

Guidance for Industry

Testing Limits in Stability Protocols for Standardized Grass Pollen Extracts

Additional copies are available from:
Office of Communication, Training and Manufacturers Assistance (HFM-40)
1401 Rockville Pike, Rockville, MD 20852-1448
(Tel) 800-835-4709 or 301-827-1800
(Internet) <http://www.fda.gov/cber/guidelines.htm>

**U. S. Department of Health and Human Services
Food and Drug Administration
Center for Biologics Evaluation and Research (CBER)
November 2000**

TABLE OF CONTENTS

Note: Page numbering may vary for documents distributed electronically.

I. INTRODUCTION	1
II. CURRENT LOT RELEASE CRITERIA	2
III. LOT RELEASE VERSUS STABILITY PROTOCOL	3
IV. A MODIFIED STABILITY PROTOCOL	4
V. RETESTING	5
VI. DEALING WITH TEST FAILURE	6
VII. EXTENSION OF DATING	7

Guidance for Industry

Testing Limits in Stability Protocols for Standardized Grass Pollen Extracts

This guidance document represents FDA's current thinking on testing limits in stability protocols for standardized grass pollen extracts. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. An alternative approach may be used if such approach satisfies the requirements of the applicable statutes, regulations, or both.

I. INTRODUCTION

Standardized grass pollen extracts must demonstrate their potency in accordance with 21 CFR 680.3(e), 610.10, and 600.3(s). Accordingly, manufacturers must demonstrate that their grass pollen extracts maintain potency and are stable until the product's expiration date. The Center for Biologics Evaluation and Research (CBER) is setting forth this document to provide guidance on developing stability protocols for standardized grass pollen extracts. A specific stability protocol that is consistent with CBER lot release is provided, along with all necessary formulas and illustrative numerical examples. This document does not change lot release criteria for these products. To the extent that prior guidance from FDA is or may be interpreted to be inconsistent with this document, this document supersedes such previous guidance. This guidance document finalizes the draft guidance entitled "Testing Limits In Stability Protocols For Standardized Grass Pollen Extracts" that was announced in the Federal Register of August 25, 1997 (62 FR 44975).

Allergenic extracts are a complex mixture of proteins obtained from natural sources. Consequently, a measure of the relative potency with respect to a standard is a critical step in lot release testing for standardized allergenic products. By their nature, such tests are relatively less precise and more costly (in both time and resources) than chemical tests for simple organic and inorganic compounds. Presently, relative potency is primarily determined by Enzyme Linked Immunosorbent Assay (ELISA). As will be shown below, the inherent variability of the measured potency associated with this assay leads to special considerations when devising lot release and stability protocols.

By way of outline and synopsis, Section 2 reviews current CBER lot release criteria and related issues of ELISA test variability. It is shown that when *only* 3 replicates are employed (rather than a 3 replicate/2 replicate sequence), the statistically equivalent range for relative potencies is **0.654-1.530**. Section 3 presents arguments as to why in certain cases of multiple testing, including those of Standardized Allergenic Extracts, the bounds for Stability Studies should be widened with respect to those for Lot Release. Section 4 develops stability bounds based on the Bonferroni adjustment for multiple tests. As an example, it is shown that for a 10-test stability study (two lots; tests on each lot at 6, 12, 18, 24 and 36 months; assays consisting of 3 replicates), the approved limits for relative potency

are **0.568-1.759**. When three lots are included in the study (a total of 15 tests), the limits are slightly expanded to **0.556-1.798**. (It is assumed in all cases that the initial time point satisfies lot release requirements.) Section 5 considers retesting and the current good manufacturing practice prohibition of “testing into compliance”. Section 6 presents options for the manufacturer when a test fails at a time point, and Section 7 explains how to extend dating.

Before proceeding to Section 2, it is useful to discuss, in general terms, and in the context of this document, the statistical notion of *manufacturer’s risk*. This term has a very specific meaning in the context of quality control and, for the nonexpert, is much clearer than related terms from statistical hypothesis testing (i.e., Type I and Type II errors). As with many statistical concepts, however, it can be misleading when used incorrectly or out of context. In the present document, manufacturer’s risk is associated with the cost of losing properly formulated product because of assay variability. In essence, to ensure that potency is consistent, CBER rejects a fraction of lots whose true relative potency is 1.0. This document shows how this risk may be taken into account when determining the number of replicates and the bounds for lot release and stability protocols.

II. CURRENT LOT RELEASE CRITERIA

The determination of relative potency (rp) in the current release testing for standardized allergenic grass pollen extracts is a two step process, which is denoted the “3+2” method. First, the rp is determined from the average of 3 replicates. If this average is between 0.699-1.431 (the 95% confidence interval for the assay, or $\alpha=0.05$), the sample passes. If it is outside this range, two additional replicates are determined. The sample passes if the average of all 5 replicates is within 0.758-1.320 ($\alpha=0.05$); otherwise, the sample fails. For later reference, these limits were obtained as

$$\text{limits} = 10^{\pm z \cdot s / \sqrt{N}} \quad (1)$$

where N is the number of replicates; $\sigma=0.1375$ is the standard deviation of the assay; z is the value of the $1-\alpha/2$ percentile of the normal distribution.

The release criteria also place an upper bound on the sample standard deviation of log relative potency, $s[\log(rp)]$:

$$s[\log(rp)] < s \sqrt{\frac{c_{.01}^2}{N-1}} \quad (2)$$

where $c_{.01}^2$ is the upper 1% critical value of the chi-squared distribution for N-1 degrees of freedom. (N is typically 3 or 5.)

The acceptance characteristics for the “3+2” determination of r_p are listed in Table 1. From this Table, the probability of acceptance for a lot of $r_p=1$ is 0.980, indicating that equivalence (to reference) is being tested at the $\alpha=0.02$, or 98% confidence level (the consequence of successive determinations at $\alpha=0.05$). These results can be used to develop alternative testing protocols with similar characteristics. For example, if *only* 3 replicates are taken, the appropriate $\alpha=0.02$ interval is **0.654-1.530**; this interval is obtained by substituting $z=2.326$ into Eq. (1). The 3 replicate procedure will be utilized for the remainder of this Guideline, because its statistical properties are easier to calculate.

Inherent in CBER’s current release testing is that about 2% of lots whose true r_p equals 1.0 (i.e., the reference itself) will fail the test. This is manufacturer’s risk discussed in Section 1. These limits ensure that a lot of $r_p > 3$ or $r_p < 0.3$ has a negligible probability of passing release testing. Likewise, there is an approximately 3% chance of passing lots with $r_p = 2$ or 0.5 (see Table 1).

For an equivalence assay (such as ELISA), the manufacturer’s risk can be reduced by increasing the number of replicates. This is not practical, however, for Standardized Allergenic Extracts: a factor of ten increase in replicate number is required to obtain a significant improvement, availability of serum pool is low, and cost of the test is high.

It is also noted that different analyses are appropriate when the bounds of the release criteria are independent of the assay (e.g., bounds of 80-120% of a set value and an assay precision of 1%).

III. LOT RELEASE VERSUS STABILITY PROTOCOL

Obvious values for upper and lower limits in a stability protocol are the lot release bounds themselves. As shown below, this approach is not feasible for allergenic extracts at present because of assay variability.

The following terms are introduced for notational simplicity:

10-test study. A protocol consisting of 2 lots tested at 5 time points (e.g., 6, 12, 18, 24 and 36 months), assuming, as always, that the initial point (0 months) has passed lot release.

15-test study. A protocol consisting of 3 lots tested at 5 time points.

ideal product. A product in which every lot is equivalent to the reference at all time points (i.e., it has full potency and there is no decay of potency with time).

(Both the *10* and *15-test* studies have been approved as stability protocols for standardized grass pollen extracts.)

Consider now a 10 test study, where failure is designated as the first instance in which a sample tests outside of CBER release limits. Given a probability of failure of 0.02 for each test, the probability that

an ideal product receives full dating is then $(0.98)^{10} = 0.82$. The acceptance rates would be lower for samples whose relative potencies differ from the reference. For example, suppose that the relative potencies of the two lots under study were 0.9 and 1.1; such lots pass CBER lot release with a probability of 0.95. Even if there were *no* decay in either lot, the probability that the grass extract would obtain full dating is $(0.95)^{10} = 0.60$.

An additional consideration for allergenic extracts is that the final product is frequently a mixture of different grasses. Since the mixture's shelf life is limited by the shortest lived component, the probability that a mixture of, for example, 7 ideal products having a 36 month shelf life is $(0.98)^{70} = 0.24$.

Returning to the terminology of risk, the preceding examples demonstrate how the manufacturer's risk can be increased to very high levels, and the statistical variability in accepted lots can be reduced to levels far below those considered necessary when devising the release criteria. Consequently, it is appropriate to modify the ELISA test limits for purposes of stability testing.

Similar considerations may apply to other products where lot release criteria and assay variability are of comparable magnitude.

IV. A MODIFIED STABILITY PROTOCOL

A statistical method of taking into account an unacceptably high rejection rate associated with multiple testing is to widen the original limits so as to maintain a specified overall probability of acceptance; this is known as the Bonferroni procedure. (This, and other adjustments for multiple comparisons, are discussed in many statistics texts. See, for example, *Principles of Biostatistics*, Pagano and Gauvreau, Duxbury Press, 1993.) Suppose that the $\alpha=0.02$ level is deemed to be an acceptable level for the stability study defined above. This implies that 2% of ideal products will fail (on average) to obtain the full 36 month dating. The Bonferroni procedure adjusts each of the 10 tests to the $\alpha/10=0.002$ level ($z=3.0902$); if there are 15 tests (3 lots in the protocol), each test is performed at the $\alpha/15=0.00133$ level ($z=3.2087$). Substituting the preceding values of z into Eq. (1) leads to the acceptance intervals **0.568-1.759** (2 lots), and **0.556-1.798** (3 lots). Failure at one time by a lot leads to a lot dating of the previous test time.

Table 2 lists the performance of the proposed limits in setting the dating period when applied to three different relative potencies. The results indicate that the ideal product will achieve full dating 98% of the time whether two or three lots are included in the protocol. The manufacturer's risk, however, is lowered as the number of lots under study is increased. In either case, a lot of $rp=0.5$ is rejected very quickly, with less than 3% of the lots remaining after 18 months and 0.2% at 36 months.

The 10-test study (defined in Section 3) for Standardized Grass Pollen Extracts is acceptable to CBER. A stability protocol consisting of more determinations (e.g., a point at 30 months) or fewer determinations (e.g., yearly points only) would have wider or narrower limits, respectively. These are straightforward to calculate from Eq. (1) and a table of the standard normal distribution.

A preparation containing a mixture of grasses should still be dated according to its shortest lived component. Hence, using this method it is anticipated, for example, that $(0.98)^7$ or 87% of mixtures with 7 components would obtain full dating.

The manufacturer could opt for the "3+2" replicate method (with the appropriate tighter intervals), but should specify this in a supplement to its PLA.

Lastly, the date of initiation of extraction is an acceptable time point for the beginning of the expiration dating period (i.e., the zero time point of a stability study).

V. RETESTING

Retesting is an important issue, in that rejection of a small fraction of otherwise acceptable lots is an inherent part of the procedure. The possibility of gross analytical error must also be recognized (although it is also presumed to be rare). Hence, retesting is permitted to establish, for example, the presence of instrument malfunction or degraded reagents. If analytical error can be demonstrated, the results of the original test can be discarded (although an investigation as to the cause of the error should also be carried out).

In general, there are two cases to consider in the context of retesting:

- A) when one of the 3 original replicates is very different from the other 2, causing the average to fall out of specifications;
- B) when all three replicates are within statistical error, but the average is out of specifications.

Case A has already been excluded by the limits placed on the standard deviation in the present release protocol (Section 2); e.g., from Eq. 2, $s[\log(rp)] < 0.2951$ when $N=3$. Consequently, a test with replicates $rp=0.55, 0.85$ and 0.20 (a failing average of 0.45), would be excluded because $s[\log(rp)]=0.32$. (Note that the average has been calculated as $\log(rp)$, and then transformed to rp .) Case A, therefore, need not be considered further when analyzing standardized grass pollen extracts.

Continuing to Case B, analytical error could be demonstrated as follows:

- i) the result of the retest consisting of at least 6 replicates should be within specifications;
- ii) the average of the retest should be significantly different from the original result, as demonstrated by a two sided t-test at the $\alpha=0.05$ level using the observed pooled variance.

Pursuant to the concepts of current good manufacturing practice, averaging the results of the original and retest is not permissible. This restriction eliminates objections associated with "testing into compliance":

if the original test results are not valid, then they should not be used; if they are valid, they cause the lot to fail, as consistent with the protocol.

The upper limit to the number of replicates in the retest is not specified here, but it should be incorporated into the testing SOP. Three replicates are not acceptable because there is no strong scientific reason to choose the second three in favor of the first. Additionally, given the variance of the test, it is difficult to distinguish two means with adequate statistical confidence with a small number of replicates. For example, suppose that the original test failed with replicates $rp=0.55, 0.85$ and 0.35 (an average of 0.547 , calculated from $\log(rp)$), and the retest (3 replicates) resulted in $rp=0.9, 1.0$ and 1.3 (an average of 1.054). Perhaps intuitively, the difference between 0.547 and 1.054 appears so large that the initial test should be rejected. This is not so. The t-test results in a p-value $=0.11$, indicating that the difference is not significant. Consequently, the original result should not be rejected and the lot fails. As a second example, an initial determination (3 replicates) with $rp=0.4$ and $s[\log(rp)]=0.2$, can be rejected if the 6 replicate retest results are $rp>1.026$ and $s[\log(rp)]=0.1375$ ($p<0.05$). The 6 replicates suggested here is a compromise between the cost of retesting and discriminatory power.

The limits for the retest are determined by the number of replicates and can be Bonferroni-adjusted for the total number of tests. For a single test (e.g., lot release) consisting of 6 replicates, the allowed interval for rp is **0.740-1.351** (Eq. 1, $\alpha=0.02$; $N=6$) and $s[\log(rp)] < 0.2389$ (Eq. 2). In a 10 test stability study ($\alpha=0.002$; $N=6$) the limits on rp for the *retested* point are **0.671-1.491**; for 15 tests ($\alpha=0.00133$; $N=6$) these limits are **0.661-1.514**.

VI. DEALING WITH TEST FAILURE

Related to retesting is the question of whether a product can ever obtain full dating if one of the test lots failed at a particular time. CBER allows this possibility, under the following conditions:

- (A) data for the full dating period is provided;
- (B) the statistical confidence is at the $\alpha=0.02$ level.

This could be accomplished by putting additional lots under study and carrying out the appropriate statistical analysis. CBER would also consider data on the failed lots for periods greater than 3 years. For example, if two lots are under test and one registers $rp = 0.560$ at 36 months, the product obtains 24 month dating (assuming that a retest indicates that the result cannot be rejected). If the study is extended to 5 years with determinations at 42, 48, 54 and 60 months, Bonferroni-adjusted limits can be calculated on the basis of 18 tests. This leads to the interval **0.551-1.815**. If the new samples are within these limits, the 36 month sample ($rp=0.560$) no longer causes failure. In fact, the product can be given a shelf life of 5 years.

VII. EXTENSION OF DATING

If a product obtains full 3 year dating, the manufacturer may choose to continue monitoring the stability with the intention of requesting an extension of the dating. In these cases new data can be analyzed using Bonferroni-adjusted values for the total number of *actual* tests. For example, if data for two lots are presented for time points 42, 48, 54 and 60 months to support 5 year dating, an adjustment for 18 tests can be used (i.e., the same limits presented in Section 6).

Table 1. The probability of acceptance for a lot of specified relative potency (rp) using the CBER "3+2" criteria described in the text. These values were calculated from Monte Carlo simulation with ten million samples.

rp	Pr(accept)	rp	Pr(accept)
0.10	0.00000	1.60	0.28166
0.20	0.00000	1.70	0.17780
0.30	0.00000	1.80	0.10663
0.40	0.00114	1.90	0.06118
0.50	0.03372	2.00	0.03375
0.60	0.20868	2.10	0.01803
0.70	0.53281	2.20	0.00933
0.80	0.81436	2.30	0.00474
0.90	0.94792	2.40	0.00230
1.00	0.97955	2.50	0.00113
1.10	0.95404	2.60	0.00055
1.20	0.87433	2.70	0.00026
1.30	0.74291	2.80	0.00012
1.40	0.58055	2.90	0.00006
1.50	0.41895	3.00	0.00003

Table 2. Probabilities of acceptance for products of representative relative potencies at different times and number of lots tested, assuming 5 determinations for each lot.

lots on rp		per test	dating \geq	dating of
test		acceptance	18 months	36 months
2	1.00	0.998	0.988	0.980
	0.75	0.935	0.818	0.716
	0.50	0.241	0.014	0.001
3	1.00	0.999	0.988	0.980
	0.75	0.949	0.855	0.770
	0.50	0.280	0.022	0.002