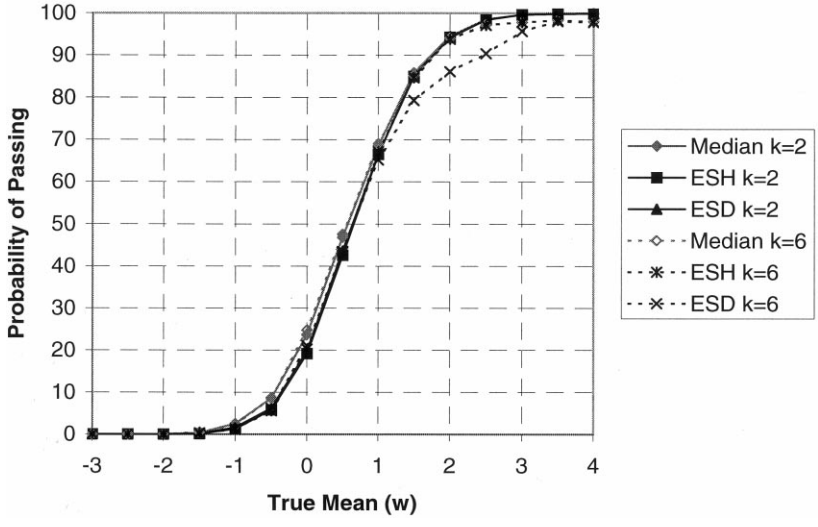

Fig. 6. Conditional probability of passing for two retests and probability of an analytical mistake = 1%. Median, median investigation procedure; ESD, extreme studentized deviate test using sample S.D.; ESH, extreme studentized deviate test using historical S.D. with 10 dof; w, true test sample mean; k, size of analytical mistake.

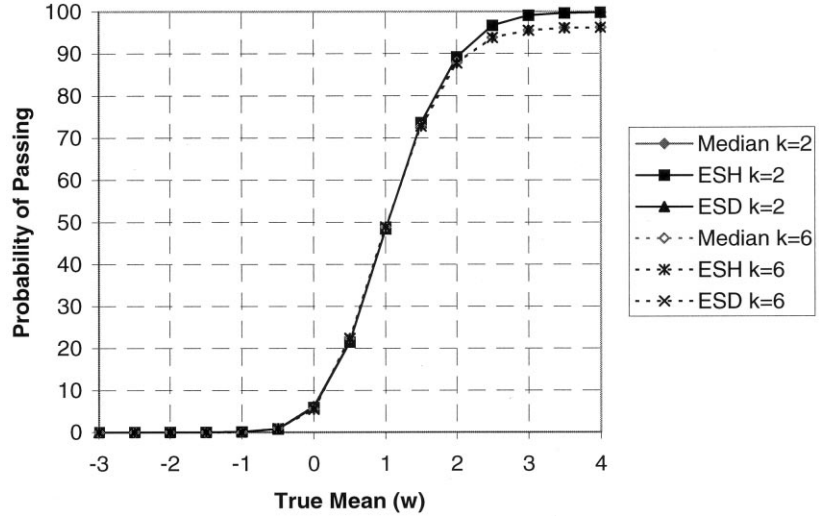

Fig. 7. Conditional probability of passing for four retests and probability of an analytical mistake = 1%. Median, median investigation procedure; ESD, extreme studentized deviate test using sample S.D.; ESH, extreme studentized deviate test using historical S.D. with 10 dof; w, true test sample mean; k, size of analytical mistake.

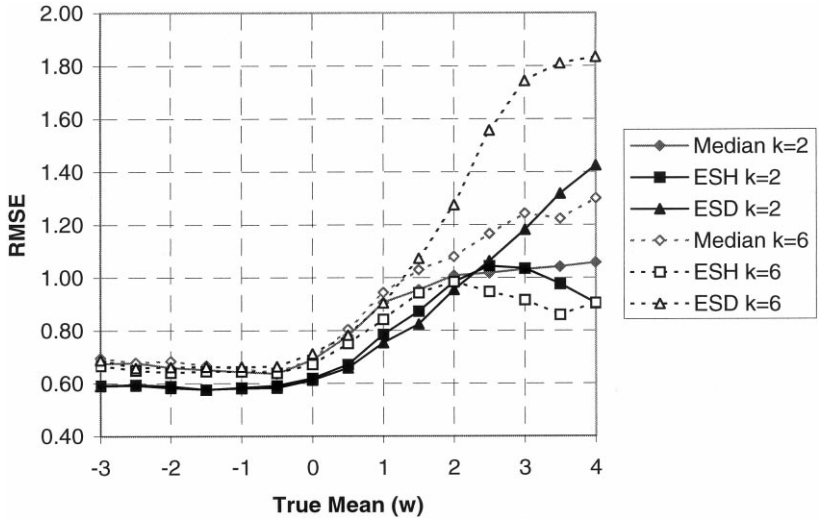

Fig. 8. Root mean square for two retests and probability of an analytical mistake = 1%. Median, median investigation procedure; ESD, extreme studentized deviate test using sample S.D.; ESH, extreme studentized deviate test using historical S.D. with 10 dof; w, true test sample mean; k, size of analytical mistake.

Plots of the results of the simulation are given in Figs. 6 and 7 for two and four retests and where $k = 2$ and $k = 6$. Similar results were observed for other cases.

The following observations were made with respect to the conditional operating characteristics:

The probability of passing for all of the procedures was very similar overall.

The median procedure and the ESD outlier procedure using an historical S.D. are more likely to pass the test than the ESD outlier procedure using the sample S.D. when the true sample mean is 2-σ or more above the limit and when a large analytical mistake is possible (see Fig. 6 for $k = 6$ and $w > 2$).

Based on these observations, we concluded that:

1. the operating characteristics of the median procedure and the ESD outlier procedure using an historical S.D. were less sensitive to the size and frequency of analytical mistakes than the ESD outlier procedure using the sample S.D.
2. the median produced passing results neither uniformly more frequently nor uniformly less frequently than the other procedures.
3. the operating characteristic of the median procedure and the ESD outlier procedure using a historical S.D. estimate were very similar even for large analytical mistakes.

A.5. Estimation properties

The second question of interest is how closely do the reported results from each procedure estimate the true sample mean w. Since the goal is to produce accurate and precise results, the performance of the procedures can be assessed using the common statistical criterion of root mean square error (RMSE), which combines both the accuracy and variability of an estimator into a one-number summary:

$$\text{RMSE} = (\text{Bias}^2 + \text{Variance})^{1/2}$$

The RMSE measures both the mean and variability of a computed result, and a smaller RMSE implies better performance of an estimator. In the simulation, Bias for any given procedure was measured by the difference between the mean of 5686 final results and the true mean w, and Variance was computed as the sample variance of the 5000 final results. Plots of the results of the simulation are included as Figs. 8 and 9 for two and

four retests, where $k = 2$ and $k = 6$. Similar results were observed for other cases.

With respect to the RMSE comparison, the following observations were made:

The RMSE from the ESD outlier test using an historical estimate of the S.D. was the lowest of all of the procedures in all situations although the median procedure performed very similarly. The median was a better estimator than the ESD outlier test using the sample S.D. when the true mean was 2-σ above the limit ($w \geq 2$). As estimators, the median procedure and the ESD outlier test using an historical S.D. estimate performed similarly regardless of the size of the potential analytical mistake, while the ESD outlier test using the sample S.D. was affected by this factor.

From these observations, we concluded that the median procedure and the ESD outlier procedure using an historical S.D. perform very similarly. The median is the preferred estimator due to the simplicity of implementation and the fact that it was not sensitive to the size and frequency of analytical mistakes.

A.6. Comparison of the procedures with a single stage

One additional aspect of interest with the two-stage procedures is the probability of a passing result compared to that from a single-stage procedure using solely the initial result. The probabilities described here are different from those generated by the simulation, since we are now dealing with all tests, rather than the subset of only those tests which had nonconforming initial results. However, for the two-stage procedures, the quantities of interest can be calculated using the following formula:

$$\begin{aligned}
&\Pr(\text{Pass}) \\
&= \Pr(\text{Pass at Stage 1}) + \Pr(\text{Pass at Stage 2}) \\
&= \Pr(\text{Init. Rslt.} \geq 0) + \Pr(\text{Init. Rslt.} < 0) \\
&\qquad \times \Pr(\text{Pass} \mid \text{Init. Rslt.} < 0) \\
&= 1 - P + [P \times \Pr(\text{Pass} \mid \text{Init. Rslt.} < 0)],
\end{aligned}$$

where $P = [\delta\Phi(k - w) + (1 - \delta)\Phi(-w)]$ and was calculated using standard normal tables.

The conditional probability in the above expression is obtained directly from the simulation results.

Using the information obtained from the simulations, we evaluated the effect on the overall probability of passing of the outlier and median procedures for different numbers of retests as compared to a single-stage procedure. For the

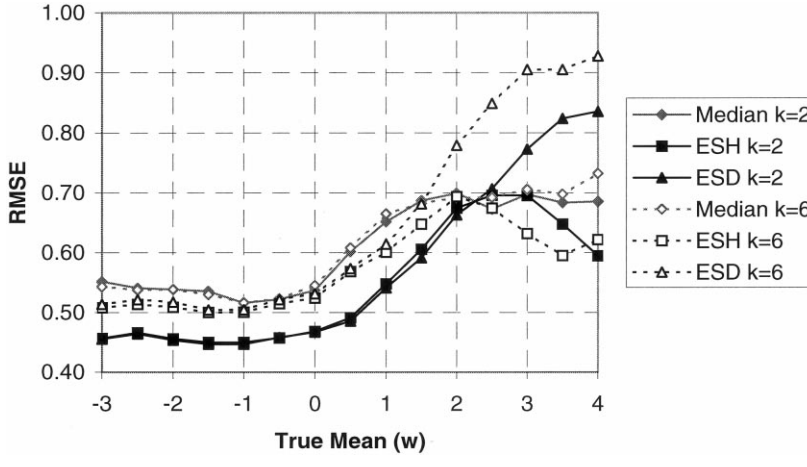

Fig. 9. Root mean square for four retests and probability of an analytical mistake = 1%. Median, median investigation procedure; ESD, extreme studentized deviate test using sample S.D.; ESH, extreme studentized deviate test using historical S.D. with 10 dof; w, true test sample mean; k, size of analytical mistake.

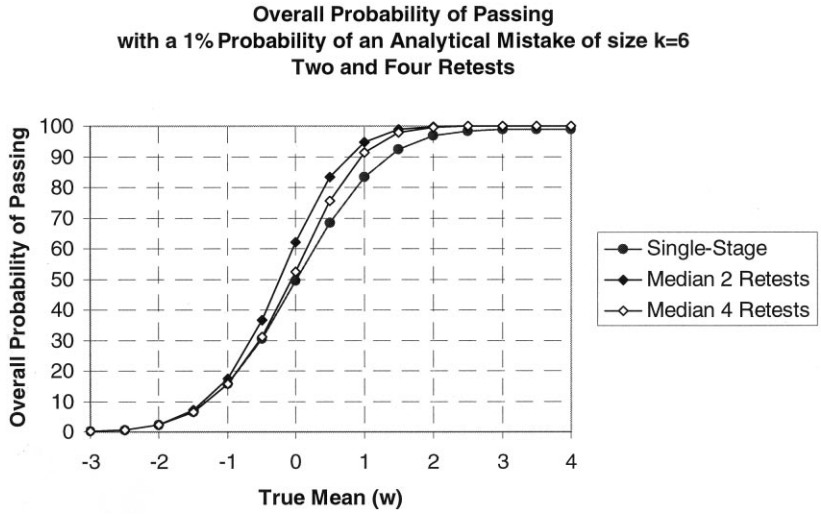

Fig. 10. Overall probability of passing with a 1% probability of an analytical mistake of size $k = 6$ for two and four retests. Single-stage procedure involving decisions based solely on initial observed result. Median 2 retests, a procedure allowing second stage and decision using median after two retests. Median 4 retests, a procedure allowing second stage and decision using median after four retests.

cases of two or four retests where $\delta = 0.01$ and $k = 6$, the results for the median procedure are shown in Fig. 10. Results for the outlier procedures and for other values of δ and k are very similar, and are not shown. We draw the following conclusions:

The outlier procedures and median procedure performed comparably with respect to the overall probability of passing when a two-stage approach was used.

There is little change in the overall probability of passing for different numbers of retests.

The addition of a second stage generated the greatest increase in probability of passing in the region where it is desirable to do so ($w \geq 0$). In the region where such an increase is undesirable, it causes little or only moderate increase.

From these observations, we conclude that if a two-stage procedure is to be used, then the median is no more likely to produce passing results than the other forms of averaging studied. In addition, little change in the probability of producing passing results is observed as the number of retests is increased.

Appendix C. SAS code utilized to generate the simulation results

```
data simvals ;
keep n del k w totalit      medval eshval esdval
                            medpas eshpas esdpas ;
niter=5000 ;
array x (i) x1-x5 ;
/* --- Critical Values for ESD Test and ESD Historical Test ----- */
```

```
array ehcrit (m) h1-h3 ; array edcrit
  (m) e1-e3 ;
h1=2.010 ; h2=2.270 ; h3=2.460 ; e1=1.15 ; e2=
  1.46 ; e3=1.67 ;
/* ------------------------------------------------------- */

do n=3,4,5 ; m=n-2 ;
  do del=.01 ;
    do k=2,4,6 ;
      do w=-3 to 4 by.5 ;
        timesneg=0; to-
talit=0 ;
       do until
(timesneg=niter) ;
/* Special Treatment for w>3 to avoid Excessive Iterations */
       if w>3 then p=0 ;
else p=uniform(0) ;
/* ------------------------------------------------------- */
       x1=w-(p<del)*k
                 +nor-
mal(0); totalit+1 ;
/* --- Start of Loop if Initial Result is Negative -------- */
       if x1<0 then do ;
         timesneg+1 ;
sum=x1 ; ssum=x1*x1 ;
min=x1; max=x1 ;
/* -------- Getting and Sorting the Results --------------- */
         do i=2 to n ;
           p=uniform(0) ;
x=w-(p<del)*k
         +normal(0) ;
           sum+x ; ssum+
x*x ;
           if x<min then
min=x ; if x>max then
max=x ;
         end ;
         ml=max ; mh=min ;


         do i=1 to n ;
           if min<x<max
  then do ;
             if x<ml then
  ml=x ; if x>mh then mh=
  x ;
               end ;
```

```
        medval
        =(ml+mh)/2 ;

        if n=5 then do ;
          do i=1 to n ; if
ml<x<mh then medval=x ;
end ;
        end ;
/* ------------------------------------------------------ */
        Mean=sum/n; s=
sqrt((ssum-mean*sum)
/(n-1)) ; r=max-min ;
/* ------ Determining Outliers Using the ESD Test --------- */
        crit
        =(mean-x1)/s ;
        if crit>edcrit
then do ;
           esdval
           =(n*mean-x1)
             /(n-1) ;
        end ;
        else esdval=
mean ;
/* ------ Determining Outliers Using
  the ESD Historical Test ---------
   */
      crit2
      =(mean-x1)
        /(((2*rangam(0,10/2))/10)
        **(1/2)) ;
      if crit2>ehcrit
then do ;
         eshval
         =(n*mean-x1)
           /(n-1) ;
      end;
      else eshval=mean ;

      if n=3 then do ;
      medpas
      =(medval>=0) and
(x2>0 and x3>0) ;
```

```
       eshpas
       =(eshval>=0) and
(x2>0 and x3>0);
       esdpas
       =(esdval>=0) and
(x2>0 and x3>0);
       end;

       if n=4 then do;
       medpas=(medval>=0) and (x2>
  0 and x3>0 and x4>0);
       eshpas=(eshval>=0) and (x2>
  0 and x3>0 and x4>0);
       esdpas=(esdval>=0) and (x2>
  0 and x3>0 and x4>0);
       end;
       if n=5 then do;
       medpas=(medval>=0) and (x2>
  0 and x3>0 and x4>0 and x5>0);
       eshpas=(eshval>=0) and (x2>
  0 and x3>0 and x4>0 and x5>0);
       esdpas=(esdval>=0) and (x2>
  0 and x3>0 and x4>0 and x5>0);
       end;
       output; totalit=
0;
     end;
/* -------- End of Loop for Initial
  Negative Result ------- */
       end; /             * End of
                            itera-
                            tion
                            loop */
     end; /               * End
                            of'w'
                            loop */
   end; /                 * End
                            of'k'
                            loop */
  end; /                  * End
                            of'del'
                            loop */
end; /                    * End
                            of'n'
                            loop */

proc means data=simvals noprint sum mean std; by n del k w;
```

```
var totalit                medval es-
                             hval es-
                             dval
                           medpas
                             eshpas
                             esdpas ;
output out=simsum                    Sum=totliter
                                     mean=dum1     medmn    eshmn    esdmn
                                                   medpass  eshpass  esdpass
                                     std=dum2      medsd    eshsd    esdsd ;
data simsum; set simsum ;
  keep n del k w totliter            Medmn         eshmn    esdmn
                                     Medsd         eshsd    esdsd
                                     Medrmse       eshrmse  esdrmse
                                     Medpass       eshpass  esd-
                                                              pass ;

medpass=round(100*medpass,.1) ;
medrmse=sqrt((medmn-w)**2+medsd**2) ;
eshpass=round(100*eshpass,.1) ;
eshrmse=sqrt((eshmn-w)**2+eshsd**2) ;
esdpass=round(100*esdpass,.1) ;
esdrmse=sqrt((esdmn-w)**2+esdsd**2) ;
```

References

[1] United States Pharmacopeia, Design and Analysis of Biological Assays, 22nd rev. ed., general chapter 111, Mack, Easton, PA, 1989, p. 1502.

[2] Analytical Methods Committee, Robust statistics...how not to reject outliers, Analyst 114 (1989) 1693–1702.

[3] R.J. Beckman, R.D. Cook, Outliers.....s, Technometrics 25 (1983) 119–149.

[4] W.J. Dixon, Ratios involving extreme values, Ann. Math. Stat. 22 (1951) 68–73.

[5] W.J. Dixon, Processing data for outliers, Biometrics 9 (1953) 74–89.

[6] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics 11 (1969) 1–21.

[7] D.C. Hoaglin, F. Mosteller, J.W. Tukey, Understanding Robust and Exploratory Data Analysis, Wiley, New York, NY, 1983.