Product Quality Research Institute Evaluation of Cascade Impactor Profiles of Pharmaceutical Aerosols, Part 3: Final Report on a Statistical Procedure for Determining Equivalence

Received: June 4, 2007; Final Revision Received: July 10, 2007; Accepted: July 10, 2007; Published: November 2, 2007

David Christopher,[1] Wallace Adams,[2] Anthony Amann,[3] Craig Bertha,[4] Peter R. Byron,[5] William Doub,[6] Craig Dunbar,[7] Walter Hauck,[8] Svetlana Lyapustina,[9] Jolyon Mitchell,[10] Beth Morgan,[11] Steve Nichols,[12] Ziqing Pan,[1] Gur Jai Pal Singh,[2,13] Terrence Tougas,[14] Yi Tsong,[15] Ron Wolff,[16] and Bruce Wyka[17]

[1]Statistics, Schering-Plough Research Institute, Kenilworth, NJ
[2]Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD
[3]ACN Pharma LLC, Naperville, IL
[4]Office of New Drug Quality Assessment, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD
[5]Virginia Commonwealth University, Richmond, VA
[6]Office of Testing and Research, Center for Drug Evaluation and Research, Food and Drug Administration, St Louis, MO
[7]Alkermes, Cambridge, MA
[8]US Pharmacopeia, Rockville, MD
[9]Drinker Biddle & Reath LLP, 1500 K St. N.W., Ste 1100, Washington DC, 20005-1209
[10]Trudell Medical International, London, Ontario, Canada
[11]Manufacturing and Supply, GlaxoSmithKline, Zebulon, NC
[12]Industrial Development, sanofi-aventis, Holmes Chapel, Cheshire, UK
[13]Current address: Watson Laboratories, Inc, Corona, CA
[14]Boehringer Ingelheim, Ridgefield, CT
[15]Office of Biometrics, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD
[16]Nektar Therapeutics, San Carlos, CA
[17]Physical & Analytical Chemistry, Schering-Plough Research Institute, Union, NJ

Abstract

The purpose of this article is to report final results of the evaluation of a chi-square ratio test proposed by the US Food and Drug Administration (FDA) for demonstrating equivalence of aerodynamic particle size distribution (APSD) profiles of nasal and orally inhaled drug products. A working group of the Product Quality Research Institute previously published results demonstrating some limitations of the proposed test. In an effort to overcome the test's limited discrimination, the group proposed a supplemental test, a population bioequivalence (PBE) test for impactor-sized mass (ISM). In this final report the group compares the chi-square ratio test to the ISM-PBE test and to the combination of both tests. The basis for comparison is a set of 55 realistic scenarios of cascade impactor data, which were evaluated for equivalence by the statistical tests and independently by the group members. In many instances, the combined application of these 2 tests appeared to increase the discriminating ability of the statistical procedure compared with the chi-square ratio test alone. In certain situations the chi-square ratio test alone was sufficient to determine equivalence of APSD profiles, while in other situations neither of the tests alone nor their combination was adequate. This report describes all of these scenarios and results. In the end, the group did not recommend a statistical test for APSD profile equivalence. The group did not investigate other in vitro tests, in vivo issues, or other statistical tests for APSD profile comparisons. The studied tests are not intended for routine quality control of APSD.

Corresponding Author: Svetlana Lyapustina, Drinker Biddle & Reath LLP, Washington, DC. Tel: 202-230-5179; Fax: 202-842-8465; E-mail: Svetlana.Lyapustina@dbr.com

KEYWORDS: Chi-square, population bioequivalence, particle size distribution, inhaler.

Introduction

Aerodynamic particle size distribution (APSD) is an important in vitro characteristic of orally inhaled and nasal drug products (OINDP), as it may affect the safety and efficacy of such products. Therefore, a US Food and Drug Administration (FDA) 1999 draft guidance for bioequivalence studies[1] recommended a statistical test for comparing APSD profiles of test (T) and reference (R) products, obtained from cascade impactor (CI) measurements, as part of the in vitro bioequivalence determination. The term "profile" in this report refers to the mean and variance of the measured active pharmaceutical ingredient (API) mass measured on each of the individual deposition sites during 30 CI runs. By its very nature, an APSD profile contains information from multiple sites and so is a multivariate (ie, multivariable) measure with no established comparison methods. The term "site" refers to the CI stages and any of the deposition sites within the entire CI train, including inhaler valve stem, mouthpiece, preseparator, and/or induction port (including the mouthpiece connector), since the inclusion of all these sites in the APSD profile comparison was recommended by the FDA draft guidance. Unless noted otherwise, the term "deposition" in this report refers to the in vitro deposition within the CI train and not to the in vivo deposition within the respiratory tract.

The statistical test proposed in the 1999 FDA guidance was based on a chi-square ratio statistic.[2] It was developed, along with the proposed critical value,[3] using the Andersen 8-stage CI (apparatus 1 in the US Pharmacopeia Chapter <601>)[4] applied to albuterol metered-dose inhaler (MDI) data. To investigate the test's applicability to a broad range of other OINDP and profile types, a working group involving scientists from the FDA, industry, academia, and the US Pharmacopeia was established through the Product Quality Research Institute (PQRI).[5] This APSD Profile Comparisons Working Group (the WG) approached the study systematically. First, the WG clarified and created an implementable algorithm for the chi-square ratio test.[6] Next,[7] the WG focused on investigating performance of the chi-square ratio test in the case of identical profiles and a set of 38 T and R profiles[8] simulated to have specific changes on a single deposition site. That work resulted in a series of observations:

- The chi-square ratio test is most sensitive to changes on the deposition sites with highest average deposition. In many cases, reaction of the test to these high-deposition sites limits the test's ability to identify changes potentially important to clinical efficacy that may occur at sites with lower deposition.
- The stability (consistency with regard to small changes in the profile) of the chi-square ratio statistic is greater when the total number of sites is large (eg, 11-13). When the total number of sites is reduced (eg, to 4-7), the chi-square ratio statistic is less stable.
- At least in some situations, the chi-square ratio may decrease (suggesting increasing "similarity" between the profiles) when in fact the difference between T and R profiles increases.
- It may be difficult to select a single critical value for the chi-square ratio that would allow consistent identification of equivalent profiles, especially in situations where equivalence or inequivalence is not immediately apparent.
- For the types of differences for which the test is likely to be used, the chi-square ratio test in itself may not have sufficient discriminating power to detect important differences in particle size distributions.

Based on these findings, and since its main charge was to evaluate the chi-square ratio test, the WG chose not to develop an alternative to the chi-square ratio test but rather to supplement it with another test in order to increase the discriminating ability of the overall statistical procedure. The proposed additional test was based on a general understanding that equivalence of impactor-sized particles is important for the overall performance of an OINDP. Therefore, the supplemental test focused on impactor-sized mass (ISM), which was defined by the WG as the sum of the API mass on all stages of the CI plus the terminal filter, but excluding the initial stage because of its lack of a specified upper cutoff size limit. The population bioequivalence (PBE) method, which is an accepted regulatory method for univariate (single-variable) measures developed by the FDA earlier,[9] was applied to ISM. The combined application of the chi-square ratio test and the ISM-PBE test (referred to as the "statistical procedure" in this report) was then studied using realistic T and R profiles. This article presents methods and results of these studies.

METHODS

Combined Application of Chi-square Ratio and PBE Tests

To evaluate the performance of the statistical procedure, the WG compared statistical outcomes for 55 realistic scenarios[10,11] against an independent decision regarding equivalence (E) or inequivalence (I) of APSD profiles in those scenarios. (The 55 realistic scenarios were generated based on the WG's survey of patterns of changes observed in real products, as explained in more detail below.) Since there is currently no unequivocal evidence linking changes in APSD to clinical outcomes, and therefore there is no "absolute basis" against which to compare the correctness of the statistical procedure, the WG used the judgment of its members (all of whom have worked with these types of products for many years) as an independent assessment. The details of this independent WG assessment are provided online[12] and briefly explained below. For the chi-square ratio test, the published algorithm[2] was implemented through an SAS (Statistical Analysis System) program[13] using the following criterion: for equivalence to be established, the chi-square ratio had to be 7.66 or less.[3] To demonstrate the influence of the change in the critical value, the test was also studied with a critical value of 2.75. The SAS program code to perform the statistical tests used in this study is available from the PQRI Web site.[14]

The PBE approach, including the form of the bioequivalence (BE) upper limit, was developed by the FDA previously.[9] (There is no lower BE limit; small values of the PBE statistic are consistent with equivalence.)

The general form of the PBE statistic is

$$\frac{(\text{Average BE Limit in Natural Log Scale})^2 + \text{Variance Terms Offset}}{\text{Scaling Variance}} \qquad (1)$$

The value of the BE upper limit used here accords with FDA's current thinking to give an upper limit, Θ_P, for PBE of

$$\Theta_p = \frac{\left[(\ln(1.11))^2 + 0.01\right]}{0.1^2} = 2.0891 \qquad (2)$$

This formula contains specific values for (1) the average BE limit, (2) the variance terms offset (ε_p), and (3) the scaling variance ($\sigma_o{}^2$) that were assigned based on the FDA's 1999 draft guidance "Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action,"[1] where the Center for Drug Evaluation and Research (CDER) recommended for specific comparative in vitro tests that the average BE limit not be larger than 90/111 (ie, the ratio of geometric means would fall within 0.90 and 1.11) and a value of 0.90, or essentially equivalently, 1.11, was tentatively recommended as the average BE limit. The variance terms offset value arises to allow some difference among the total variances that may, in practice, be inconsequential because of the low variability of in vitro measurements. The currently recommended value for ε_p is 0.01. The scaling variance value adjusts the BE criterion depending on the R product variance, using mixed scaling. CDER recommended a value of 0.1 for the scaling standard deviation and 0.01 for the scaling variance, and the currently recommended change-over point for mixed scaling is 0.10. (Because the FDA 1999 guidance, from which these numeric values were obtained, is currently a draft, these values should not be assumed to be final for the purpose of preparing applications to the FDA.) Simulated values for ISM-PBE were taken from the APSD T and R profiles and declared equivalent when Θ_P was found to be no more than 2.0891. A more detailed description of the PBE approach is available at the PQRI Web site.[15]

For a given pair of APSD profiles to be declared equivalent according to the overall statistical procedure, results from both statistical tests (chi-square ratio and ISM-PBE) had to indicate equivalence (in standard logic notation, E × E = E). If either of the statistical tests failed to show equivalence, then the statistical conclusion was recorded as "failing to show equivalence" (ie, I × E = I; E × I = I; I × I = I).

Realistic ("Target") Scenarios and Their Evaluation

To study performance of the statistical procedure, the WG developed "target" profiles[7,10,11] that could be subjected to the statistical tests. In the studies reported here, the WG focused on realistic changes likely to be observed in APSD profiles. The patterns of realistic changes were learned through a confidential survey of WG members, who in turn may have surveyed their respective organizations. A total of 14 different patterns of changes were received, and that information was used to generate the set of realistic scenarios for this study. The situations provided in this survey were blinded with respect to the company and product, but they did give numerical and descriptive information about the typical changes in APSD CI profiles observed with orally inhaled products (eg, "an increase in larger particle content in the ISM due to a change in formulation"). These descriptions were applied to the Andersen CI profiles of a real MDI and a real dry powder inhaler (DPI) to produce families of realistic simulated profiles, with interstage correlations, modeled on actual data and with the changes patterned upon the provided real-life scenarios. APSD data were normalized to total recovery as required by the FDA chi-square ratio test.[1]

An example of a family of realistic scenarios is shown in Figures 1, 2, and 3. Scenario 2b (Figure 2) represents the observed change based on actual data (compared with R, the T profile shows increased deposition at sites 8 and 9, and decreased deposition at sites 10 and 11). To assess the discriminating abilities of the test, profiles were created in which this difference was either minimized (scenario 2a, Figure 1) or exaggerated (scenario 2c, Figure 3). Other examples of scenarios are provided in the detailed minutes of the WG.[16] In this fashion, from the 14 original scenarios and systematically designed changes, the WG obtained 55 simulated realistic profiles, which are available at the PQRI Web site in detail[10,11] and in summary form.[17] While these 55 scenarios are not inclusive of all possible profiles for all OINDP, they were deemed, based on the experience of the WG members, to be sufficiently representative of the scope of formulations

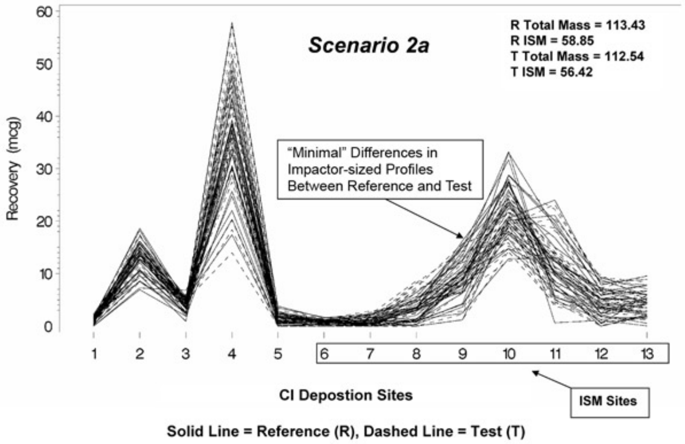

Figure 1. Profile of Scenario 2a, showing deposition and variability on each of 13 cascade impactor sites. ISM indicates impactor-sized mass.

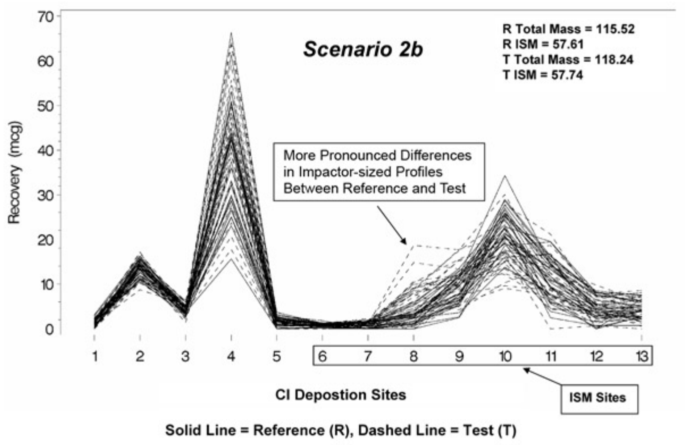

Figure 2. Profile of Scenario 2b, showing deposition and variability on each of 13 cascade impactor sites. ISM indicates impactor-sized mass.

on the market or in development. Namely, the underlying profiles (from an MDI and a DPI) were distinctly different, the applied patterns of change were varied qualitatively (across the 14 different scenarios provided through the survey), and the systematic minimization and exaggeration of changes explored quantitative extremes within those scenarios. The goal of the study was to examine test behavior rather than product behavior, so the number of underlying profiles, while limited, was believed sufficient to provide a realistic assessment of the performance of the statistical procedure.

Each of the 55 simulated profiles was evaluated by the WG, which comprised pharmacologists, pharmacists, chemists, aerosol physicists, regulatory reviewers, product developers, and statisticians, all with experience in the area of OINDP regulation. Of the 14 evaluations received from the WG members, 8 were purely qualitative and 6 were based on some quantitative procedure; all are described in the detailed compilation of the WG evaluation.[12] For the evaluation, WG members assumed that certain changes in APSD profiles could be consistently translated into changes in pulmonary delivery, which in turn might be translated into changes in clinical outcomes, and that the administered drug(s) could be either bronchodilators (beta-adrenergic or anticholinergic) or anti-inflammatory steroids. They were instructed further to attempt to adopt a "regulatory perspective" and to declare each pair of simulated T and R profiles as E or I. The frequency of each decision (E or I) (eg, the fraction of the WG members who considered the profiles to be equivalent) was calculated across the WG for each scenario and later used as a comparator for the E or I decisions made by the statistical procedure.

The statistical procedure was applied to the same 55 profiles. For each of the 55 scenarios, 1000 pairs of simulated R and T profiles were generated and the chi-square ratio and ISM-PBE tests were applied to each pair. The results were then summarized as frequencies of the answers of a given type (E or I) for each test separately and for the combined tests. The outcomes of the WG members' assessments were then compared with the outcomes of the statistical evaluations.

Results and Discussion

Chi-squared Ratio Applied Alone and in Combination with PBE Test

The distribution of the 95th percentiles of the chi-square ratio means, obtained from 1000 simulations applied to each of the 55 scenarios, is presented as a box-and-whisker plot in Figure 4. The edges of the boxes are at the 25th and 75th percentiles of each distribution (the interquartile range); the central horizontal line in each box is drawn at the 50th percentile (median); the vertical lines (or "whiskers") extend from the box as far as the data extend, to at most 1.5 times the interquartile range; and the more extreme data values are represented by crosses.

A reminder: for a given T-to-R comparison, a chi-square ratio below the critical value indicates that the test declares these T and R to be equivalent, while any value above the critical value indicates that T and R are failing to show equivalence according to the chi-square ratio test. In Figure 4, a solid horizontal line corresponding to the critical value of 7.66 (as proposed by the FDA) was well above the entire distribution of chi-square ratio results for scenarios 1 through 36 and substantially above most of the distribution for the rest of the 55 scenarios (except for scenarios 37 and 47, where it was above a portion of the distribution). This means that for scenarios 1 to 36, the chi-square ratio test showed no discrimination and declared all of the T and R pairs to be equivalent. For scenarios 37 to 55, the chi-square ratio showed some discrimination but mostly declared the T and R pairs to be equivalent. This shows that when the chi-square ratio's

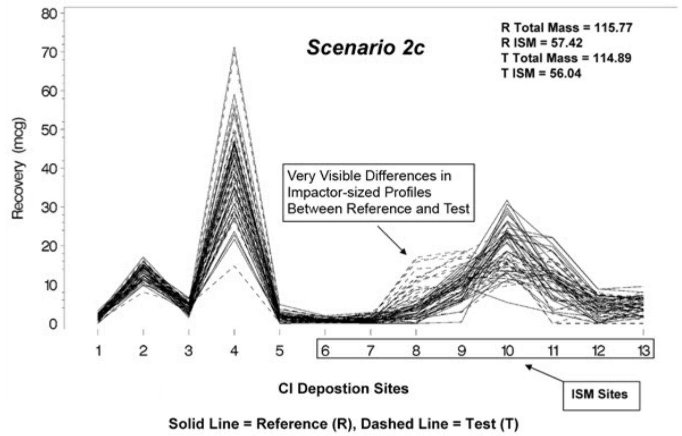

Figure 3. Profile of Scenario 2c, showing deposition and variability on each of 13 cascade impactor sites. ISM indicates impactor-sized mass.

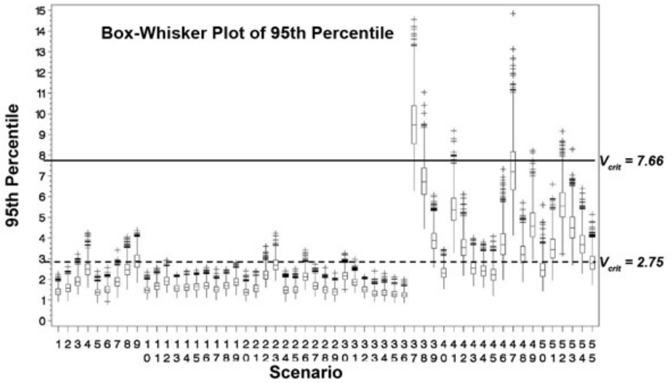

Figure 4. Box-and-whisker plot of the 95th percentile of the chi-square ratio means, for the studied 55 scenarios.

critical value is 7.66, the discriminatory ability of the combined statistical procedure is largely due to the ISM-PBE component, as will be discussed in more detail below.

To explore the possibility of improvement with a different critical value, the WG chose an alternative critical value of 2.75 (dashed line in Figure 4) to reevaluate the discriminating ability of the chi-square ratio test across the same scenarios. The value of 2.75 was selected because it was neither so high as to be above all chi-square ratio distributions for scenarios 1 to 36, nor so low as to be always below distributions for scenarios 37 to 55. In other words, the value of 2.75 seemed to offer the best possibility for an increased discrimination (rather than summarily failing all scenarios or declaring them all equivalent). The reader, however, can examine any other critical value by moving the dashed line to the desired ordinate in Figure 4 and following the same line of argument as presented below.

A summary of all results for the chi-square ratio test with the critical values of 7.66 and 2.75 is presented in Table 1 alongside the results when combined with the ISM-PBE test.

The 2 "Chi-Square Ratio Alone" columns of Table 1 show the effect of changing the critical value on the outcome of the chi-square ratio test. The results reflect the choice of 2.75 as a compromise across the 2 sets of scenarios discussed earlier. For scenarios 1 to 36, the chi-square ratio test still passed most of the T and R data sets as equivalent. For scenarios 37 to 55, the chi-square ratio test with a critical value of 2.75 failed most of the T and R pairs.

Table 1 also shows that the combined test (ie, the "statistical procedure") produced results that were generally in better agreement with the WG than either the chi-square ratio test alone or the ISM-PBE test alone. Differences between the decisions of the WG and the statistical procedure (listed in the 9th and 10th columns of Table 1) were explored in greater detail as follows: agreement to within 50% (or 0.50) of the overall frequency of equivalence declarations between the WG and the statistical procedure was interpreted as adequate for the statistical procedure to make correct decisions. In cases where the difference was more than 0.50, the WG concluded that the statistical tests were not generally capable of making the correct decision and therefore judgment based on reviewer's experience and other information about the products (eg, details of the device, exact formulation, results of other in vitro and in vivo tests)[18] would be necessary to determine APSD equivalence. Of the 55 scenarios, 10 scenarios (18%) fell into this latter category using 7.66 as a critical value (V_{crit}) and 12 (22%) fell into this category when 2.75 was used as V_{crit}. For example, with respect to Figures 1, 2, and 3, WG members generally agreed with the statistical procedure (difference $\leq$ 0.5) for Scenarios 2a and 2b but not 2c, where the statistical procedure showed equivalence 89% of the time, while only 21% of the WG members declared the profiles equivalent (difference = 0.68). Overall, however, the change from a critical value of 7.66 to 2.75 resulted in less consistency between the decisions made using the statistical procedure and those based on the WG's judgment. Even though for some of the scenarios (eg, 12a1), the statistical procedure became more closely aligned with the WG's judgment, statistical decisions on several other scenarios were reversed and became farther removed from the WG's judgment.

Further examination of Table 1 shows that in most cases the outcome of the statistical procedure was controlled by the ISM-PBE test when the chi-square critical value was 7.66 (exceptions were 12a0 and 12a1, which represented very low variability for both R and T profiles, and 13c, which had large differences in the high-deposition sites outside the impactor). This is what would be expected from Figure 4 for the choice of V_{crit} = 7.66. Stated differently, for 52 of the 55 scenarios the statistical assessment of profile equivalence depended on only the total API mass deposited inside the impactor, regardless of the relative amounts deposited on the various stages (or particle size ranges) within the CI. Moreover, in some of the cases in which the ISM-PBE test controlled the outcome of the statistical procedure, the result was not consistent with the WG assessment, suggesting that neither the ISM-PBE test alone nor the overall statistical procedure is adequate in those cases. With the choice of 2.75 as V_{crit}, the situation is the same for scenarios 1 to 36. For scenarios 37 to 55, however, this critical value leads to the chi-square ratio test failing almost all data sets. The chi-square ratio is then the dominating component of the combined procedure.

Table 2 shows a subset of results from the 55 scenarios for which the difference between the WG assessment and the statistical assessment (using 7.66 as the chi-square ratio's critical value) was 50% or greater (10 scenarios). The majority of these scenarios have visibly different R and T profiles for the impactor stages.[10,11] However, the total amounts deposited within the impactor differed by no more than 7%

Table 1. Summary of the Proportions and Proportion Differences With Which Scenarios Were Declared "Equivalent" by the WG and by the Statistical Procedure (Chi-Square Ratio With Critical Value 7.66 and the ISM-PBE Tests)*

				Chi-Square Ratio Alone		Chi-Square Ratio + ISM		Differences: Combined vs WG		Comments for Cases "Difference > 0.50": Combined, 7.66 vs 2.75
ID	Scenario	WG Assessment "E"	ISM-PBE Alone "E"	V_{crit} = 7.66 "E"	V_{crit} = 2.75 "E"	V_{crit} = 7.66 "E/E"	V_{crit} = 2.75 "E/E"	V_{crit} = 7.66	V_{crit} = 2.75	
1	1a	**1.00**	0.70	1.00	1.00	**0.70**	**0.70**	–0.30	–0.30	
2	**1b**	**0.79**	0.22	1.00	1.00	**0.22**	**0.22**	**–0.57**	**–0.57**	No change
3	1c	**0.07**	0.01	1.00	0.99	**0.01**	**0.01**	–0.06	–0.06	
4	1d	**0.00**	0.00	1.00	0.75	**0.00**	**0.00**	0.00	0.00	
5	1aa	**1.00**	0.81	1.00	1.00	**0.81**	**0.81**	–0.19	–0.19	
6	1bb	**0.79**	0.48	1.00	1.00	**0.48**	**0.48**	–0.31	–0.31	
7	1cc	**0.36**	0.01	1.00	0.99	**0.01**	**0.01**	–0.35	–0.35	
8	1dd	**0.00**	0.00	1.00	0.77	**0.00**	**0.00**	0.00	0.00	
9	1ee	**0.00**	0.00	1.00	0.38	**0.00**	**0.00**	0.00	0.00	
10	2a	**0.79**	0.89	1.00	1.00	**0.89**	**0.89**	0.10	0.10	
11	2b	**0.50**	0.92	1.00	1.00	**0.92**	**0.92**	0.42	0.42	
12	**2c**	**0.21**	0.89	1.00	1.00	**0.89**	**0.89**	**0.68**	**0.68**	No change
13	2aa1	**0.71**	0.88	1.00	1.00	**0.88**	**0.88**	0.17	0.17	
14	2bb1	**0.64**	0.80	1.00	1.00	**0.80**	**0.80**	0.16	0.16	
15	2cc1	**0.50**	0.74	1.00	1.00	**0.74**	**0.74**	0.24	0.24	
16	**2dd1**	**0.29**	0.89	1.00	1.00	**0.89**	**0.89**	**0.60**	**0.60**	No change
17	2aa2	**0.64**	0.90	1.00	1.00	**0.90**	**0.90**	0.26	0.26	
18	**2bb2**	**0.29**	0.89	1.00	1.00	**0.89**	**0.89**	**0.60**	**0.60**	No change
19	**2cc2**	**0.14**	0.94	1.00	1.00	**0.94**	**0.94**	**0.80**	**0.80**	No change
20	4a	**1.00**	0.88	1.00	1.00	**0.88**	**0.88**	–0.12	–0.12	
21	4b	**1.00**	0.89	1.00	1.00	**0.89**	**0.89**	–0.11	–0.11	
22	4c	**0.21**	0.68	1.00	0.96	**0.68**	**0.66**	0.47	0.45	
23	4d	**0.14**	0.45	1.00	0.57	**0.45**	**0.29**	0.31	0.15	
24	5a	**0.93**	0.89	1.00	1.00	**0.89**	**0.89**	–0.04	–0.04	
25	5b	**0.86**	0.86	1.00	1.00	**0.86**	**0.86**	0.00	0.00	
26	5c	**0.29**	0.50	1.00	0.97	**0.50**	**0.49**	0.21	0.20	
27	**7a**	**0.29**	0.89	1.00	1.00	**0.89**	**0.89**	**0.60**	**0.60**	No change
28	7b	**0.50**	0.95	1.00	1.00	**0.95**	**0.95**	0.45	0.45	
29	7c	**0.93**	0.92	1.00	1.00	**0.92**	**0.92**	–0.01	–0.01	
30	**10a**	**0.14**	1.00	1.00	0.97	**1.00**	**0.97**	**0.86**	**0.83**	No change
31	**10b**	**0.29**	1.00	1.00	1.00	**1.00**	**1.00**	**0.71**	**0.71**	No change
32	10c	**0.50**	1.00	1.00	1.00	**1.00**	**1.00**	0.50	0.50	
33	10d	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	0.00	0.00	
34	11a	**0.64**	0.78	1.00	1.00	**0.78**	**0.78**	0.14	0.14	
35	11b	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	0.00	0.00	
36	11c	**1.00**	1.00	1.00	1.00	**1.00**	**1.00**	0.00	0.00	
37	12a0	**0.07**	1.00	0.07	0.00	**0.07**	**0.00**	0.00	–0.07	
38	**12a1**	**0.14**	1.00	0.82	0.00	**0.82**	**0.00**	**0.68**	–0.14	2.75 "agrees"
39	**12a2**	**0.86**	1.00	1.00	0.01	**1.00**	**0.01**	0.14	**–0.85**	7.66 "agrees"
40	12a3	**1.00**	1.00	1.00	0.88	**1.00**	**0.88**	0.00	–0.12	
41	12b0	**0.29**	0.03	0.99	0.00	**0.03**	**0.00**	–0.26	–0.29	
42	**12b1**	**0.86**	0.93	1.00	0.06	**0.93**	**0.06**	0.07	**–0.80**	7.66 "agrees"
43	12b2	**0.86**	1.00	1.00	0.70	**1.00**	**0.70**	0.14	–0.16	
44	12b3	**1.00**	1.00	1.00	0.83	**1.00**	**0.83**	0.00	–0.17	
45	**13a**	**1.00**	0.40	1.00	0.90	**0.40**	**0.37**	**–0.60**	**–0.63**	No change
46	**13b**	**0.57**	0.38	1.00	0.08	**0.38**	**0.03**	–0.19	**–0.54**	7.66 "agrees"
47	13c	**0.36**	0.41	0.63	0.00	**0.26**	**0.00**	–0.10	–0.36	

Table 1. Cont.

				Chi-Square Ratio Alone		Chi-Square Ratio + ISM		Differences: Combined vs WG		
ID	Scenario	WG Assessment "E"	ISM-PBE Alone "E"	V_{crit} = 7.66 "E"	V_{crit} = 2.75 "E"	V_{crit} = 7.66 "E/E"	V_{crit} = 2.75 "E/E"	V_{crit} = 7.66	V_{crit} = 2.75	Comments for Cases "Difference > 0.50": Combined, 7.66 vs 2.75
48	13d	**0.21**	0.00	1.00	0.19	**0.00**	**0.00**	–0.21	–0.21	
49	13e	**0.07**	0.00	1.00	0.00	**0.00**	**0.00**	–0.07	–0.07	
50	13f	**0.93**	0.84	1.00	0.74	**0.84**	**0.65**	–0.09	–0.28	
51	13g	**0.50**	0.83	1.00	0.11	**0.83**	**0.10**	0.33	–0.40	
52	14a1	**0.00**	0.00	0.98	0.00	**0.00**	**0.00**	0.00	0.00	
53	14a2	**0.07**	0.00	1.00	0.00	**0.00**	**0.00**	–0.07	–0.07	
54	14a3	**0.43**	0.03	1.00	0.03	**0.03**	**0.00**	–0.40	–0.43	
55	**14a4**	**0.71**	0.37	1.00	0.49	**0.37**	**0.22**	–0.34	–0.49	7.66 "closer"

*The bolded values are to help guide the reader to the appropriate results for the combined test and for those instances where the combined test results are "importantly" different from the WG assessment. WG indicates the Aerodynamic Particle Size Distribution Profile Comparisons Working Group; ISM, impactor-sized mass; PBE, population bioequivalence; E, equivalent.

between T and R. The WG assessment for 8 of the 10 scenarios invariably indicated that they should not be considered equivalent; however, the ISM-PBE assessment, based solely on the total deposition in the impactor, uniformly resulted in a decision of equivalence for these 8, consistent with the way this test is designed. An example of this is scenario 2c (Figure 3), where it can be seen that although the ISM values for R and T differ by less than 3%, the profiles are visibly quite different. It appeared that for these 8 scenarios the WG members consistently judged the relative differences between R and T to be important enough to be considered inequivalent, while the ISM-PBE test was unable to detect profile differences since it compared only total amounts deposited inside the impactor.

For Scenario 2c (Figure 3), even though the WG considered the products to have important differences in deposition profile within the CI, the chi-square ratio test indicated equivalence between profiles because of high and nearly equal deposition on site 4 (outside the impactor), despite an almost 3-fold difference in mean deposition on site 8 and an almost 30% lower mean deposition on site 10 (both of these sites are within the impactor-sized portion of the profile).

In the remaining 2 of 10 scenarios, namely 1b and 13a, the majority of the WG judged the profiles to be equivalent, while the statistical tests consistently indicated inequivalence, because of the ISM-PBE results. The R vs T differences in ISM for these scenarios were 21% of the reference ISM for 1b, and 12% for 13a (as calculated from the last 2 columns of Table 2). These differences were not considered to be important by the WG members but were large enough to be consistently detected as inequivalent by the ISM-PBE test.

When the above exercise was repeated using a critical value of 2.75 for the chi-square ratio test, 12 instances of differences greater than 0.50 between the statistical tests and the WG judgment were found. Nine of those cases were identical to, or at most no more than 0.03 different from, the results seen when 7.66 was used as a critical value. Three additional instances (12a2, 12b1, 13b) disagree with the WG judgment using the 2.75 critical value, whereas the 7.66 critical value results do not disagree, and in only one instance did the 2.75 critical value result agree with the WG judgment when the 7.66 result did not (12a1). Overall, the test with V_{crit} = 2.75 performed slightly worse than the test with V_{crit} = 7.66, relative to the agreement with the WG judgment.

Generally speaking, particles inside the impactor, especially those with smaller aerodynamic size, may affect the efficacy of delivered inhaled drugs. Particles deposited outside the impactor roughly represent the nonrespirable portion of the dose and therefore may play some role in the safety of a product. The exact definitions of respirable and nonrespirable particles, and their correlation to safety and efficacy, remain a subject of debate. Nevertheless, based on this general understanding, the value of the chi-square ratio test could lie in its ability to detect differences in the high-deposition sites, which for most orally inhaled drugs occur outside the impactor and may be tentatively linked to safety considerations.

The value of using the ISM-PBE test as proposed here lies in its ability to detect differences in the total deposition inside the impactor. A potential disadvantage of introducing this additional ISM-PBE test is that the T product must satisfy 2 tests (chi-square ratio and ISM-PBE), which may increase the overall rate of incorrectly declaring 2 profiles inequivalent unless special statistical adjustments are made in each

Table 2. Scenarios for Which the Difference Between WG Assessment and Statistical Assessment Is Greater Than 0.50 (50%) Using 7.66 as the Chi-Square Critical Value*

		Proportion of Assessments					ISM Deposition Mean (%RSD)	
ID	Scenario	WG Assessment "E"	Chi-Square V_{crit} = 7.66 "E"	ISM-PBE "E"	Combined Chi-Sq/ISM-PBE "E × E"	Difference WG vs Method	Reference	Test
2	1b	**0.79**	1.00	0.22	**0.22**	**0.57**	59 (15)	47 (15)
12	2c	**0.21**	1.00	0.89	**0.89**	**–0.68**	57 (17)	56 (14)
16	2dd1	**0.29**	1.00	0.89	**0.89**	**–0.60**	56 (17)	56 (14)
18	2bb2	**0.29**	1.00	0.89	**0.89**	**–0.60**	54 (17)	53 (15)
19	2cc2	**0.14**	1.00	0.94	**0.94**	**–0.80**	57 (15)	55 (12)
27	7a	**0.29**	1.00	0.89	**0.89**	**–0.60**	64 (13)	67 (12)
30	10a	**0.14**	1.00	1.00	**1.00**	**–0.86**	86 (7)	88 (5)
31	10b	**0.29**	1.00	1.00	**1.00**	**–0.71**	87 (7)	87 (5)
38	12a1	**0.14**	0.82	1.00	**0.82**	**–0.68**	72 (8)	67 (4)
45	13a	**1.00**	1.00	0.40	**0.40**	**0.60**	25 (11)	22 (10)

*The bolded values are to help guide the reader to the comparisons between the WG assessment and the outcome of the combined statistical procedure. WG indicates the Aerodynamic Particle Size Distribution Profile Comparisons Working Group; E, equivalent; ISM, impactor-sized mass; RSD, relative standard deviation; PBE, population bioequivalence.

test component to keep the overall probability of that error at a predetermined low value.

In this study, the WG compared the statistical outcomes of the chi-square ratio test with and without the addition of the ISM-PBE test with the judgment of expert WG members. This approach was adopted because currently there are no clinical data that would allow estimation of some "permitted difference" in APSD (considering both mean and variability) from the R product that would result in equivalent clinical performance. A limited number of studies have been published about the relationship between mass median aerodynamic diameter and deposition patterns in the lung, but these have involved small numbers of subjects, showed limitations in methodology, and occasionally led to discrepant conclusions.[19-22] Thus, currently there is no predictive relationship between changes in APSD and clinical results. Without a clinically defined "permitted difference" or a "gold standard" to which the outcome of the statistical procedure could be compared, the experienced judgment of the WG members using "target profiles" described previously[7] was the best available option. While the WG recognizes the limitations in its estimation of equivalence, the general consistency of responses among the members, each of whom conducted the evaluation independently using his or her own methods, lends further credence to the WG conclusions.

The WG did not address issues of in vivo equivalence, although the WG acknowledges that a complete bioequivalence package that includes in vivo information may assist in evaluating the importance of the in vitro APSD results. Similarly, this study did not address other in vitro tests that might be recommended by current regulatory guidance documents. The WG arbitrarily set the threshold for "agreement" between statistical evaluation and the WG assessment as being a difference in frequency ≤ 0.50. Obviously, if this "condition for agreement" had been set differently, a different number of cases would have been seen where the statistical procedure produced results that were inconsistent with the WG judgment. Even so, the general conclusions in this PQRI report would remain unchanged, as can be seen from a careful examination of the results presented in Table 1.

From these comparisons, neither the statistical tests alone nor their combination seemed able to discriminate important shifts in the potentially "respirable" portion of the profile when the overall API mass deposited inside the impactor remained the same (eg, the highest-deposition sites do not shift from outside to inside the impactor). For the chi-square ratio test, this is in part related to the choice of the critical value, V_{crit}. The FDA originally proposed 7.66 based on the albuterol MDI data it considered. As Figure 4 confirms, the value of 7.66 was not a good choice for the scenarios considered here. However, an alternate choice that was discriminatory for the profiles at the right side of Figure 4 would still be too large for the scenarios to the left. Similarly, if the critical value were chosen for the scenarios at the left, it would be too low for the other scenarios, mostly declaring all T-R pairs to not be equivalent. For instance, the considered alternative critical value of 2.75, which was chosen so as not to be too low for the right-hand scenarios nor too high for the left-hand scenarios, showed inferior performance compared with the critical value of 7.66. As can be seen from Figure 4, any selected critical value (2.75, 7.66, or any other) would cut across statistical distributions for profiles judged both E and I, and so there is no single critical value that can discriminate differences of interest with this statistical procedure.

Another possible way to improve performance of the chi-square ratio test could be to apply that test to only the impactor-sized portion of the profile. However, this approach was rejected because it cannot improve the situation because of the decreased number of sites, which decreases the performance (stability, consistency, and discriminating power) of the chi-square ratio test, as was shown in earlier PQRI studies.[23]

Similarly, fine-tuning of the ISM-PBE method by changing its acceptance criteria may only change the sensitivity of the ISM-PBE test to the changes in the overall API mass sized by the impactor; it will not affect the inability of the test to detect shifts in the profile.

In summary, adjusting the acceptance criteria (for the PBE test) or the critical value (for the chi-square ratio test), or applying the chi-square ratio test to only the impactor-sized portion of the profile, cannot address the identified deficiencies, because of the nature and design of these tests.

Important Note on Using Appropriate Data for APSD Comparisons

In the present evaluation, it was assumed that R and T data were comparable, that is, obtained under the same conditions and in such a way that non-product-related variability was minimized and statistically balanced between R and T, such that any observed differences were mostly due to true differences between the products. Because CI measurements can be notoriously dependent on factors not related to the product (eg, analyst, impactor, laboratory, environmental conditions, and method[24]), it is important to ensure in practice that the 2 sets of data to be compared are indeed comparable. Therefore, regardless of the statistical or other test used for APSD comparison, the APSD profiles should be obtained from CI testing in cohorts, such that variabilities and biases introduced into APSD results by different analysts and impactors are balanced. (This approach is similar to the crossover designs used in bioequivalence studies to address subject-to-subject variability.) Without such balancing, comparison of APSD profiles obtained from CI testing is generally inappropriate and may be misleading, regardless of the statistical method to be employed.

The requirement of simultaneous testing in cohorts also makes the discussed tests inappropriate for quality control purposes. Current quality control tests typically compare CI deposition data from groupings of stages (which are specific to a particular product) to preset numerical limits (also specific to a product). By contrast, the equivalence tests considered here require head-to-head comparisons between 2 products tested in cohorts (ie, both products tested on the same day by the same analyst using the same impactors, etc) through a balanced design, which is not possible in a quality control context of batch release or in stability testing to determine whether a batch has changed over time.

CONCLUSIONS

The PQRI WG evaluated the capability of the combined statistical procedure (a chi-square ratio test plus an ISM-PBE test) for detecting differences in APSD CI profiles between T and R that may be considered important for the establishment of in vitro bioequivalence. It was found that the studied statistical procedure may lead to conclusions that are at odds with expert judgments regarding the equivalence of T and R profiles. In situations where the variabilities of T and R profiles are commensurate, the statistical procedure may be overly sensitive to the differences in mean depositions that were not deemed important by the WG (as discussed, eg, for 1b and 13a); in other situations, the statistical tests in their current construction were unable to detect profile shifts. The change of the acceptance criteria (for the PBE test) or the selection of a different critical value (for the chi-square ratio test) could not resolve these deficiencies; therefore, no recommendations are made by the WG for APSD profile comparisons.

ACKNOWLEDGMENTS

The WG appreciates the opportunity provided by PQRI for scientists from the government, industry, academia, and the US Pharmacopeia to address a complex scientific regulatory issue through a collaborative scientific process. The authors also thank the editor and anonymous reviewers for helpful comments.

REFERENCES

1. FDA/CDER. Draft guidance for industry: bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action. Food and Drug Administration Web site. 1999; Available at: http://www.fda.gov/cder/guidance/2070DFT.pdf. Accessed March 28, 2007.

2. Tsong Y. Statistical comparison of particle size distribution profiles. 2004; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/DC01-475116-v2-Yi_Tsong_Statistical_Archive_PQRI_Profile_Comparisons.DOC. Accessed June 22, 2007.

3. Tsong Y. Profile analysis of cascade impactor data: proposed FDA approach. Presentation to OINDP Subcommittee of the Advisory Committee for Pharmaceutical Science. Rockville, MD: US Pharmacopeia; April 26, 2000. Available at: http://www.fda.gov/ohrms/dockets/ac/00/slides/3609s1e.ppt. Accessed October 19, 2006.

4. United States Pharmacopeial Convention. Chapter 601. Aerosols, metered-dose inhalers, and dry powder inhalers. In: *USP30-NF25*. Rockville, MD: USP; 2007:220–240.

5. DeLuca PP, Lyapustina S. Product Quality Research Institute reports. *AAPS PharmSciTech [serial online]*. 2007;8:article 6.

6. Adams WP, Christopher D, Lee DS, et al. Product Quality Research Institute evaluation of cascade impactor profiles of pharmaceutical

aerosols, Part 1: background for a statistical method. *AAPS PharmSciTech [serial online].* 2007;8:article 4.

7. Christopher D, Adams WP, Lee DS, et al. Product Quality Research Institute evaluation of cascade impactor profiles of pharmaceutical aerosols, Part 2: evaluation of a method for determining equivalence. *AAPS PharmSciTech [serial online].* 2007;8:article 5.

8. PQRI Profile Comparisons Working Group. Systematic study of 38 scenarios. 2005; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/38%20Systematic%20Scenarios.pdf. Accessed November 1, 2006.

9. FDA. CDER. Statistical information from the June 1999 Draft Guidance and Statistical Information for *In Vitro* Bioequivalence Data Posted on August 18, 1999. Available at: http://www.fda.gov/cder/guidance/5383stats.pdf. Accessed January 14, 2007.

10. PQRI Profile Comparisons Working Group. Realistic scenarios 1-33 of 55. 2005; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/Realistic%20Scenarios%201-33%20of%2055.pdf. Accessed October 19, 2006.

11. PQRI Profile Comparisons Working Group. Realistic scenarios 34-55 of 55. 2005; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/Realistic%20Scenarios%2034-55%20of%2055.pdf. Accessed October 19, 2006.

12. PQRI APSD Profile Comparisons Working Group's assessment. 2006; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/PQRI_APSD_Profile_Comparisons_Working_Groups_Assessment.XLS. Accessed June 22, 2007.

13. SAS Institute Inc. *SAS System for Windows, Release 9.1 (TS1M3).* Cary, NC: SAS Institute Inc; 2003.

14. PQRI APSD Profile Comparisons Working Group. SAS code implementing the chi-square ratio method. 2006; Available at: http://www.pqri.org/structure/psdp.asp. Accessed May 29, 2007.

15. PQRI APSD Profile Comparisons Working Group. Description of the PBE approach. 2007; Available at: http://www.pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/DC01-526048-v1-PBE_Description_for_PQRI_Prof_Comp.pdf. Accessed May 29, 2007.

16. PQRI APSD Profile Comparisons Working Group. Minutes of the teleconference on 13 July 2005:3-8. Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/071305min.pdf. Accessed March 28, 2007.

17. PQRI APSD Profile Comparisons Working Group. Qualitative descriptions with thumbnail plots of the 55 realistic scenarios. 2006; Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/DC01-525635-v1-Thumbnail_plots_prof_comp_scenarios.pdf. Accessed May 17, 2007.

18. FDA/CDER. Second draft guidance for industry: bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action. *Food and Drug Administration Web site.* 2003; Available at: http://www.fda.gov/cder/guidance/5383DFT.pdf. Accessed June 22, 2007.

19. Zanen P, Go LT, Lammers J-WJ. The optimal particle size for β-adrenergic aerosols in mild asthmatics. *Int J Pharm.* 1994;107 211–217.

20. Zanen P, Go LT, Lammers J-WJ. The optimal particle size for parasympathicolytic aerosols in mild asthmatics. *Int J Pharm.* 1995;114:111–115.

21. Zanen P, Go LT, Lammers J-WJ. Optimal particle size for beta 2 agonist and anticholinergic aerosols in patients with severe airflow obstruction. *Thorax.* 1996;51:977–980.

22. Usmani OS, Biddiscombe MF, Barnes PJ. Regional lung deposition and bronchodilator response as a function of β_2-agonist particle size. *Am J Respir Crit Care Med.* 2005;172:1497–1504.

23. Lee DS. Searching for the holy grail of a single PSD profile comparator. Respiratory Drug Delivery Conference IX; April 25-29, 2004; Palm Desert, CA. Available at: http://pqri.org/commworking/minutes/pdfs/dptc/psdpcwg/Addl/DC01-504108-v1-Identical Profiles_and_Stabilty_Slides.PPT. Accessed October 19, 2006.

24. Christopher D, Curry P, Doub B, et al. Considerations for the development and practice of cascade impaction testing, including a mass balance failure investigation tree. *J Aerosol Med.* 2003; 16:235–247.