An Assessment of the 4-6-20 Rule for Acceptance of Analytical Runs in Bioavailability, Bioequivalence, and Pharmacokinetic Studies

Robert O. Kringle[1]

Received May 28, 1993; accepted September 30, 1993

A recent conference report described a decision rule, hereafter referred to as the 4-6-20 rule, for acceptance/rejection of analytical runs in bioavailability, bioequivalence, and pharmacokinetic studies. This procedure requires that quality control specimens at three concentrations (low, medium, and high) be assayed in duplicate in each run. For run acceptance, at least four of the six assay values must be within ±20% of their respective nominal concentrations, and at least one of the two values at each concentration must be within these limits. An inherent flaw in this decision rule is that the risk of rejecting runs, when the assay performance has in fact not deteriorated, varies for each assay and is neither known nor controlled. In this paper simulation methods are used to evaluate the operating characteristics of the 4-6-20 rule in comparison to those of classical statistical quality control procedures.

KEY WORDS: quality control; Shewhart control; multivariate control; operating characteristics; power.

INTRODUCTION

After an analytical method has been validated to show that its performance characteristics are acceptable prior to use in bioanalytical studies, the performance of the method in routine applications is monitored with quality control (QC) methods. A recent conference report (1) provided guiding principles for both validation studies and quality control methods. This paper is an evaluation of the *QC* procedures recommended in the conference report. The conference recommendations for *validation* studies will be evaluated in a forthcoming paper.

The conference report describes a decision rule, hereafter referred to as the 4-6-20 rule, for acceptance/rejection of analytical runs in bioavailability, bioequivalence, and pharmacokinetic studies (1). This procedure requires that QC specimens at three concentrations (low, medium, and high) be assayed in duplicate in each run. For run acceptance, at least four of the six assay values must be within ±20% of their respectively nominal concentrations, and at least one of the two values at each concentration must be within these limits. An inherent flaw in this decision rule is that the risk of rejecting runs, when the assay performance has in fact not deteriorated, varies for each assay and is neither known nor controlled. Although the conference recommendations for validation studies include the use of various statistical methods, the recommendations for QC are completely void of statistical methods. In this paper classical statistical QC procedures are presented as alternatives to the 4-6-20 procedure, and simulation methods are used to evaluate and compare the operating characteristics of the 4-6-20 rule and the statistical procedures.

METHODS

Shewhart Control Procedure

One of the oldest, simplest, and most commonly used QC procedures for monitoring the stability of a process is the Shewhart (2) procedure for the mean and standard deviation (or alternatively, the range) of n measurements on a QC specimen in each run. The control limits for the run mean, allowing for both between-run and within-run variance components, are

$$\mu \pm z_{1-\alpha/2}[\sigma_B^2 + \sigma_W^2/n]^{1/2} \quad (1)$$

where the population mean, μ, and the variance components, σ_B^2 and σ_W^2, are assumed known but, in practice, are estimated with prior data in what is called the baseline period. On a run-by-run basis, the Shewhart procedure for the run mean is a statistical hypothesis test that the mean of the method has not changed. The false rejection rate, α, is controlled by the value of the standard normal deviate $z_{1-\alpha/2}$. The commonly used standard Gaussian critical value, $z_{1-\alpha/2} = 3$, yields a false rejection probability $\alpha = 0.0027$.

The upper control limit for the within-run standard deviation is

$$[\sigma_W^2 \chi^2_{\nu,1-\alpha}/\nu]^{1/2} \quad (2)$$

On a run-by-run basis, the Shewhart procedure for the within-run standard deviation is a statistical hypothesis test that the within-run variation of the method has not increased. The false rejection rate, α, is controlled by the value of the chi-square deviate $\chi^2_{\nu,1-\alpha}$ with $\nu = n - 1$ degrees of freedom (df). For $n = 2$, the chi-square critical value $\chi^2_{\nu,1-\alpha} = 9.0$ yields a false rejection probability $\alpha = 0.0027$.

The baseline period for the Shewhart procedure should consist of 30 or more runs and should be frequently updated with the accrual of routine runs. If there are far fewer than 30 runs, then the Gaussian and chi-square critical values in (1) and (2) should be replaced by the appropriate critical values of the t and F distributions to control the false rejection probability.

If the Shewhart procedure is applied to each of the three control specimens, then there are six statistics (three means and three standard deviations) to monitor for each run. If a run is rejected when one or more of the six statistics exceeds its control limits, then the overall false rejection probability is

$$\alpha_0 = 1 - \prod_{i=1}^{6} (1 - \alpha_i) \quad (3)$$

[1] Sterling Winthrop Pharmaceuticals Research Division, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426-0900.

0724-8741/94/0400-0556$07.00/0

assuming that the six statistics are independent. An approximation that holds for small α_i is $\alpha_0 = \Sigma\alpha_i$. The probability α_0 gives the frequency of runs that will unnecessarily be rejected or repeated, and therefore it should be an acceptably small value. If each α_i is 0.0027, then α_0 is 0.0161. The three within-run standard deviations are always independent from each other and from the means, but the three means are not mutually independent for assays with significant between-run covariances for the three QC specimens. In such cases Eq. (3) provides bounds for α_0, using all six terms for the upper bound and only four terms (for the three standard deviations and one mean) for the lower bound.

Multivariate Control Procedure

For assays where independence is not satisfied, a multivariate analogue (3) of the Shewhart procedure provides exact control of the overall false rejection rate and, in addition, greater statistical power to detect real changes in assay performance. There are two QC statistics, one for monitoring the run mean vector and one for monitoring the within-run variability. The statistic for monitoring the vector $\bar{\mathbf{x}}$ of $p = 3$ control specimen means for a specific run is

$$T_M^2 = (\bar{\mathbf{x}} - \boldsymbol{\mu})'(\Sigma_B + \Sigma_W/n)^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (4)$$

where the population mean vector, $\boldsymbol{\mu}$, and the between-run and within-run p-dimensional variance–covariance components, Σ_B and Σ_W, are estimated from baseline data. The upper control limit for T_M^2 is $\chi^2_{\nu,1-\alpha}$ with $\nu = p$ df, where p is the number of control specimens. The false rejection rate α is controlled by the value of the chi-square deviate $\chi^2_{\nu,1-\alpha}$.

The statistic for monitoring the within-run variability for a specific run is

$$T_D^2 = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})\,\Sigma_W^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) \quad (5)$$

The upper control limit for T_D^2 is $\chi^2_{\nu,1-\alpha}$ with $\nu = p(n-1)$ df. The false rejection rate α is controlled by the value of the chi-square deviate $\chi^2_{\nu,1-\alpha}$.

If a run is rejected when either T_M^2 or T_D^2 exceeds its control limit, then the overall false rejection probability is $\alpha_0 = 1 - (1 - \alpha_M)(1 - \alpha_D)$, which is well approximated by $\alpha_0 = \alpha_M + \alpha_D$ for small values. The baseline period for the multivariate procedure should consist of at least 30 runs; otherwise the chi-square critical values should be replaced with appropriate critical values of the F distribution.

Simulation Model for Inherent Assay Bias and Variability

The *operating characteristics* (false rejection rates and power probabilities) of the 4-6-20, Shewhart, and multivariate procedures were evaluated by simulation. To evaluate and compare the false rejection rates for the three procedures, the following model was used to generate random observations for the QC specimens:

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \mathbf{r}_i + \mathbf{e}_{ij} \quad (6)$$

where $\mathbf{x}_{ij}$ is the assay result vector (3×1, representing the three control samples) for the ith run and the jth replicate ($j = 1,2$); $\boldsymbol{\mu}$ is the 3×1 vector of true means of assay results; $\mathbf{r}_i$ is the 3×1 vector of random run effects for the ith run, assumed to be $N(\mathbf{0},\Sigma_B)$; and $\mathbf{e}_{ij}$ is the 3×1 vector of random errors for the ith run and the jth replicate, assumed to be $N(\mathbf{0},\Sigma_W)$.

Further assumptions to simplify the simulation study were as follows

(a) $\boldsymbol{\mu} = \beta\mathbf{c}$ where c is the vector of nominal concentrations and $100(\beta - 1)$ represents the percentage multiplicative *inherent assay bias;*
(b) $\mathbf{c} = (100, 100, 100)$ without loss of generality;
(c) Σ_W and Σ_B are diagonal matrices; and
(d) the total standard deviation (square roots of diagonal elements of $\Sigma_T = \Sigma_B + \Sigma_W$) is a constant percentage of the mean (elements of $\boldsymbol{\mu}$), i.e., the *inherent assay variability* is a constant *coefficient of variation* ($\%CV = 100\sigma/\mu$).

Five thousand runs were simulated according to the model in Eq. (6), using the RANNOR function in SAS (4), for each combination of total CV (1 to 15% in increments of 1%) and inherent bias (0, +5, +10, and +15%). These maximum values (15% for total CV and +15% for bias) represent the extremes for a valid assay, according to the recommendations (1). The proportion of the total variance due to the between-run and within-run components was set at 20 and 80%, respectively. For the Shewhart procedure, each of the six α values was set to 0.0027, so the overall α_0 was 0.0161. For the multivariate procedure, each of the two α values was set to 0.0080 to make the overall α_0 the same as that for the Shewhart procedure.

Simulation of Changes in Assay Bias or Variability

To investigate the power of the three QC procedures to detect changes in *assay bias,* 5000 runs were simulated with the model

$$\mathbf{x}_{ij} = \gamma\boldsymbol{\mu} + \mathbf{r}_i + \mathbf{e}_{ij} \quad (7)$$

where γ represents a constant multiplicative increase in the assay results, for each of three total CVs (3, 8, and 15%) and a series of values of $\gamma \geq 1.0$. It was assumed that there was no inherent bias (i.e., $\beta = 1$). Thus, changes in average assay results were represented by changes in assay bias from 0 to $100(\gamma - 1)\%$. The proportion of the total variance due to the between-run and within-run components was 20 and 80%, respectively. The α_0 value for both the Shewhart and the multivariate procedures was 0.0161.

To investigate the power of the three QC procedures to detect changes in *assay variability,* 5000 runs were simulated with the model

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \lambda(\mathbf{r}_i + \mathbf{e}_{ij}) \quad (8)$$

where λ represents a constant multiplicative increase in the assay standard deviation, for each of three inherent total CVs (3, 8, and 15%) and a series of values of $\lambda \geq 1.0$. It was assumed that there was no inherent bias (i.e., $\beta = 1$). The proportion of the total variance due to the between-run and within-run components was 20 and 80%, respectively. The α_0

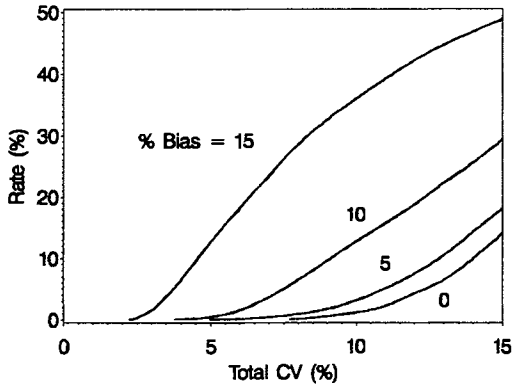

Fig. 1. False rejection rates (rates at which runs are rejected when neither bias nor CV has changed) for the 4-6-20 rule, as a function of inherent bias and inherent CV.

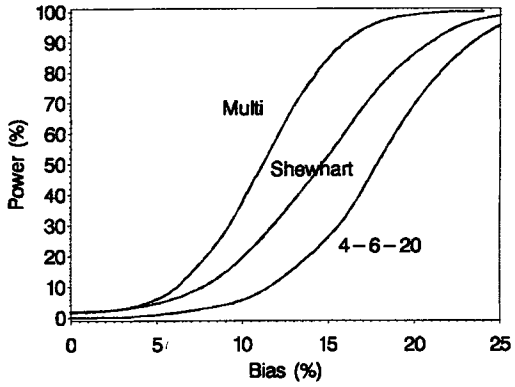

Fig. 3. Power of three QC rules to detect actual changes in assay bias from an inherent 0% bias. Inherent total CV = 8%.

value for both the Shewhart and the multivariate procedures was 0.0161.

RESULTS

False Rejection Rates

False rejection rates for the 4-6-20 procedure, as a function of the inherent total CV and the inherent bias, are shown in Fig. 1. For small values of total CV and bias, the false rejection rate is essentially zero. This indicates that essentially no runs will be unnecessarily repeated or rejected. However, there is a cost for this "undercontrol" in terms of reduced power to detect real changes in assay performance, as will be shown in the next two sections. As the CV or bias increases, the false rejection rate increases. For bias <5% and total CV <10%, the false rejection rate remains low, 2% or less. However, for some assays with analytical performance well within the validation acceptance criteria, say 10% CV and 10% bias, the false rejection rate is high, approximately 12%. Thus, 12% of routine runs will be unnecessarily rejected or repeated. For assays with performance at the validation acceptance limits, 15% CV and 15% bias, the false rejection rate is approximately 50%. In contrast to these uncontrolled and sometimes high false rejection rates with the 4-6-20 procedure, the false rejection rate for the Shewhart and multivariate procedures is controlled at α_0 = 1.6% for all values of inherent total CV and inherent bias.

Power to Detect Changes in Assay Bias

Power probabilities, or rejection rates, for detecting actual changes in the assay bias are shown in Figs. 2–4. The inherent assay bias is assumed to be zero, which is approximately true for many assays. For an assay with a total CV of 3% (Fig. 2), the power of the 4-6-20 procedure is much lower than that of the Shewhart procedure, which is slightly less than that of the multivariate procedure. The 4-6-20 procedure has negligible power (<1%) to detect changes in assay bias as large as 15%, whereas the Shewhart and multivariate procedures have excellent power (>99%) to detect changes in assay bias of 10% or more. The 4-6-20 procedure has good power (say, 80%) to detect a change in bias of 21%, whereas the Shewhart and multivariate procedures have good power to detect changes of 7 and 5%, respectively.

For an assay with a total CV of 8% (Fig. 3), the power of the 4-6-20 procedure is less than that of the Shewhart procedure, which is less than that of the multivariate procedure. The bias changes that can be detected with good power (say, 80%) are approximately 21, 19, and 14% for the 4-6-20, Shewhart, and multivariate procedures, respectively.

For an assay with a total CV of 15% (Fig. 4), the power of the 4-6-20 procedure is greater than that of the Shewhart

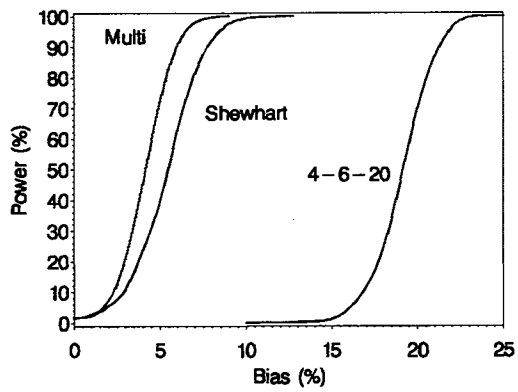

Fig. 2. Power of three QC rules to detect actual changes in assay bias from an inherent 0% bias. Inherent total CV = 3%.

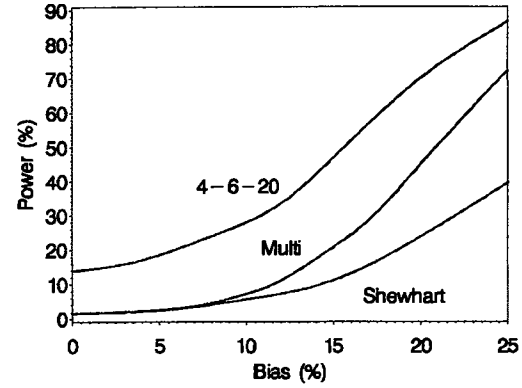

Fig. 4. Power of three QC rules to detect actual changes in assay bias from an inherent 0% bias. Inherent total CV = 15%.

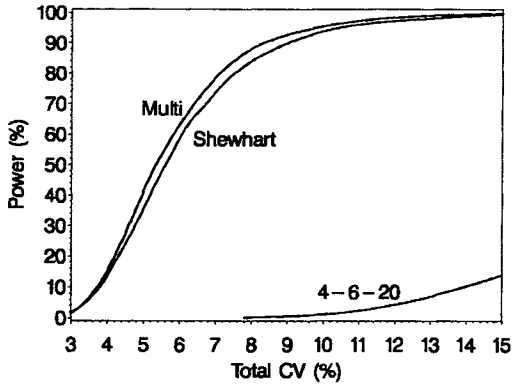

Fig. 5. Power of three QC rules to detect actual changes in assay CV from an inherent 3% total CV. Inherent bias = 0%.

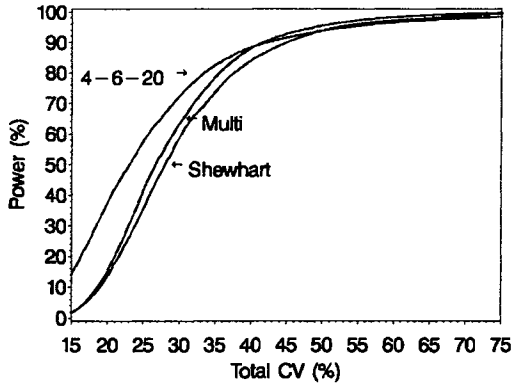

Fig. 7. Power of three QC rules to detect actual changes in assay CV from an inherent 15% total CV. Inherent bias = 0%.

and multivariate procedures, but this advantage is the result of the high false rejection rate of 14%.

Power to Detect Changes in Assay Variability

Power probabilities, or rejection rates, for detecting actual changes in the total CV are shown in Figs. 5–7. For an assay with an inherent total CV of 3% (Fig. 5), the power of the 4-6-20 procedure is much lower than that of the Shewhart procedure, which is slightly less than that of the multivariate procedure. If the total CV increases 5-fold (to 15%) the 4-6-20 procedure has only 12% power to detect this change in total CV, whereas the Shewhart and multivariate procedures have good power (>80%) to detect 2.5-fold (to 7.5%) or greater increases in total CV.

For an assay with an inherent total CV of 8% (Fig. 6), the power of the 4-6-20 procedure is lower than that of the Shewhart procedure, which is slightly less than that of the multivariate procedure. If the total CV increases twofold (to 16%) the 4-6-20 procedure has only 18% power to detect this change in total CV, whereas the Shewhart and multivariate procedures have moderate power (60%) to detect this change.

For an assay with an inherent total CV of 15% (Fig. 7), the power curves of the three procedures are similar, but the 4-6-20 procedure has a high false rejection rate of 14%, compared to the 1.6% rate of the two statistical procedures.

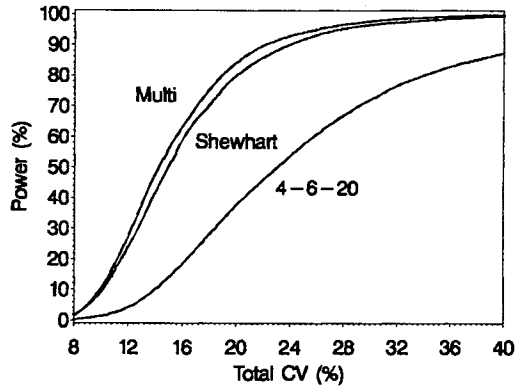

Fig. 6. Power of three QC rules to detect actual changes in assay CV from an inherent 8% total CV. Inherent bias = 0%.

DISCUSSION

A basic flaw with the 4-6-20 procedure is that the frequency of falsely rejected runs depends upon the inherent bias and variability of the assay. For assays with moderate (or greater) inherent bias and variability, the false rejection rate is too high, thus requiring the unnecessary repetition/rejection of runs. For assays with low bias and variability, the false rejection rate is too low, thus compromising the power to detect important deteriorations in assay performance. To remedy these problems, the Shewhart or multivariate statistical QC procedure should be used, with a fixed false rejection rate of acceptably small size.

These arguments are similar to those which led to the dismissal of the 75/75 rule for bioequivalence assessment. Haynes (5,6) and Metzler and Huang (7) showed by simulation that the probability of incorrectly rejecting bioequivalence of two formulations that are truly equal in average bioavailability depends on the intersubject variability of the formulations. Recently, Dobbins and Thiyagarajan (8) demonstrated this link between the significance level (incorrect rejection rate) and variability by placing the 75/75 rule in the framework of a statistical hypothesis test.

It may be appropriate to establish criteria to allow acceptance of runs wherein a statistical QC rule has been violated, but the analytical performance change is considered small relative to the scientific need. Factors to consider include the type and magnitude of the rule violation, the use of the assay results, the study design, and the type of statistical analyses to be performed. For example, the statistical error term in a crossover study includes only assay and intrasubject variability, whereas the error term in a parallel study also includes intersubject biological variability. Thus, an analytical performance change of a given type and magnitude will generally be more serious for a crossover study than for a parallel study.

ACKNOWLEDGMENTS

The author is grateful to Yakov Fuxman for generating the simulations and the graphs and to Fred Snikeris for helpful comments.

REFERENCES

1. V. P. Shah *et al.* Analytical methods validation: Bioavailability, bioequivalence, and pharmacokinetic studies. *J. Pharm. Sci.* **81**:309–312 (1992).
2. W. A. Shewhart. *Economic Control of Quality of Manufactured Product,* D. Van Nostrand, New York, 1939.
3. H. Hotelling. Multivariate quality control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis (eds.), *Techniques of Statistical Analysis,* McGraw–Hill, New York, 1947, pp. 111–184.
4. SAS Institute Inc. *SAS Language: Reference* Version 6, 1st ed., SAS Institute, Cary, NC, 1990.
5. J. D. Haynes. Statistical simulation study of a new proposed uniformity requirement for bioequivalence studies. *J. Pharm. Sci.* **70**:673 (1981).
6. J. D. Haynes. FDA 75/75 rule: A response. *J. Pharm. Sci.* **72**:99–100 (1983).
7. C. M. Metzler and D. C. Huang. Statistical methods for bioavailability and bioequivalence. *Clin. Res. Pract. Drug Regul. Affairs* **1**:109–132 (1983).
8. T. W. Dobbins and B. Thiyagarajan. A retrospective assessment of the 75/75 rule in bioequivalence. *Stat. Med.* **11**:1333–1342 (1992).