# Regulatory Focus

Tom Layloff

## Do You Believe in Statistics?

A reader asked that specification setting be addressed in this column. This article is the first of several that will be presented on the topic. The process of limit-setting is one of the most interesting aspects of analytical testing and product regulation. It involves the following components:

1. Which product attributes and limits are critical to establish product quality?
2. How should those attributes be assessed?
3. What variance levels are acceptable?

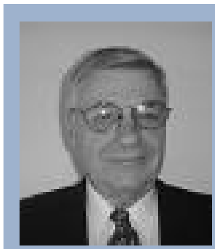
At first glance, this seems to be an area that would be relatively straightforward, but it sometimes is fraught with complications linked to personal biases, historical anachronisms, technological accessibility, politics, market barriers, and push–pull beaucratic posturing.

Problem 1 above would seem to be the most easily resolved. Physicians and other health practitioners should determine which quality attributes are critical to ensure the safety and efficacy of the product. Toxicologists should review the toxicology data on the active pharmaceutical ingredients (APIs) and excipient impurities to establish rational daily intake-based limits that keep the risks within acceptable bounds. Fortunately, most pharmaceutical products have a relatively wide use range that does not require critical administration. Similarly, API precursors and degradants generally are not significantly more toxic than the API, and excipients are frequently food-related products present in the normal diet at much higher levels than from pharmaceutical sources. This topic will be the subject of a later article.

Point 2 would also seem to be relatively straightforward. However, the answer to this question suffers from the old adage, "Call the carpenter and he brings a hammer; call the plumber and he brings a wrench." The answer depends in part on which technologist you pose the question to and to whom the answer will be given—push–pull bureaucratic posturing. If you ask a mass spectrometer expert to make a suggestion, the MS will likely be the tool of first choice. Similarly, if the reviewing regulator has a strong MS background, he or she will likely prefer an MS assessment because he/she is more knowledgeable in that technique. Again, this failing is not catastrophic in large part because the various assessment technologies yield similar answers—frequently, however, at markedly different costs. This topic will also be the subject of a later article.

> "THE PROCESS OF LIMIT-SETTING IS ONE OF THE MOST INTERESTING ASPECTS OF ANALYTICAL TESTING AND PRODUCT REGULATION."

Point 3 is the subject of this article because the available assessment technology sets the baseline for any limit-setting efforts. All assessment technology pro-



***Dr. Tom Layloff** is Principal Program Associate in the Center for Pharmaceutical Management, Management Sciences for Health (MSH), which addresses pharmaceutical quality issues in international commerce and developing nations; and is Adjunct Professor of Chemistry, St. Louis University (Missouri). Prior to joining MSH he was employed by the United States Pharmacopeia (USP) as Vice-President and Director of the Pharmaceutical Division. He served in the U.S. FDA Center for Drug Evaluation and Research (CDER) as Associate Director for Standards Development and as Director of the FDA's leading pharmaceutical surveillance laboratory (St. Louis, MO). He was elected to the USP Committee of Revision and he served on two Chemistry Revision Sub-Committees and as Chair of the General Chapters Sub-Committee, Member of the Reference Standards Sub-Committee, Member of the Chemistry Revision Sub-Committee, and Member of the Division of Standards Development Executive Committee (policy-setting body for USP standards) and as Chair of that Committee. He is very active in the WCBP (formerly Well-Characterized Biotechnology Pharmaceuticals) symposium series where he is serving as Co-Chair of the 2002 meeting and Chair of the WCBP Permanent Organizing Committee. Dr. Layloff is Past-President and Fellow of AOAC International and Fellow of the American Association of Pharmaceutical Sciences. He is a member of the Sigma Xi and Phi Lambda Upsilon honorary societies. He received BA/BS degrees in Chemistry and an MS degree in Organic Chemistry from Washington University (St. Louis, MO) and a Ph.D. in Analytical Chemistry from the University of Kansas (Lawrence). Please send comments or topics suggestions for this column to tom@layloff.net; home page: www.layloff.net. The author would like to thank Dr. Ferria D. Harrison for helpful suggestions and comments.*

cesses have inherent variabilities that arise from random and systematic errors. Each step in an analytical process contributes an uncertainty component* to the combined standard uncertainty of the overall process. In pharmaceutical analyses, the individual uncertainty components generally are not estimated and the combined standard uncertainty is used as the criterion for method acceptability. In addition, pharmaceutical assessment protocols generally require the use of a procedural standard reference material containing the analyte of interest, which, along with relatively wide acceptance limits, makes these protocols legally defensible.

However, in the assessment of limits, there seems, in some instances, to be an inconsistency between the specified product limits and the acceptable performance of the assessment technologies that give rise, at a minimum, to analytical inefficiencies. For example, it is interesting to consider a case that appears in the *United States Pharmacopeia* (USP 24-NF 19, 2001).** The *USP* specifies the following chromatographic requirement: "Replicate injections of a standard preparation used in the assay or other standard solution are compared to ascertain whether requirements for precision are met. Unless otherwise specified in the individual monograph, data from five replicate injections of the analyte are used to calculate the relative standard deviation, $S_r$; if the r requirement is 2.0% or less, data from six replicate injections are used if the relative standard deviation requirement is more than 2.0%."[1] This concept is carried forward into the *USP* monograph requirements, e.g., for the active pharmaceutical ingredient: "Acyclovir contains not less than 98.0 percent and not more than 101.0 percent of $C_8H_{11}N_5O_3$, calculated on the anhydrous basis. . . .Chromatographic system (see Chromatography <621>). . . the relative standard deviation for replicate injections is not more than 2.0%." A maximum assessment relative standard deviation of not more than 2% statistically means that, in an array of analyses, about two-thirds of the results would lay between 98 and 102% (±1 standard deviation) while approx. one-third would lay outside those limits; 1 of 20 results would lay outside 96–104% (±2 standard deviations), and 3 of 1000 would lay outside 94–106% (±3 standard deviations). For standard deviation ranges, see Ref. 2. Therefore, when considering an array of chromatographic analytical data obtained through replicate injections with the above acceptance limits and at the maximum allowable assessment standard deviation, one would expect that more than one-third of the individual analyses would fail the upper or lower limit, with the larger number failing the upper limit due to the asymmetry (+1% and –2%). In regulatory parlance, the failing data are called out-of-specification (OOS) events, which requires follow-up investigations to ascertain the source of the OOS. In this instance, the OOS events occur because the product acceptable limits and assessment tolerances are not well linked. However, if a number of individual analyses are averaged, there will be convergence to a more accurate value, if the assessment bias is small. (See *Appendix 1* for examples of replicate analyses requirements.)

An analogy would be a shooting gallery with a rifle having a poor focus. It is difficult to hit a small target with a poor focus, but if one makes enough attempts, the target will be hit on the average, although none of the individual attempts may have hit it. In the above instance, an assessment method would need to have a relative standard deviation of 0.5 to hit the targeted 98–101% 997 times out of 1000, e.g., [–2] (100%–98% = 2% below the target) + [+1] (101%–100% = 1% above the target) = 3 or ±1.5%, with respect to the symmetric target. For the 1.5% to be equal to ±3 relative standard deviations, the relative standard deviation would have to be 0.5%. In the above case, to have the analytical result be outside the 98–101% (sample at 99.5%) range, only 3 times in 1000 would the assessment technology need to have a relative standard deviation of 0.5%, i.e., have a much better focus.

"ALL ASSESSMENT TECHNOLOGY PROCESSES HAVE INHERENT VARIABILITIES THAT ARISE FROM RANDOM AND SYSTEMATIC ERRORS."

The relative standard deviation obtained in this within-laboratory series of experiments is called the repeatability (*Appendix 2*) relative standard deviation (RSD). In the above example, the statistical limits are for replicate injections only. If the analytical process also involved separate weighings, dilutions, etc., the combined standard uncertainty would be very similar to the uncertainty component arising from the chromatographic process alone, e.g., weighing and dilution uncertainty components are typically less than one part per thousand (0.1%), which generally do not contribute significantly to the overall error budget for chromatographic analytical methods. To better assess the robustness and measurement uncertainty of analytical methods, AOAC International (Gaithersburg, MD, formerly the Association of Official Analytical Chemists) has established a highly structured, multilaboratory, multisample collaborative study protocol.[4] This multilaboratory protocol allows the assessment of the among-laboratory RSD, which is called the reproducibility (Appendix 2) RSD in addition to the repeatability. In the AOAC International protocol, the same samples are analyzed using the same procedures and reference materials in the participating laboratories. The originator of the study then critically reviews the collaborative study results before they are compiled and statistically evaluated. After evaluation, a collaborative study report is prepared for submission to the AOAC for publication.

Dr. William Horwitz, in the late 1970s, conducted a retrospective review of published pharmaceutical collaborative studies to determine if there were significant

*This nomenclature is taken from the EURACHEM (www.eurachem.bam.de/) guide on "Quantifying uncertainty in analytical measurement, 2nd ed. (2000). The text is available to download in the html format at www.measurementuncertainty.org or pdf format at www.eurachem.bam.de/guides/quam2.pdf. The guide was produced by a joint EURACHEM/CITAC [Co-operation on International Traceaility in Analytical Chemistry] Measurement Uncertainty Working Group in collaboration with representatives from AOAC International, IAEA [International Atomic Energy Agency], and EA [Environmental Agency].

**In the United States and a number of other adopting countries, the *USP* is the legislated benchmark for pharmaceutical products named therein, and those products must conform to the cited standards using the methods specified in the monographs.

Table 1

**Results of repeatability and reproducibility findings from collaborative studies on different APIs**

| | **Com-** | | **Repeat-** | **Reproduc-** |
|---|---|---|---|---|
| | **pounds** | **Studies** | **ability** | **ibility** |
| **Method*** | **(no.)** | **(no.)** | **(% RSD)** | **(% RSD)** |
| LC | 26 | 18 | 1.8 | 2.9 |
| GC | 8 | 4 | 1.3 | 2.6 |
| SPCTR | 5 | 5 | 1.1 | 2.5 |
| AUTO | 10 | 7 | 1.3 | 2.2 |
| Total/avg. | 49 | 34 | 1.5 | 2.6 |

(Averages weighted for number of compounds)

*SPCTR, spectrophotometric methods; AUTO, automated methods. Eighteen LC studies were reported for the analysis of 26 compounds. The average repeatability (% RSD) for those studies was 1.8%, and the average reproducibility (%RSD) was 2.9%.

variance differences that could be related to the weight fraction of the analyte in the matrix. In this elegant study, he compiled the repeatability and reproducibility findings from the 34 collaborative studies that had been conducted on 49 different APIs. The results of these findings are presented in *Table 1*.[5]

It is striking to note that the within-laboratory relative standard deviations (repeatability) average 1.5% with a low of 0.5% on one study with an among-laboratory relative standard deviation (reproducibility) of 2.6% with a low of 1% in that previously reported low study, i.e., repeatability of 0.5% and reproducibility of 1% for the lowest case example. On average, the among-laboratory relative standard deviation (reproducibility) is 1.7 times the within-laboratory relative standard deviation (repeatability).

In the *USP* example cited above, the maximum allowable repeatability is 2%. In an among-laboratory assessment, one would therefore expect a reproducibility 1.7 × 2% or 3.4%. This type of relative standard deviation would be expected when comparing results of different testing laboratories using the same methods and samples, e.g., different laboratories in the same or different firms (commerce issue) or a firm's laboratory compared to an FDA laboratory (regulatory issue). Once again, performing replicate analyses and averaging the results of the array can minimize the impact of the poor focus assessment technology.* In this instance, the

> “IF A TIGHTER LIMIT FOR A GIVEN ATTRIBUTE IS DESIRABLE, IT IS USEFUL TO CONSIDER ASSESSMENT TECHNOLOGIES WITH LOWER VARIANCE (BETTER FOCUS) TO REDUCE THE NEED FOR REPLICATE ASSESSMENTS.”

*Taking the mean of replicate sample tests reduces the variability by one over the square root of the number of replicates. In this case, taking the mean of 4 would reduce the variability by 2, which would reduce the error from about one-third to about one-twentieth.

among-laboratory assessment with the API acceptance limits poses an interesting problem.

### *Conclusion*

Analytical acceptance limits should not be set to be more stringent than that reasonably attainable with the defined assessment technology, ±3 RSD (among-laboratory relative standard deviation). If a tighter limit for a given attribute is desirable, it is useful to consider assessment technologies with lower variance (better focus) to reduce the need for replicate assessments.* Setting the limits at lower levels without appropriate replicate definitions could lead to unnecessary OOS laboratory findings. A reasonable estimate in pharmaceutical analyses is that the reproducibility is approx. 1.7 times the repeatability.

*If the assessment technology variance is significant compared to the acceptance limits, it is useful to define a replicate analyses requirement in a standard operating procedure to avoid the appearance of testing into compliance or answer shopping.

## References

1. USP 24-NF 11, 2001. <621> Chromatography, system suitability.
2. www.neatideas.com/stdev.htm.
3. International Vocabulary of Basic and General Terms in Metrology; ISO/TAG 4 1994, as cited in the EURACHEM Guide.
4. www.aoac.org/vmeth/omamanual/omamanual.htm.
5. Horwitz W. JAOAC 1977; 60:1355–63.

**AG/PT**

Appendix 1

How many replicate analyses should be specified in the laboratory standard operating procedures (SOPs)?

It depends.

Considering the *USP* example cited in this article, the operational constraints are:

Acceptance target = the upper acceptance value
–lower acceptance value
Acceptance target = 101%–98% = 3%
Assessment repeatability (RSD) = 2%*

If the within-laboratory acceptable OOS rate** is 1 in 20 failures, then the assessment focus is equal to the acceptance target divided by 4 (±2 standard deviations) = 0.75%. If the within-laboratory acceptable OOS rate is 3 in 1000 failures, then the assessment focus is equal to the acceptance target divided by 6 (±3 standard deviations) = 0.5%.

With an assessment repeatability of 2% and an assessment focus of 0.75%, the number of replicates required to meet an OOS level of 1 in 20 would be equal to the assessment repeatability divided by the assessment focus squared, i.e. $(2/0.75)^2 = 7$. If the OOS level is reduced to 3 in 1000, the replicate requirement increases to $(2/0.5)^2 = 16$.

As noted previously, in order to reduce the incidence of OOS findings, it would be useful to keep the assessment repeatability less than the acceptance target divided by 6 (±3 standard deviations) or increase the acceptance target, thereby lowering the assessment focus to better convergence with the assessment repeatability.

In among-laboratory testing, the replicate test requirements needed to control OOS findings increases by a factor of about 3 (1.7 × 1.7), but that is another story.

*It should be noted that chromatographic procedures, which use the same portioning medium repeatedly, e.g., HPLC and GC, tend to change partitioning characteristics with repeated use, which may require reassessing the method repeatability over time. The tailing and resolution factors become less favorable with continued use, i.e., the partitioning effectiveness lessens.

**The acceptable OOS rate is the number of failures due to the statistical considerations.

Appendix 2[3]

*Repeatability (of results of measurements):* Closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement.

Notes: These conditions are called repeatability conditions. Repeatability conditions include: 1) the same measurement procedure, 2) the same observer, 3) the same measuring instrument used under the same conditions, 4) the same location, and 5) repetition over a short period of time. Repeatability may be expressed quantitatively in terms of dispersion characteristics of the results.

*Reproducibility (of results of measurements):* Closeness of the agreement between the results of measurements of the same measurand carried out; under changed conditions of measurement.

Notes: A valid statement of reproducibility requires specification of the conditions changed. The changed conditions may include: 1) principle of measurement, 2) method of measurement, 3) observer, 4) measuring instrument, 5) reference standard, 6) location, 7) condition of use, and 8) time. Reproducibility may be expressed quantitatively in terms of dispersion characteristics of the results. Results are here usually understood to be corrected results.