Short communication

# Metrologic approaches to setting acceptance criteria: Unacceptable and unusual characteristics

Walter W. Hauck*, Darrell R. Abernethy, Roger L. Williams

*US Pharmacopeia, Documentary Standards Development, 12601 Twinbrook Parkway, Rockville, MD 20852-1790, United States*

ABSTRACT

Decisions regarding acceptance criteria in regulatory or compendial contexts are among the most difficult to make. Acceptance criteria aid in the identification, on the one hand, of materials with *unacceptable* characteristics that should not pass the tests and procedures or, on the other hand, of *unusual* characteristics that indicate materials that are unlikely to pass the tests and procedures. For relatively complex procedures metrological approaches can differentiate between intra- and inter-laboratory variation and clarify unacceptable and unusual data. Such testing requires collaborative studies in which each participating laboratory essentially compares itself to the other laboratories in the collaborative study. Laboratories that use the reference standard established by the collaborative study are conducting a performance verification test in which they compare their capabilities to those of laboratories in the collaborative study. This paper considers aspects of a series of complex issues involving unacceptable/unusual characteristics primarily in the context of USP's work but with implications for manufacturing science via considerations of process capability and Quality by Design and to measurement science. Ultimately, acceptance criteria support the availability of good quality, safe, and effective medicines for patients and consumers.

© 2008 Elsevier B.V. All rights reserved.

## 1. Introduction

The United States Pharmacopeia (USP) creates publicly available documentary and physical standards that allow testing by first parties (e.g., manufacturers), second parties (e.g., purchasers), and third parties (e.g., independent bodies such as governments) to ensure the quality of a medicine (drug or biologic) and its ingredients [1]. USP's documentary standards are expressed in monographs that contain introductory requirements that are followed by the public specification—the tests, procedures, and acceptance criteria that form the core of the monograph. Of particular relevance to this paper is that a key part of any specification is its acceptance criteria—the pass/fail boundaries that determine whether a medicine and its ingredients are suitable for their intended uses.

Acceptance criteria can be identified as falling in one of two general classes. In the first are acceptance criteria that aid in the identification of unacceptable materials by identifying *unacceptable* characteristics, namely those that identify materials that *should not* pass the necessary tests and procedures for safety and/or efficacy reasons. In contrast are acceptance criteria that identify *unusual* characteristics, that is, materials that are *unlikely* to pass the tests

and procedures because of, for example, poor process control. A major challenge is to know which class is applicable in a given context.

The purpose of this paper is to examine these challenges for setting acceptance criteria in the context of USP's standards-setting activities. Although the considerations are applicable to all of USP's activities (medicines and their ingredients, dietary supplements and their ingredients, and food ingredients), focus will be on solid dosage forms. Particular attention will be given to collaborative studies that help USP develop reference materials that can be used in performance tests for procedures that help differentiate unusual and unacceptable drug products.

## 2. Limits: unacceptable values in testing

One general approach to establishing upper and lower acceptance limits is based on the idea of setting limits to exclude unacceptable results—ones that yield negative therapeutic benefit. For bioequivalence, as an example, the US standard for acceptance is that the ratio of (geometric) means of the test and reference formulations' bioavailabilities should be within 80–125% [2]. This reflects a clinical judgment about how two dosage forms, either with unique components and composition and/or arising from separate methods of manufacture, can differ without impairing the efficacy or safety for the consumer.

* Corresponding author. Tel.: +1 301 816 8390; fax: +1 301 816 8373.
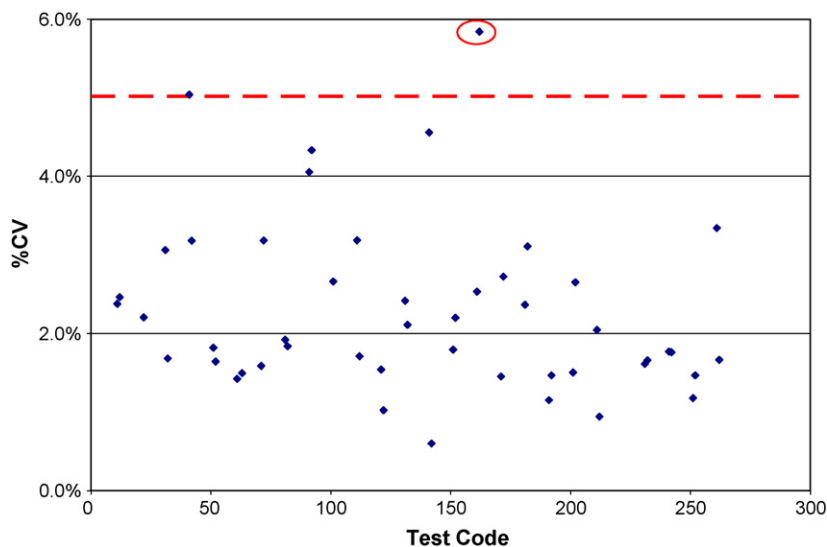*E-mail address:* wh@usp.org (W.W. Hauck).

**Fig. 1.** Example of a control chart. The test code is an arbitrary code identifying a test of six tablets. The dashed line is the upper control limit (5.0%). The %CV of the circled point (5.8%) exceeds the control limit.

Data on which these acceptance limits are set optimally arise during the drug development process. If available, these data help define the therapeutic window (range) for a medicine in the population and support regulatory and compendial decisions about acceptance criteria. Unfortunately, such data are frequently lacking for a medicine, even ones studied recently. Further, therapeutic windows for an individual are almost certainly not available early in drug development. These can be much narrower than the population window, as is the case, for example, for warfarin. The goal of a quality specification is to ensure consistency in daily dosing for an individual, so individual therapeutic windows for a medicine would be especially valuable. In the absence of needed data and information, USP sets its acceptance criteria as best it can, frequently relying on publicly available information from FDA, manufacturers, and the biomedical literature.

## 3. Limits: unusual values and capability-based approaches

Instead of relying on ranges based on an externally derived standard when setting acceptance criteria, scientists alternatively can use a capability-based approach that relies on the variability of the test procedure. This leads to acceptance criteria that arise from the capability of the procedure and thus identifies unusual analytical results. The approach also is used to set acceptance criteria for articles for which manufacturing capability define acceptance criteria. The following elaborates on this approach.

One common technique in quality control is the use of control charts that follow some quantity over time [3]. Control limits often are set to indicate when an unusual value has occurred. The charts also can be examined for trends. Although control charts are mentioned here in the context of manufacturing, they have broad applicability. For example, hospitals use similar methods to track outcomes such as infection rates over time. USP collaborative studies also use control charts; see Fig. 1 for an example.

In Fig. 1, each point is the percent coefficient of variation (%CV) for a dissolution experiment involving six tablets. The x-axis is an arbitrary code that identifies the experiment. The data of Fig. 1 thus are not a time series, as in manufacturing applications, but still they remain a useful means of identifying unusual values. The control limits for this chart were set at probability values corresponding

to what would be ±3 sigma for means, i.e., to 0.00135. Only the upper control limit is shown (dashed line at 5.0%) because, for %CV, low values are acceptable. One experiment resulted in a %CV of 5.8%, which was above the control limit and thus represented an unusual value in the context of these data where the median %CV was 1.9%.

Control limits such as the one for Fig. 1 are statistical tolerance intervals [4] and are usually determined assuming the variability is known rather than estimated from a study. A property of tolerance intervals is their coverage. That is, they cover a specified proportion of the expected distribution of values with a specified degree of confidence. This coverage property makes tolerance intervals desirable for setting acceptance criteria. The coverage allows control of the false failure rate. For example, if one sets the coverage at 99%, then one expects only 1% of future values to fall outside the tolerance interval if the procedure is conducted in a manner similar to that used in setting the limits. Traditional two-sigma limits for a mean correspond to a coverage of approximately 95%, and three-sigma limits correspond to approximately 99.7%. By picking the control limits or tolerance interval coverage, a manufacturer or equivalent body can specify how often it can expect to fail if the process or procedure remains unchanged.

Missing from capability-based approaches, however, is consideration of what should and should not fail. A capability-based approach can possibly lead to acceptance criteria that are either so broad that they raise consumer concerns or are narrower than needed (because analytical variability usually is less than clinical variability). For the data of Fig. 1, for example, all the %CV's are small for dissolution experiments. USP General Chapter *The Dissolution Procedure: Development and Validation* <1092> suggests that a %CV exceeding 10% is highly variable [5]. By that external standard, all the data of Fig. 1 could be considered "acceptable" even though one experiment is identified as "unusual". Another disadvantage of capability-based approaches is that if the process or method changes, then the acceptance criteria must be reevaluated.

## 4. Proficiency approaches, collaborative tests, and acceptance criteria

The laboratory tests specified in *USP* monographs range from relatively simple (high-performance liquid chromatography) to rel-

atively complex (dissolution of solid oral dosage forms), which has important implications for setting acceptance criteria. For dissolution testing the sample preparation step involves placing the tablet in a vessel of solvent, stirring that solution, and taking samples at specified times. When they are performing such complex procedures, laboratory personnel find value in periodic competency testing. For dissolution, following requests from industry and FDA, in 1979 USP introduced RS tablets, formerly termed calibrator tablets, for use in periodic performance verification tests (PVT) of dissolution, formerly termed an apparatus suitability test. This typically is conducted every six months [6].

As a proficiency test, PVT is in accord with ISO Guide 43-1, which describes proficiency testing as the use of interlaboratory comparisons to "determine the performance of individual laboratories for specific tests or measurements and to monitor laboratories' continuing performance" [7]. The USP PVT is an interlaboratory activity because the acceptance limits are set from a collaborative study conducted on new or continuing lots of specially prepared RS materials. Earlier considerations herein regarding setting acceptance criteria apply both as well to PVTs. USP PVT acceptance criteria are currently set using a multiple of the reproducibility standard deviation from the collaborative study and hence are a capability-based approach. Failing results at an individual laboratory are those that are then unusual relative to the results from laboratories participating in the collaborative studies. Statistically, this actually is a test of the null hypothesis of no difference between the individual laboratory and the collaborative results versus an alternative of some difference. If the intent of a PVT really is to demonstrate similarity to the collaborative results, then these statistical hypotheses are not appropriate, and a statistical equivalence test is called for. In an equivalence test, the null hypothesis becomes one of dissimilarity and the alternative is that of similarity. Equivalence testing then leads to the *unacceptable* approach to setting acceptance criteria discussed earlier. That is, acceptable differences define what "similar" results are.

Variability (capability) and safety/efficacy considerations are not mutually exclusive. For example, Hauck et al. proposed an approach to setting dissolution criteria based on the variability in dissolution observed in clinical trials and similar batches [8]. For an approved drug product the variability in clinical trial batches obviously was acceptable, so criteria for the approved product can be set to ensure little, if any, change from clinical trial materials.

As an example of how USP could use the *unacceptable* approach, one can consider vertical diffusion cells for measuring the rate of release from semi-solid topical dosage forms (a candidate PVT under consideration by USP). FDA has set limits of 75–133% for a ratio of medians to compare two formulations—a decision that presumably reflects some understanding of how dissimilar two formulations can be without adversely influencing safety and efficacy outcomes [9]. USP might simply apply this acceptance criterion in a PVT designed to assess the integrity of the use of the vertical diffusion cell to assess laboratory performance. The alternative course would be to determine acceptance criteria based on *unusual* values, as now occurs for the PVT described in General Chapter *Dissolution* <711>.

## 5. Discussion

The heart of a monograph lies in its tests, procedures, and acceptance criteria—and of these probably the most challenging to set are the acceptance criteria. For the most part, these are default criteria required to be met over the shelf-life of the article, usually 98.0–102.0% for the drug substance and 90.0–110.0% for the drug product. Are these acceptance criteria sensible? Because of USP's application of a PVT to in vitro performance, PVT acceptance criteria might better be based on unacceptable results, because in vitro performance is used increasingly to define in vivo outcomes. For the current *USP* dissolution PVT for a non-solution orally administered dosage form, acceptance criteria based on those for content uniformity (85–115%) or bioequivalence (80–125%) thus might be reasonable.

Although we have focused on USP applications, the considerations have broad application. For example the design space concept ultimately rests on a set of specifications and their acceptance criteria. How should those criteria be set? Capability-based approaches to determining acceptance limits are a relatively easy means for setting acceptance criteria because the data arise readily in a collaborative study or in the manufacturing process. In contrast, setting limits based on prespecified *unacceptable* values is more difficult because the latter depend on an external standard, such as an understanding of the therapeutic window in a population or an individual that is usually not established and likely will differ for each article. Because capability-based approaches identify the *unusual*, they are reasonable for acceptance criteria whenever the unusual is also unacceptable. In manufacturing quality control, for example, the goal is to identify change. Hence, unusual is unacceptable because it may indicate that something in the manufacturing process has changed. If a pharmaceutical manufacturing process undergoes change, manufacturers must have prespecified definitions of unusual (which is, in this case, unacceptable) departures from the design space that warrant interventions up to and including batch rejection.

One foundation for choosing between unusual or unacceptable as the basis for acceptance criteria is the nearness to the consumer. Thinking this way, manufacturers can develop drug product criteria that are related in some way to safety and efficacy. There would similarly be less justification for criteria for the substance to be related to safety and efficacy. Criteria for the substance might, instead, be one step removed by being related to the criteria for the product. That is, the acceptance limits for the substance might be based on a percentage of those for the product.

Overall, decisions regarding acceptance criteria, either in a regulatory or compendial context, are among the most difficult to make. At times they are made in isolation without consideration of allied criteria, and at other times they may appear to be set arbitrarily. USP is actively conducting metrologically based experiments and is carrying on discussions with industry and regulatory personnel to chart scientifically valid approaches that will help practitioners and patients obtain good-quality, safe, and effective medicines and also help regulatory personnel comprehend the compendial bases of decision-making.

## Acknowledgments

## References

[1] National Institute of Standards and Technology, Fed. Regist. 65 (2000) 48894–48902.
[2] FDA, Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations, 2002, accessed at www.fda.gov/cder/guidance/4964dft.pdf (accessed 29 August 2008).
[3] D.C. Montgomery, Introduction to Statistical Quality Control, 4th ed., Wiley, New York, 2001.

[4] G.J. Hahn, W.Q. Meeker, Statistical Intervals, A Guide for Practitioners, Wiley-Interscience, New York, 1991.
[5] US Pharmacopeia, USP 31–NF 26, The Dissolution Procedure: Development and Validation, US Pharmacopeia, Rockville, MD, 2008.
[6] G. Deng, A. Ashley, W.E. Brown, J.W. Eaton, W.W. Hauck, L.C. Kikwai, M.R. Liddell, R.G. Manning, J.M. Munoz, P. Nithyanandan, M.J. Glasgow, E. Stippler, S.Z. Wahab, R.L. Williams, Pharm. Res. 25 (2008) 1100–1109.
[7] ISO, Guide 43-01, Proficiency Testing by Interlaboratory Comparisons—Development and Operation of Proficience Testing Schemes, 2nd ed., ISO, Geneva, Switzerland, 1997.
[8] W.W. Hauck, T. Foster, E. Sheinin, T. Cecil, W. Brown, M. Marques, R.L. Williams, Pharm. Res. 22 (2005) 182–187.
[9] V.P. Shah, S.W. Shaw, D.D. Norton, J. Elkins, R. Hanson, Pharm. Forum 32 (2006) 1590–1596.