AAPS PharmSciTech (2019) 20: 249
DOI: 10.1208/s12249-019-1443-7

Research Article

Theme: Paul Myrdal Memorial Issue - Pharmaceutical Formulation and Aerosol Sciences
Guest Editors: Philip J. Kuehl and Stephen W. Stein

Cascade Impactor Equivalence Testing: Comparison of the Performance of the Modified Chi-Square Ratio Statistic (mCSRS) with the Original CSRS and EMA's Average Bioequivalence Approach

Abhinav Kurumaddali,[1] David Christopher,[2] Dennis Sandell,[3] Helen Strickland,[4] Beth Morgan,[4] Juergen Bulitta,[1] Christopher Wiggenhorn,[5] Stephen Stein,[5] Svetlana Lyapustina,[6] and Günther Hochhaus[1,7]

Received 17 January 2019; accepted 3 June 2019; published online 8 July 2019

Abstract. The performances of three statistical approaches for assessing *in vitro* equivalence was evaluated with a set of 55 scenarios of realistic test (T) and reference (R) cascade impactor (CI) profiles (originally employed by the Product Quality Research Institute to evaluate the chi-square ratio statistic: CSRS) by comparing the outcomes against experts' opinion (surrogate for the truth). The three methods were (A) a stepwise aerodynamic particle size distribution (APSD) equivalence test integrating population bioequivalence (PBE) testing of impactor-sized mass (ISM) with the CSRS (PBE-CSRS approach), previously suggested by the USFDA; (B) the combination of PBE testing of single actuation content and ISM with the newly suggested modified CSRS (PBE-mCSRS approach), a method employing reference variance scaling; and (C) EMA's average bioequivalence (ABE approach). Based on Monte-Carlo simulations, both PBE-CSRS and ABE approaches resulted in high misclassification rates, the former with highest false-pass rate and the latter with highest false-fail rate at both ≥ 50% and ≥ 80% classification threshold values (the % of simulations or experts necessary to judge a given scenario as equivalent). Based on DeLong's tests, the PBE-mCSRS approach showed significantly better overall agreement with experts' opinion compared to the other approaches. Comparison of CSRS with mCSRS (both without PBE) suggested that the more discriminatory characteristics of the mCSRS method is based on the integration of variance scaling into the mCSRS method. Contrary to the ABE approach, the application of PBE-mCSRS approach for assessing APSD profiles of three dry powder inhaler (DPI) formulations supported the pharmacokinetic bioequivalence assessment of these formulations.

KEY WORDS: aerodynamic particle size distribution; bioequivalence; cascade impactor; modified chi-square ratio statistic (mCSRS); receiver operating characteristic curves (ROC).

INTRODUCTION

Assessing the bioequivalence of traditional oral dosage forms does not generally represent a challenge, as established guidelines recommend the assessment of the systemic drug exposure (AUC and C_{max}) between the test (T) and reference (R) formulations. In contrast, it is quite challenging in the case of locally acting drug products, such as inhalation drugs, as the active pharmacological ingredient (API) is directly delivered to the site of action; thus, blood plasma concentrations are judged by many stakeholders to be less relevant for bioequivalence decisions (1). It is well established that the aerodynamic particle size distribution (APSD) of inhaled formulations plays a crucial role in determining the pulmonary deposited dose and regional lung deposition pattern (2–5). Hence, the international regulatory agencies such as the Food and Drug Administration (FDA, USA), Health Canada (HC, Canada), European Medicines Agency (EMA, European Union), Therapeutic Goods Administration (TGA, Australia) *etc.* recommend *in vitro* equivalence testing using cascade impactors such as the Andersen cascade impactor (ACI) or the next-generation impactor (NGI, see Fig. 1) as one of the key steps in the approval of "generic" (or

Guest Editors: Philip J. Kuehl and Stephen W. Stein

[1] Department of Pharmaceutics, College of Pharmacy, University of Florida, Box 100494, Gainesville, Florida, 32610, USA.
[2] Biostatistics, Merck & Co., West Point, PA, USA.
[3] S5 Consulting, Blentarp, Sweden.
[4] Statistics, GlaxoSmithKline, Raleigh, NC, USA.
[5] R&D, 3M Drug Delivery Systems Division, 3M Company, St. Paul, MN, USA.
[6] Pharmaceutical Practice Group, Drinker Biddle and Reath LLP, Washington, DC, USA.
[7] To whom correspondence should be addressed. (e–mail: Hochhaus@cop.ufl.edu)

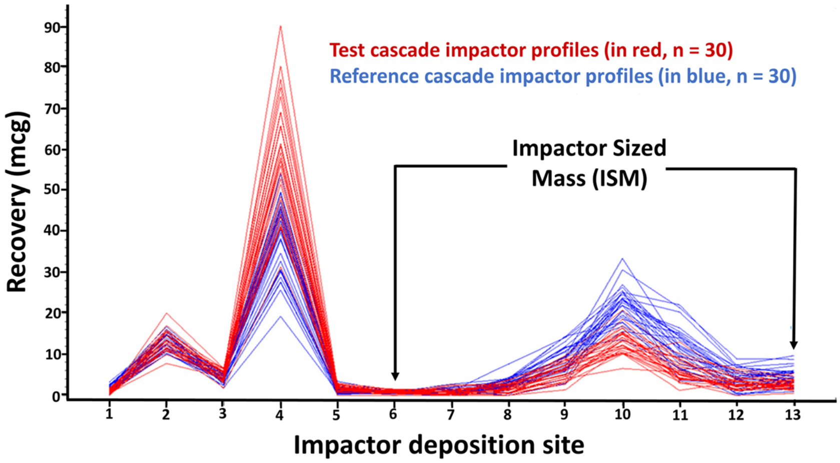

Fig. 1. Representation of Andersen Cascade Impactor (ACI) profiles obtained from typical test (T) and reference (R) inhalation products of sample size 30 each

"follow-on", or "second-entry") inhalation drug products (1). On a theoretical level, the cascade impactor profile analysis of test (T) and reference (R) products should consider the shape of the cascade impactor profile (Fig. 1) as well as absolute cumulative dose entering the impactor (impactor-sized mass, or ISM). In addition, the single actuation content is of relevance to ensure that the total dose leaving the dosage form is equivalent, as it is also relevant for the orally available drug. Methods to statistically evaluate equivalence of multivariate vectors with correlated elements such as T and R cascade impactor profiles are complex, and the statistical methods used for evaluating the shape of the cascade impactor profiles are not specified within FDA or Health Canada official guidance (1,6–8).

In June 1999, FDA issued a guidance entitled "Draft Guidance for Industry: Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action", recommending the use of a chi-square ratio statistic (CSRS), a univariate cumulative assessment metric for evaluating the equivalence in shape of T and R cascade impactor profiles (based on relative stage depositions) (9,10). In the same guidance, FDA also proposed the use of Population Bioequivalence (PBE) criterion for comparing the single actuation content and impactor-sized mass of T and R formulations (9–12). The performance of the combination of PBE applied to ISM and the CSRS test applied to the full cascade impactor (CI) profile (PBE-CSRS approach) was evaluated by a Product Quality Research Institute Working Group (PQRI WG) focused on APSD comparisons, using a set of 55 PQRI-developed scenarios of realistic T and R CI profiles. Since there was no definitive basis established by industry or regulatory agencies for determining APSD equivalence, the PQRI WG compared the outcomes of the CSRS approach for the 55 scenarios against an independent assessment of experts' opinion. The working group concluded that the CSRS approach could not discriminate consistently between what experts judged to be equivalent and non-equivalent cascade impactor profiles (2,10). More specifically, the working group found that the use of a fixed critical value within the CSRS test (defined in the FDA CSRS approach) for making pass/fail decisions and the instability of the CSRS when applied to a reduced number of deposition sites compromises the discriminatory ability, and therefore the utility of this approach for making relevant equivalence decisions (2,10).

The European Medicines Agency (EMA) recommends the method of Average Bioequivalence (ABE) for testing the *in vitro* bioequivalence of T and R cascade impactor profiles in its 2009 guidance (8,13). The ABE statistical procedure can be applied to deposition data of individual impactor stages or justified groups of stages. Evidence of equivalence is based on confidence intervals of T/R ratios within a window of ± 15% (8,13). It is noteworthy that given the stringent acceptance criteria set in the EMA guidance and the multiple test comparisons to be performed for T-R profiles (one test per stage or group), the R product tested against itself generally fails to meet the bioequivalence criteria (13). Hence, it was of interest to compare the outcome of the EMA method with those of alternative tests and to determine the effect of less stringent acceptance criteria on the outcomes of this approach.

To overcome the limitations of the CSRS for relevant decision-making identified in the PQRI WG report, University of Florida (UoF) in collaboration with FDA developed a modified version of the CSRS (mCSRS) for comparing the T and R cascade impactor profiles. Unlike CSRS, the mCSRS was shown to be stable even when applied to a reduced number of cascade impactor stages that are more relevant to lung deposition (2). Most importantly, the critical value is scaled according to the variability of the reference product (quantified by a cumulative metric called reference variance scaling, RVS) following the same idea that was the basis for extending the ABE approach into the PBE test (4).

For this article, the three statistical approaches, ABE, PBE-CSRS and PBE-mCSRS were applied to the same sample of CI profiles, namely the 55 PQRI scenarios, originally used by the PQRI working group. The results of all three statistical approaches were compared against the experts' opinion (surrogate for the truth) for all the 55 PQRI scenarios using quadrant (scatter) plots. Each statistical approach was evaluated for its accuracy/validity (measured by true pass rate and true fail rate which were defined in terms of agreement with experts' opinion). The ability of each statistical approach to discriminate between equivalent and non-equivalent T and R cascade impactor profiles and its agreement with the experts' opinion was quantified by means of receiver operating characteristic (ROC) curves (14–16). To gain more insight into the applicability of the ABE approach, the effect of relaxing the EMA acceptance criteria on the outcome was studied by either widening the ± 15% acceptance limit or reducing the confidence level. Finally, the behavior of the three statistical tests in relation to the variability of the R formulation and the potential *in vivo* relevance of the statistical tests were assessed.

METHODS

Overall Strategy

The predictive performance of the three statistical approaches in evaluating the equivalence of cascade impactor profiles was tested by analyzing the 55 PQRI scenarios described below, of T and R cascade impactor profiles and comparing the results with evaluations of subject matter experts presented in the PQRI WG report.

1. Average bioequivalence approach (ABE): Assessment of individual stages or groups of stages using the ABE approach, a standard statistical equivalence test, as described by EMA in its guidance (8).
2. Chi-square ratio statistic approach (PBE-CSRS): This method tests first the equivalence of T and R products in impactor-sized mass through the population equivalence (PBE) approach followed by evaluating the equivalence in the shape of the cascade impactor profiles (based on relative stage depositions) by the chi-square ratio test (10).
3. Modified chi-square ratio statistic approach (PBE-mCSRS): This method tests first the equivalence of T and R products in single actuation content and impactor-sized mass by population equivalence (PBE) approaches followed by evaluating the equivalence in the shape of the cascade impactor profiles (based on relative stage depositions) by the modified chi-square ratio test (4).

In addition, the potential *in vivo* relevance of results generated by the statistical evaluation of cascade impactor (CI) profiles was explored by comparing the ABE, PBE and mCSRS-based CI profile equivalence outcomes of three experimental dry powder inhaler (DPI) formulations of fluticasone propionate (FP; A, B, and C) to their corresponding PK bioequivalence (PK BE) outcomes using data previously presented (17). Details of the DPI formulation development, *in vitro* assessment, and PK studies will be published elsewhere, while statistical outcomes are also reported here, for completeness of the present discussion.

Description of the PQRI Scenarios

To compare the above methods, this study used 55 realistic scenarios of 30 T and 30 R simulated cascade impactor profiles previously published by the Product Quality Research Institute Working Group (PQRI WG) and results of the evaluation of these by subject matter experts, who judged these profiles as equivalent or non-equivalent (10).

The 55 scenarios were developed by the PQRI WG based on statistical variance component analysis of blinded sets of cascade impactor data from actual products. This variance component analysis produced for each set of data (*e.g.*, albuterol MDI) the mean and variance for each CI deposition site, plus a variance-covariance matrix which characterized the interrelationship among the deposition sites. Using these values, simulated datasets were produced that closely mimicked all the important characteristics of the APSD profiles from an actual product. By changing the values for deposition site means and/or variance (but maintaining the interrelationship among deposition sites), different scenarios were simulated that ranged from the observed profiles to profiles with various combinations of differences between T and R in mean deposition and variability. In brief, the 55 PQRI scenarios were comprised of three main classes:

Class I: It includes scenario nos. 1–44, each scenario representing 30 T and 30 R cascade impactor profiles obtained using an Andersen Cascade Impactor (ACI) containing 13 deposition sites (deposition sites 6 through 13 representing impactor-sized mass, ISM, sum of amount deposited on ISM deposition sites, the deposition sites with specified upper cut-off size) operated at a flow rate of 28.3 L/min.

Class II: It includes scenario nos. 45–51, each scenario representing 30 T and 30 R cascade impactor profiles obtained using an Andersen Cascade Impactor (ACI) containing 11 deposition sites (deposition sites 4 through 11 representing impactor-sized mass, ISM) operated at a flow rate of 60 L/min.

Class III: It includes scenario nos. 52–55, each scenario representing 30 T and 30 R cascade impactor profiles obtained using the next-generation impactor (NGI) containing 10 deposition sites (deposition sites 3 through 10 representing impactor-sized mass, ISM) operated at a flow rate of 60 L/min. These profiles were both directly assessed by subject matter experts and analyzed by each of the three statistical approaches.

Evaluation of the PQRI Scenarios by Subject Matter Experts

This study builds upon the previously published PQRI report on the 55 scenarios of cascade impactor profiles and their visual (not statistical) evaluation by subject matter experts (who represented experienced product developers, bioequivalence researchers and regulatory affairs professionals from industry, academia, pharmacopeia, and FDA) (10). As described in a previous PQRI publication, for each scenario, 14 independent evaluations were received from subject matter experts, who visually reviewed pairs of CI profiles and adopted a "regulatory perspective" for concluding equivalence or not based on the assumption that certain changes in CI profiles could be consistently translated into *in vivo* pulmonary deposition changes, which in turn might affect the clinical outcomes (10). Reasons for having to use this subjective way of assessing the profiles and consequently the statistical test to be evaluated were given in the same publication together with more information on the subject-matter expertise of the experts involved.

For the purpose of comparison, an overall pass was assigned for a given scenario when the percent of PQRI WG members (experts) that classified T and R profile of a given scenario as equivalent exceeded the specified threshold value (for example ≥ 50% and ≥ 80%). The experts' opinion (at ≥ 50% and ≥ 80% threshold values) was defined as a surrogate for "the truth" when evaluating the performance of the three statistical approaches (ABE, PBE-CSRS and PBE-mCSRS approaches).

Application of the Three Statistical Approaches to the PQRI Scenarios

To evaluate the performance of the statistical approaches, for a given scenario of the 55 studied, 1000 sets, each consisting of 30 T and 30 R cascade impactor profiles, were generated by Monte Carlo simulations as described in the previous publications (2–4). Briefly, information on the population means and standard deviations of drug amounts on all deposition sites along with the population inter-site correlations between all the deposition sites of the cascade impactor profiles was used to generate 1000 random samples of 30 T and 30 R cascade impactor profiles under the assumption of multivariate normal distribution of the drug amounts on all deposition sites in SAS software. These 1000 replicates of a given scenario were subjected to the statistical tests. The three statistical approaches applied to each of the 1000 datasets in all the 55 PQRI scenarios are described below:

1. Average bioequivalence approach (ABE)

The ABE approach was applied to each of the 1000 datasets within all the 55 PQRI scenarios as recommended in the 2009 EMA guidance using the statistical software R (version 3.4.4). Briefly, for each dataset of 30 T and 30 R cascade impactor profiles, all of the deposition sites in a cascade impactor profile were divided into four groups (13):

- Group 1: deposition sites with no defined upper cut-off diameter (deposition sites 1–4, 1–3, and 1–2 for PQRI scenarios 1–44, 45–51, and 52–55 respectively)
- Group 2: deposition sites representing coarse mass (deposition sites 5–7, 4–6, and 3–4 for PQRI scenarios 1–44, 45–51, and 52–55 respectively)
- Group 3: deposition sites representing fine particle mass (deposition sites 8–10, 7–9, and 5–7 for PQRI scenarios 1–44, 45–51, and 52–55 respectively)
- Group 4: deposition sites representing extra-fine particle mass (deposition sites 11–13, 10–11, and 8–10 for PQRI scenarios 1–44, 45–51, and 52–55 respectively)

The T/R ratio 90% confidence intervals (equations shown below) for each group of deposition sites were constructed by the geometric mean ratio (GMR) method (13,18). Within each dataset, the T was declared equivalent to R if and only if the lower and upper bounds of the T/R ratio 90% confidence intervals (LB, UB) for all four stage groups were maintained within EMA's ±15% acceptance limits (0.85, 1.18). To study the effect of relaxing the EMA acceptance limits on the outcome of the statistical approach, the analysis was extended by evaluating whether the 90% confidence intervals were maintained within the following T/R ratio ranges: 0.80–1.25 (±20% acceptance limit), 0.75–1.33 (±25% acceptance limit), 0.70–1.43 (±30% acceptance limit), and 0.60–1.67 (±40% acceptance limit). Further, we calculated 70% and 80% confidence intervals and evaluated whether these were maintained within the T/R range of 0.85–1.18 (± 15% acceptance limit). This procedure was applied to all the 1000 replicates of 30 T and 30 R cascade impactor profiles in each scenario, and the T profile within a given scenario (1000 datasets) was judged as equivalent to the R profile if more than or equal to 50% or 80% of the 1000 replicates met the ABE approach criteria.

$$\mathrm{T/R\ ratio\ (100–\alpha)\%CI}: e^{(MeanDiff \pm ME)}$$

$$MeanDiff = \left(\frac{\sum_{i=1}^{n_T} T_i}{n_T} - \frac{\sum_{j=1}^{n_R} R_j}{n_R} \right)$$

$$ME = t_{(1-\alpha/2),df} \cdot \sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}$$

where T_i = natural logarithm transformed deposition of the *i*th ($i = 1, \ldots n_T = 30$) cascade impactor profile for the T product within each group

R_j = natural logarithm transformed deposition of the *j*th ($j = 1, \ldots n_R = 30$) cascade impactor profile for the R product within each group

s_T = standard deviation of the natural logarithm transformed deposition of the *i*th ($i = 1, \ldots n_T = 30$) cascade impactor profile for the T product within each group

s_R = standard deviation of the natural logarithm transformed deposition of the *j*th ($j = 1, \ldots n_R = 30$) cascade impactor profile for the R product within each group

α = type I error

$t_{(1-\alpha/2)}$ = quantile of t-distribution corresponding to (1 − α/2) probability and Wald-Statterthwite's degrees of freedom (df)

$$df = \frac{\left(\frac{s_R{}^2}{n_R} + \frac{s_T{}^2}{n_T} \right)^2}{\frac{1}{n_R-1}\left(\frac{s_R{}^2}{n_R} \right)^2 + \frac{1}{n_T-1}\left(\frac{s_T{}^2}{n_T} \right)^2}$$

2. Chi-square ratio statistic approach (PBE-CSRS)

Results of a previously published study were used in which the CSRS approach was applied to the 55 PQRI scenarios as follows in two steps (10):

Step 1: To compare the impactor-sized mass (ISM) of T and R products, the population bioequivalence (PBE) method was applied to each of the 1000 datasets of all the 55 PQRI scenarios using the reference- or constant-scaled linearized PBE criterion (shown below) approach described in the FDA's "draft guidance on Budesonide" (which specified a constant critical value of 7.66) using the statistical software R (version 3.4.4) (7). First, for each cascade impactor profile, ISM was computed. For each dataset of 30 T and 30 R cascade impactor profiles, 95% upper confidence bound of the reference- or constant-scaled linearized PBE criterion for ISM (U_{95}) was computed. The T was declared equivalent to R if and only if the U_{95} was found to be less than or equal to zero. If a given dataset (consisting of 30 T and 30 R profiles of a given scenario) lacked equivalence in ISM, the overall test for this dataset was defined as failed.

Linearized criteria:

$$\boldsymbol{\eta_1}=(\boldsymbol{\mu_T}-\boldsymbol{\mu_R})^2+(\boldsymbol{\sigma_T^2}-\boldsymbol{\sigma_R^2})-\boldsymbol{\theta_p}\cdot\boldsymbol{\sigma_R^2}<0 \qquad \textit{for}\ \boldsymbol{\sigma_R}>\boldsymbol{\sigma_{T0}}$$

$$\boldsymbol{\eta_2}=(\boldsymbol{\mu_T}-\boldsymbol{\mu_R})^2+(\boldsymbol{\sigma_T^2}-\boldsymbol{\sigma_R^2})-\boldsymbol{\theta_p}\cdot\boldsymbol{\sigma_{T0}^2}<0 \qquad \textit{for}\ \boldsymbol{\sigma_R}\leq\boldsymbol{\sigma_{T0}}$$

where

$\boldsymbol{\mu_T-\mu_R}$: mean difference of T (log scale) and R (log scale) products

$\boldsymbol{\sigma_T^2,\sigma_R^2}$: total variance of T and R products

$\boldsymbol{\sigma_{T0}}$: Regulatory constant = 0.1

$\boldsymbol{\theta_p}$: Regulatory constant calculated as following:

$$\boldsymbol{\theta_p}=\frac{[\boldsymbol{ln}(\mathbf{1.11})]^2+\mathbf{0.01}}{\mathbf{0.1^2}}=\mathbf{2.089}$$

Step 2: The chi-square ratio statistic algorithm (as described in the FDA June 1999 draft guidance for industry) was applied in SAS software to all given datasets of a given scenario if ISM was judged as equivalent (9,10). First, all the cascade impactor profiles were normalized, *i.e.*, all deposition sites were expressed in percent of total mass deposited (TM; the sum of amount deposited on all deposition sites). From each dataset of 30 T and 30 R normalized cascade impactor profiles, 500 triplets (2 R profiles: R_k and R_m; $k \neq m$; $k = 1,\ldots, 30$; $m = 1,\ldots, 30$; and 1 T profile: T_j; $j = 1,\ldots, 30$) were resampled with replacement and the CSRS of each triplet was calculated using the computational form shown below.

$$\boldsymbol{CSRS_{jkm}}=\frac{\sum_{i=1}^{p}\frac{\left(\boldsymbol{T_{ij}}-\frac{\boldsymbol{R_{ik}+R_{im}}}{\mathbf{2}}\right)^{\mathbf{2}}}{\frac{\boldsymbol{T_{ij}}+\frac{\boldsymbol{R_{ik}+R_{im}}}{\mathbf{2}}}{\mathbf{2}}}}{\sum_{i=1}^{p}\frac{(\boldsymbol{R_{ik}-R_{im}})^{\mathbf{2}}}{\frac{\boldsymbol{R_{ik}+R_{im}}}{\mathbf{2}}}}$$

where p = number of deposition sites of the cascade impactor profile.

T_{ij} = normalized deposition on the *i*th deposition site of the *j*th cascade impactor profile for the T product.

R_{ik} and R_{im} = normalized deposition on the *i*th deposition site of the *k*th and *m*th cascade impactor profile respectively where the *k*th and *m*th cascade impactor profiles represent two different samples of the same R product.

Subsequently, the mean of the 500 CSRS's was calculated and this procedure was repeated for 300 times to obtain the distribution of the mean of CSRS. The *T* was declared equivalent to *R* if and only if the 95th percentile of the distribution of the mean of CSRS was found to be less than the fixed critical value of 7.66, as described in the FDA's 1999 draft guidance (which specified a constant critical value of 7.66) (9,10). Within the chi-square ratio statistic approach, for each of the 1000 datasets, the T was declared equivalent to R only if it met the bioequivalence criteria of both population bioequivalence and chi-square ratio statistic test (*i.e.*, steps 1–2). Finally, the T profile within a given scenario (1000 datasets) was judged as equivalent to the R profile if more than or equal to 50% or 80% of the 1000 datasets showed equivalence.

3. Modified chi-square ratio statistic approach (PBE-mCSRS)

The procedure for this statistical approach was described in a previous publication (4). Briefly, this approach involves the following three steps:

Step 1: This step was identical to Step 1 under CSRS approach as described above except that the PBE was applied to the total mass (TM, sum of amount deposited on all deposition sites, as surrogate for single actuation content). The T product within a dataset of 30 T and 30 T profiles was declared equivalent to R if and only if the 95% upper confidence bound of the reference- or constant-scaled linearized PBE criterion for total mass (U_{95}) was found to be less than or equal to zero. If a given dataset (consisting of 30 T and 30 R profiles of a given scenario) lacked equivalence in single actuation content, the overall test for this dataset was defined as failed.

Step 2: PBE for ISM was performed for all given datasets of a given scenario if single actuation content was judged as equivalent. The statistical procedure was identical to that in step 1 under mCSRS approach (described above) except that the PBE was applied to ISM instead of single actuation content. Again, the T was declared equivalent to R if and only if the 95% upper confidence bound of the reference- or constant-scaled linearized PBE criterion for ISM (U_{95}) was found to be less than or equal to zero. If a given dataset (consisting of 30 T and 30 R profiles of a given scenario) lacked equivalence in ISM, the overall test (2 PBEs and 1 mCSRS) for this test unit was defined as failed.

Step 3: The modified chi-square ratio statistic algorithm was applied to all given test units of a given scenario if TM and ISM was judged as equivalent. This algorithm was applied only to the ISM deposition sites as described in the previous publications using the statistical software R (version 3.4.4) (2–4). This involves two steps:

Step 3a (calculation of the test statistic): First, all the ISM cascade impactor profiles of an individual run were normalized, *i.e.*, the percent of ISM mass (obtained by dividing the absolute deposition on each ISM deposition site by the sum of amount deposited on all ISM deposition sites) deposited on each deposition site "i" (i = 1,…,p) was calculated for a given profile. From each dataset of 30 T and 30 R normalized ISM cascade impactor profiles, 2000 bootstrapped replicates of 30 T and 30 R cascade impactor profiles were obtained. For each of the 900 pairs of test (T_j; j = 1,…,30) and reference (R_i; i = 1,…,30) cascade impactor profiles in a bootstrapped replicate, the mCSRS was calculated using the computational form

$$\boldsymbol{mCSRS_{jk}} = \frac{\sum_{i=1}^{p} \frac{\left(\boldsymbol{T_{ij}} - \overline{\boldsymbol{R}}_{\boldsymbol{i}}\right)^{2}}{\overline{\boldsymbol{R}}_{\boldsymbol{i}}}}{\sum_{i=1}^{p} \frac{\left(\boldsymbol{R_{ik}} - \overline{\boldsymbol{R}}_{\boldsymbol{i}}\right)^{2}}{\overline{\boldsymbol{R}}_{\boldsymbol{i}}}}$$

where p = number of deposition sites of the normalized ISM cascade impactor profile

T_{ij} = normalized deposition on the ith deposition site of the jth cascade impactor profile for the T product.

R_{ik} = normalized deposition on the ith deposition site of the kth cascade impactor profile for the R product.

$\overline{R}_i$ = sample mean of the i^{th} deposition site of all ISM normalized R cascade impactor profiles in a dataset.

Subsequently, the median of the 900 mCSRS's was computed for all the 2000 bootstrapped replicates resulting in a distribution for the median of the mCSRS. From this distribution, the 90% bias corrected and accelerated (BCA) upper confidence bound of the median of the mCSRS was computed that served as the test statistic for this procedure (4).

Step 3b (Calculation of the critical value): As described in a previous publication, the critical value for this procedure depends on the maximum allowable difference between T and R (acceptance limit) and the variability of the R product (4). For each dataset, the variability of the R product was estimated by computing the reference variance scaling metric (RVS, equation given below) of the normalized ISM cascade impactor R profiles.

$$RVS = \sqrt{\frac{\sum_{i=1}^{p} \overline{\boldsymbol{R}}_{\boldsymbol{i}} * \boldsymbol{CV_i^2}}{\sum_{i=1}^{p} \overline{\boldsymbol{R}}_{\boldsymbol{i}}}}$$

where RVS = Reference Variance Scaling metric for each dataset of 30 T and 30 R normalized ISM cascade impactor profiles.

CV_i = coefficient of variation (%) of the *i*th deposition site of the normalized ISM cascade impactor profiles of the R product (also called RSD, relative standard deviation, the sample standard deviation expressed in percent of the average); p = total number of ISM deposition sites.

Subsequently, the critical value for each of the 1000 datasets at ± 10%, ± 15%, ± 20%, ± 25%, and ± 30% acceptance limits were computed using the equations shown below derived in a previous publication (4):

$$C10 = 0.993 + 124 * RVS^{-2}$$

$$C15 = 0.970 + 294 * RVS^{-2}$$

$$C20 = 0.949 + 536 * RVS^{-2}$$

$$C25 = 0.916 + 856 * RVS^{-2}$$

$$C30 = 0.896 + 1245 * RVS^{-2}$$

where C10, C15, …, C30 = critical values at ± 10%, ± 15%, …, ± 30% acceptance limits respectively for each dataset of 30 T and 30 R normalized ISM cascade impactor profiles.

RVS = Reference Variance Scaling metric for each dataset of 30 T and 30 R normalized ISM cascade impactor profiles.

The T was declared equivalent to R if and only if the test statistic (from Step 3a) was found to be less than the critical value (from Step 3b). Within the mCSRS approach, for each of the 1000 datasets, the T was declared equivalent to R only if it met the bioequivalence criteria of both the population bioequivalence test and mCSRS test at ± 25% acceptance limit (*i.e.*, steps 1–3). Finally, the T profile within a given scenario (1000 datasets) was judged as equivalent to the R profile if more than or equal to 50% or 80% of the 1000 datasets showed equivalence.

Comparison of the Outcomes of the 55 PQRI Scenarios from the Three Statistical Approaches to that of the Experts' Opinion

Results of the statistical tests for a given scenario (% of the 1000 datasets resulting in equivalence) were plotted against the experts' opinion (% of subject matter experts classifying R and T profiles of a given scenario as equivalent). Classifying a scenario as "equivalent" either at the 50% (T and R are judged as equivalent if more than or equal to 50% of the datasets of a given scenario suggested equivalence) or at the 80% threshold level (T and R are judged as equivalent if more than or equal to 80% of the datasets of a given scenario were suggested to be equivalent) and considering the experts' opinion (at ≥ 50% or ≥ 80% threshold values) as surrogate for "the truth," each scenario could fall into one of the four quadrants of the scatter plots (see Fig. 2): (1) top-right quadrant PP (experts' opinion: pass; statistical approach: pass), (2) top-left quadrant FP (experts' opinion: fail; statistical approach: pass), (3) bottom-left quadrant FF (experts' opinion: fail; statistical approach: fail), and (4) bottom-right quadrant PF (experts' opinion: pass; statistical approach: fail). Subsequently, the estimates of percent agreement, false pass rate (complement of true fail rate, also defined as complement of specificity), and false fail rate (complement of true pass rate, also defined as complement of sensitivity) were computed for each statistical approach. Percent agreement with experts' opinion was defined as the percent of the 55 scenarios that fall into PP and FF categories at a specified threshold value (≥ 50% or ≥ 80%). False pass rate was defined as the percent of scenarios for which the statistical test suggested "pass" while the experts classified them as "fail" (*i.e.*, the number of scenarios falling into the FP quadrant divided by the number of scenarios present in FP and FF quadrants) at a specified threshold value (≥ 50% or ≥ 80%) (14). False fail rate was defined as the percent of scenarios for which the statistical test suggested fail while the experts classified them as pass (*i.e.*, the number of scenarios falling into the PF quadrant divided by the number of scenarios present in PF and PP quadrants) at a specified threshold value (≥ 50% or ≥ 80%) (14).

To get an estimate of the accuracy of each statistical approach (ABE, PBE-CSRS, and PBE-mCSRS) in comparison to the experts' opinion (at ≥ 80% threshold value), receiver operating characteristic (ROC) curves were constructed using the R statistical software package "pROC" (14). The accuracy of each statistical approach was estimated from the ROC curves (see Fig. 2) as the area under the ROC curve (AUC), and the 95% confidence intervals for the AUC were calculated by DeLong method (14). Statistical significance testing of the difference among AUC of the ROC curves was performed at 5% significance level using the R statistical software package "pROC" based on the non-parametric DeLong's test for comparing correlated ROC curves and pairwise comparisons were made based on Bonferroni-adjusted p values (19–21). In addition, to determine the relative performance of the three approaches at high sensitivity values, point estimates and 95% CI of the specificities at 90% and 95% sensitivities were computed for each approach using the R statistical software package "pROC." Finally, to understand the behavior of the statistical tests for evaluating the equivalence in shape of the cascade impactor profiles in relation to R formulation variability, the linear relationship between the outcomes (pass rate) of CSRS test alone, mCSRS alone for the 55 PQRI scenarios, and mean reference variance (MRV, a cumulative estimate of the R formulation variability for a given scenario) was assessed and compared with that of the ABE approach and the experts' opinion (see Fig. 4). MRV for a given scenario was obtained by calculating the arithmetic mean of reference variance scaling (RVS, equation given above) of each of the 1000 replicates of 30 R cascade impactor profiles.

Cascade Impactor Profiles of Three Experimental DPI Formulations (as Assessed by ABE, PBE, and mCSRS Statistical Tests) and the Outcome of Pharmacokinetic (PK) Study

Cascade impactor profiles (Fig. 6) of three experimental DPI-FP formulations generated within 1 month of storage at room temperature by standard USP methodology using a next-generation impactor (NGI) at 60 L/min ($N = 10$ for each formulation),were evaluated by ABE, PBE, and mCSRS (using a critical value of 25% cutoff) statistical tests. Within the mCSRS statistical approach, PBE was assessed for ISM (ISM defined as drug deposited on NGI stage 1 through filter) and fine particle dose (FPD) less than 3 μm, which is defined as drug deposited on NGI stage 4 through filter. Results were compared with those of a clinical PK study, presented at a recent conference (17) and employing the same three DPI-FP formulations. Within the PK analysis, $AUC_{0\text{-inf}}$ was used as metric for the extent of absorption while dose-adjusted C_{max} as metric sensitive to regional lung deposition differences (17).

RESULTS

When the ABE method was applied to compare T and R profiles of the 55 scenarios, only 4 scenarios were judged to be equivalent (threshold level ≥ 50%). A larger number of equivalent scenarios was suggested by the PBE-mCSRS method (15 scenarios), by PQRI subject matter experts (31 scenarios) and the PBE-CSRS method (36 scenarios) with a ≥ 50% threshold value. Quadrant plots compared the results obtained with ABE, PBE-CSRS, and PBE-mCSRS with those of the subject matter experts' opinion (Fig. 2, classification threshold ≥ 50%). The percent agreement with experts' opinion, false pass rate, and false fail rate for the three statistical approaches at ≥ 50% and ≥ 80% classification threshold are further summarized in Tables I and II, respectively.

The ROC curves for the three statistical approaches (Fig. 3) along with the corresponding analysis (AUC [95% DeLong's confidence interval], a cumulative measure of the accuracy of the statistical approaches) are shown in Fig. 3 and Table III, indicating the highest accuracy for the PBE-mCSRS method. To compare the performance of ROC curves at high sensitivity values, 95% CI of specificities for each approach at 0.90 and 0.95 sensitivity values are shown in Table IV, indicating the best performance of PBE-mCSRS approach with higher specificity values.

Because of the high failing rate of the ABE when EMA's criteria were used (confidence interval 90%; range of bioequivalence limit: 0.85–1.18), outcomes using different criteria were compared with the experts' opinions (Table V). Similarly, how differences in the mCSRS acceptance criteria would change the outcome of PBE-mCSRS test was also evaluated (Table VI).

Since the PBE-CSRS and PBE-mCSRS methods, as proposed by Christopher et al. (9,10) and Weber et al. (4), include PBE assessments for ISM (CSRS and mCSRS) and

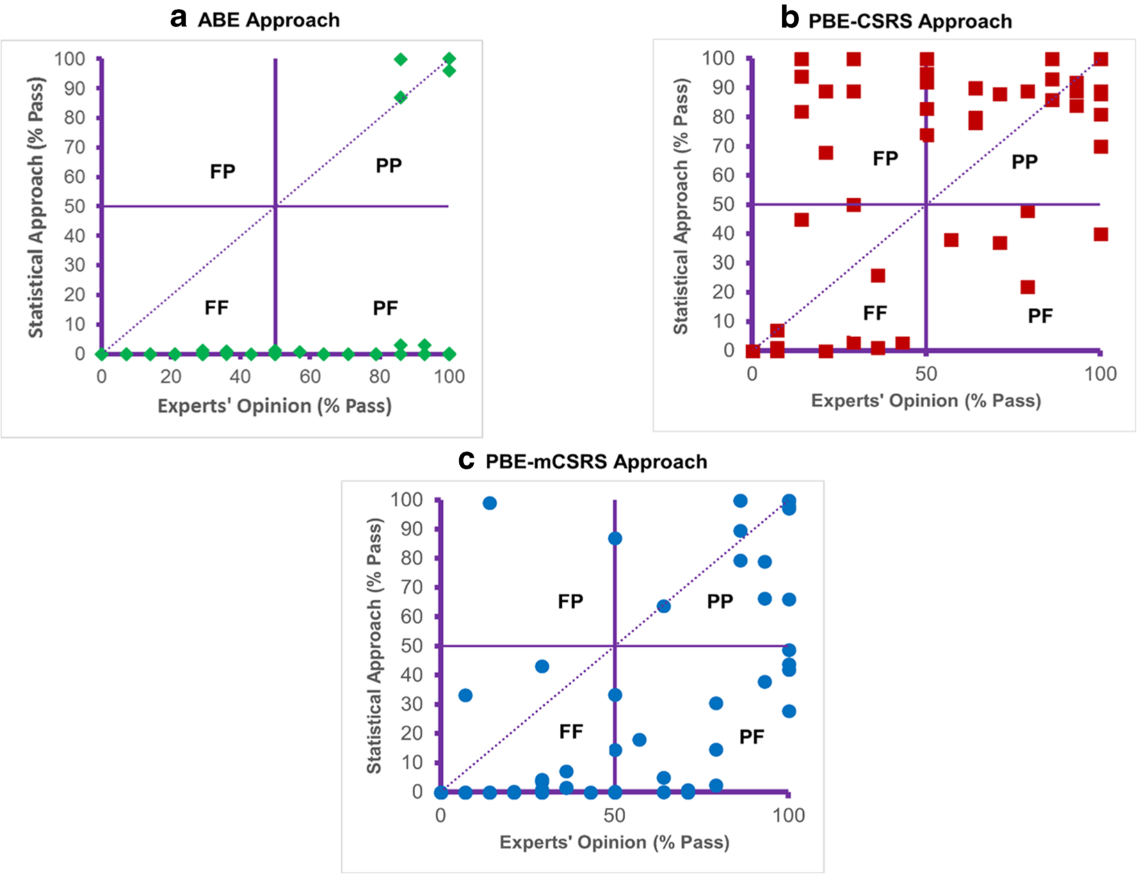

Fig. 2. Scatter plots comparing the results of the three statistical approaches. **a** Average bioequivalence approach (ABE). **b** Chi-square ratio statistic approach (PBE-CSRS). **c** Modified chi-square ratio statistic approach (PBE-mCSRS); x-axis: Percent of experts that declared T equivalent to R for each scenario; y-axis: Percent of 1000 simulated datasets that met the T and R equivalence criteria as per the particular statistical approach for each scenario. Four quadrants: (1) Higher right quadrant PP (experts' opinion: pass; statistical approach: pass), (2) higher left quadrant FP (experts' opinion: fail; statistical approach: pass), (3) lower-left quadrant FF (experts' opinion: fail; Statistical approach: fail), (4) lower-right quadrant PF (experts' opinion: pass; statistical approach: fail). Quadrants are based on a passing criterion of ≥ 50% (A scenario was judged as equivalent if greater than or equal to 50% of the 1000 data sets or greater than or equal to 50% of the experts judged a given scenario equivalent). Expert opinions have previously been reported (9,10)

single actuation content (mCSRS), it was of interest to evaluate the discriminatory power of CSRS and mCSRS alone to identify differences in shape only (not including PBE assessments). Results (% of the 1000 datasets for a given scenario passing) were plotted against the mean reference variance (Fig. 4c for CSRS; and 4d for mCSRS). While considering both, shape and amount, results provided by the expert's (Fig. 4a) and the ABE method (Fig. 4b) are shown for comparison. Overall, the CSRS method lacked any discriminatory power as T and R profiles are judged to be equivalent for most scenarios. A higher discriminatory power was observed for the mCSRS method across a wide range of reference variances. Plotting the difference between passing rate of CSRS and mCSRS for a given scenario *vs* the

Table I. Agreement of ABE, PBE-CSRS, and PBE-mCSRS Approaches with the Experts' Opinion at ≥ 50% Threshold[a]

Statistical approach	Number of 55 PQRI scenarios that met the equivalence criteria	Agreement with experts' opinion	False pass rate	False fail rate
ABE	4	28/55 = 50.9%	0/24 = 0%	27/31 = 87.1%
PBE-CSRS	36	40/55 = 72.7%	10/24 = 41.7%	5/31 = 16.1%
PBE-mCSRS	15	37/55 = 67.3%	1/24 = 4.2%	17/31 = 54.8%

[a] A scenario was judged as equivalent if for a given scenario greater than or equal to 50% of the experts or greater than or equal to 50% of the 1000 datasets indicated equivalence between T and R profiles

Table II. Agreement of ABE, PBE-CSRS, and PBE-mCSRS Approaches with Experts' Opinion at ≥ 80% threshold[a]

Statistical approach	Number of 55 PQRI scenarios that met the equivalence criteria	Agreement with experts' opinion	False pass rate	False fail rate
ABE	4	42/55 = 76.4%	0/38 = 0%	13/17 = 76.5%
PBE-CSRS	31	37/55 = 67.3%	16/38 = 42.1%	2/17 = 11.8%
PBE-mCSRS	10	44/55 = 80%	2/38 = 5.3%	9/17 = 52.9%

[a] A scenario was judged as equivalent if for a given scenario greater than or equal to 80% of the experts or greater than or equal to 80% of the 1000 datasets indicated equivalence between T and R profiles

observed mean reference variance (Fig. 5) suggested that CSRS and mCSRS judgments differ especially at higher variance (Fig. 5).

Results of CI profile comparisons of three experimental DPI-FP formulations (Fig. 6) obtained by ABE, and results for PBE (ISM: NGI stage 1-filter and FPD < 3 μm: NGI stage 4-filter) and mCSRS test are shown together with the previously reported outcomes of a PK BE study in Table VII (17). In addition, the pros and cons of the three statistical approaches are summarized in Table VIII.

DISCUSSION

In this paper, the outcomes from three statistical approaches were compared to the evaluations of subject matter experts from the PQRI working group. While expert's classification and the ABE method considered the absolute drug amounts on given stages or groups for the equivalence decision, CSRS and mCSRS express stage depositions relative to the cumulative deposited amount (%TM and %ISM, respectively) and therefore only evaluate the shape of the profiles. As outlined in the original publications, it was therefore necessary for CSRS and mCSRS tests to include additional tests into the assessment which probe for dose related differences (single actuation content and/or impactor-sized mass) (4,10). PBE tests for ISM and single actuation content (SAC) are therefore included in the decision-making process.

Methods and data handling were identical to the ones originally proposed in relevant publications or guidance (4,8,10,13). This led to the situation that input data were not always the same. The ABE method considered all deposition sites from which subsequently defined stage groups were generated. The PBE-CSRS test considered all available deposition sites but restricted the PBE method to ISM stages. The PBE-mCSRS test considered potential differences in the SAC, amount of drug deposited on ISM stages, and the shape of the ISM deposition sites. While generally the SAC is determined in separate experiments, we derived the SAC as the sum of all deposition sites (drug deposited in USP throat, pre-separator (if used), and all deposition stages) (7). The assessment of the single actuation content (SAC) as an integrated component within the mCSRS approach refers to the total amount of drug released from the inhalation drug product. This evaluation ensures that the test drug product delivers an equivalent amount of drug relative to the reference product as determined in a specified test as outlined in US Pharmacopeia (USP) 25, < 601 > (22). Unfortunately, SAC was not generated for the 55 PQRI scenarios and hence the total mass (TM, the sum of drug on all accessories and deposition sites of cascade impactor) was used as the best available surrogate for SAC, following the procedure described in a previous publication (4). The total mass represents the sum of individual cascade impactor deposition sites plus inlet and pre-separator depositions, resulting in similar but probably somewhat more variable estimates. Hence, the use of TM instead of SAC (when the reference variability is less than test variability) within the PBE-mCSRS approach will result in a more conservative evaluation, as equivalence will be more difficult to achieve because of higher variability.

The statistical outcomes of the three methods were compared with the historical judgment of expert members of the PQRI working group. There are certain limitations to using experts' opinion as the truth such as lack of complete information on the methods (especially the subjectivity employed for assessing the variability of CI profiles) in evaluating the equivalence of the CI profiles (10). However, considering that no satisfactory predictive *in vitro*–*in vivo* relationship between APSD differences and clinically acceptable differences in lung dose/regional lung deposition of T and R formulations has been established and no alternative statistical test has been validated as a "gold" standard (10), these subjective evaluations were used as the best available surrogate for truth. This was feasible because of the vast experience of the subject matter experts. However, re-evaluation of scenarios suggested that some of the decisions of the experts might have been debatable. For example, in the case of scenario number 32, within which both T and R profiles had identical mean TM, identical mean ISM, and identical variability (%CV), 50% of the subject matter experts concluded that T and R profiles are non-equivalent (while as per the PBE-mCSRS approach, 87% of the simulated T and R datasets met the equivalence criteria). Another example is scenario number 38, which 87% of the subject matter experts concluded that T and R profiles are non-equivalent despite less than 10% difference in the mean TM and mean ISM of T and R profiles, low test and low reference variability (while as per the PBE-mCSRS approach, 99% of the simulated T and R datasets met the equivalence criteria).

One of the challenges in studies like this is to obtain representative datasets describing "real-life" scenarios. We

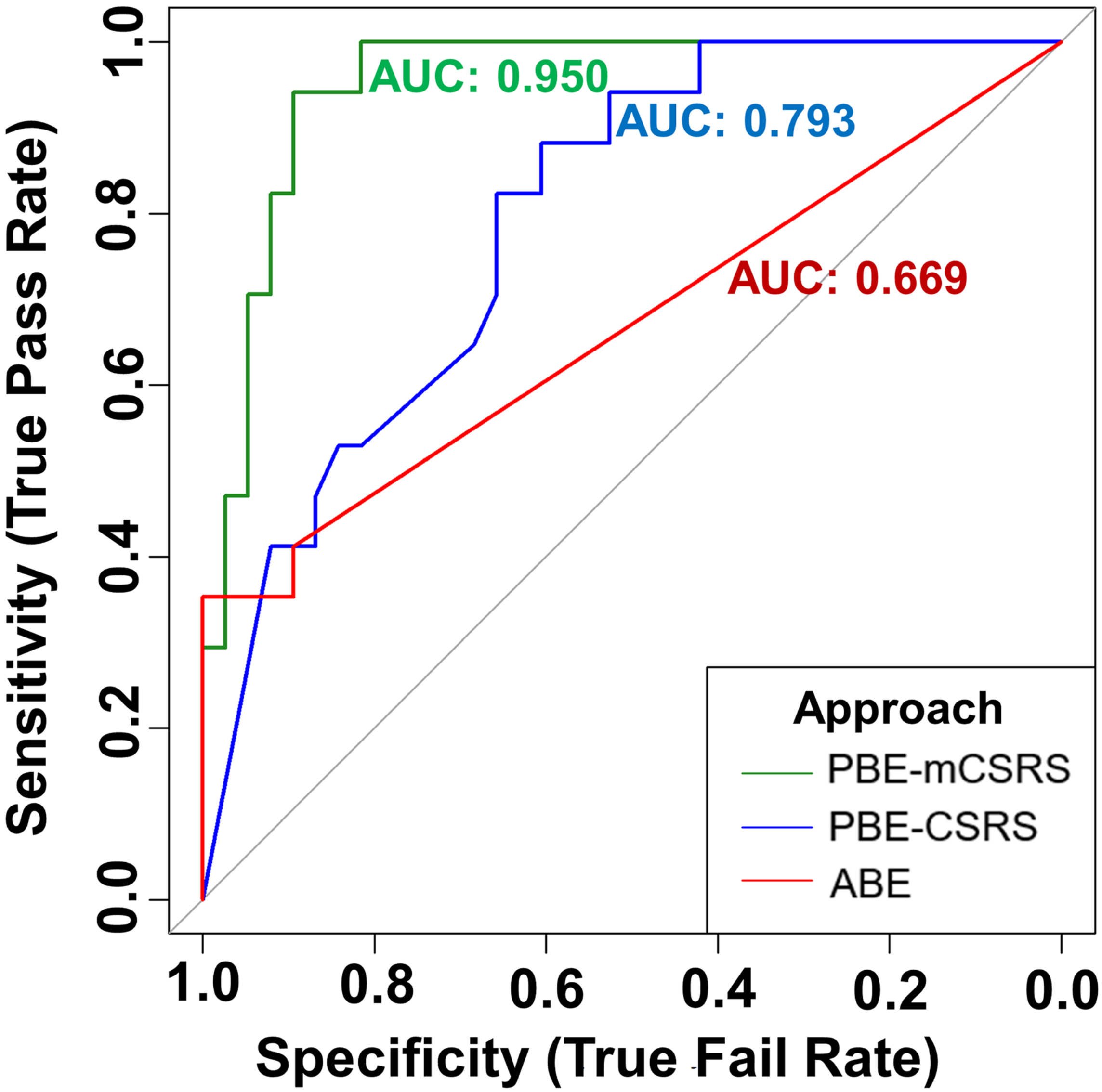

Fig. 3. Receiver operating characteristic (ROC) curves for the three statistical approaches. **a** Average bioequivalence approach (ABE, in red). **b** Chi-square ratio statistic approach (PBE-CSRS, in blue). **c** Modified chi-square ratio statistic approach (PBE-mCSRS, in green) against the experts' opinion (threshold value for experts' opinion is set to be 80%, *i.e.*, if greater than or equal to 80% of the experts declared equivalency, the particular scenario was considered truly equivalent and *vice-versa*) obtained from the R package "pROC". Please note that the direction of the x-axis is reversed. Thus, the x-axis represents false pass rate (the complement of true fail rate). Area under the ROC curves calculated by DeLong's method, AUC [95% confidence interval]–ABE approach: 0.669 [0.534, 0.803]; PBE-CSRS approach: 0.793 [0.675, 0.912]; PBE-mCSRS approach: 0.950 [0.898, 1.000]

decided to use the 55 scenarios originally suggested by PQRI as these were generated by the PQRI working group after receiving information on a total of 14 real life pairs of cascade impactor profiles (patterns observed before and after change) that served as the foundation to further generate a more complete set of scenarios. Considering this, we believe that the dataset remains a representative sample of T and R cascade impactor profiles of orally inhaled formulations on market or in development (10). While the log-normal parametric distribution assumption or non-parametric bootstrapping of actual observations are plausible options for simulation of new datasets for profile comparison investigations, we had to stay with the datasets evaluated by the experts. This multivariate normal distribution assumption used by the PQRI group when generating the original scenarios was based on a previous report which concluded that within all the 55 PQRI scenarios, the absolute recovery amounts (*i.e.*, the actual CI data used in this study) follow an approximately normal distribution on each deposition site of CI profile (23). Further, the 55 scenarios were simulated by the PQRI working group based on their judgment that a multivariate normal distribution would fairly represent real

Table III. Pairwise Comparisons of the Area Under the ROC Curves (AUC) for the Three Statistical Approaches

Comparison	DeLong's "Z" test statistic	*p* value	Bonferroni-adjusted *p* value
PBE-mCSRS *vs* PBE-CSRS	2.84	0.0045	0.0135*
PBE-mCSRS *vs* ABE	3.86	0.0001	0.0003*
PBE-CSRS *vs* ABE	1.44	0.1487	0.4461

*Statistically different AUCs at 5% significance level

data with different shapes and inter-correlation structure between stages. As these were the scenarios the experts evaluated and in order to be consistent with the previous publications, we stayed with the multivariate normal distribution assumption in our continued evaluation of the statistical procedures.

We first compared the outcomes obtained from the three statistical approaches (including PBE tests, where applicable) with the evaluations from subject matter experts (Fig. 2). The ABE approach as suggested by EMA applied at its ± 15% acceptance limit showed poor agreement with the experts' opinion (Tables I, II, and V). Sandell previously reported that it is unlikely to show profile equivalence even when reference product is tested against itself (24). We confirmed Sandell's results, when the 55 scenarios were analyzed using single stage analysis, as none of the 55 scenarios revealed equivalence (data not shown). When we performed the analysis with groups (see methods section for details of grouping) rather than with individual stages, the ABE method suggested equivalence for only 4 of the 55 scenarios (Fig. 2a). Because of the very small number of equivalent scenarios, it was difficult to probe for relationships between variance and passing rate. However, the four scenarios that showed equivalence, exhibited the smallest variance (Fig. 4b).

The main reason for the poor agreement with the expert judgments is that the ABE approach involves the individual assessment of multiple stages or groups, all of which must meet the equivalence criteria. For example, scenario no. 35 and no. 36 which had pass rates greater than 95% based on experts' opinion, mCSRS and CSRS approaches, resulted in 0% pass rate based on the ABE approach owing to the lack of equivalence in group 2 (representing coarse mass) and group 4 (representing extra-fine particle mass). It should be noted that both group 2 and group 4 represents a significantly small proportion of the total mass deposited in the cascade impactor, prone to high analytical variability and might not be clinically relevant. Since the EMA's approach places equal weight on all the four groups, it led to high false fail rate which is in agreement with the previously reported literature (3,13). Thus, our data together with those from Sandell further underline that EMA's ABE approach and acceptance criteria are too restrictive and unrealistic, even if the grouping approach is applied (24). Less stringent criteria (Table V) were able to increase the pass rate (T and R profiles in 7, 10, 14, and 27 scenarios were judged to be equivalent at ± 20%, ± 25%, ± 30%, and ± 40% difference acceptance limits respectively) and the agreement with the expert opinion (56.4% agreement at acceptance limit of ± 40% compared to 50.9% at ± 15%) which is only slightly lower than the 67.3% agreement observed for mCSRS approach (Tables I and V). While under these conditions (acceptance limit: ± 40%), the false fail rate was reduced from 87.1% to 45.2%, slightly lower than the 54.8% for the PBE-mCSRS approach, the false pass rate sky-rocketed to 41.7% (compared to 4.2% for the PBE-mCSRS approach), a value that is not acceptable considering patient safety concerns. Overall, EMA's ABE approach using multiple stage (group) comparisons is too stringent when the current acceptance criteria are applied (false fail rate too high) or do not provide enough patient protection if criteria are loosened. We were unable to identify any acceptable compromise. The solution might be, as suggested by Sandell and in line with the approach taken by FDA for PBE and by Weber et al. for mCSRS, to scale the acceptance criteria for each stage or group according to the variability of the R products, so the same ± 15% limits are not used for all endpoints regardless of their variability.

The PBE-CSRS approach suggested the largest number of scenarios for which T and R products were judged to be equivalent (36 scenarios, Fig. 2b at 50% threshold level). While this resulted in a high agreement with the subject matter experts' opinions (72.7% for the 50% threshold and 67.3% for the 80% threshold; Tables I and II), it also translated into the highest number of false-positive decisions (41.7%, for the 50% threshold, 42.1% for the 80% threshold Fig. 2b, Tables I and II). This method is therefore unlikely to ensure patient's safety in a consistent manner. This re-analysis of the scenarios using slightly different approaches for assessing the method performance than originally reported by PQRI is in full support of the original conclusions (10). As apparent from Fig. 4c, most of the 55 PQRI scenarios showed 100% pass rate independent of the reference variance. Thus, the discriminatory power of this method is purely driven by the ISM-PBE. The inability of the CSRS method to identify

Table IV. 95% CI of Specificities for the Three Statistical Approaches at 0.90 and 0.95 Sensitivity Values

Statistical approach	Sensitivity	Specificity (95% CI)
ABE	0.90	0.15 (0.11, 0.25)
PBE-CSRS	0.90	0.55 (0.32, 0.76)
PBE-mCSRS	0.90	0.89 (0.74, 0.97)
ABE	0.95	0.08 (0.05, 0.13)
PBE-CSRS	0.95	0.45 (0.26, 0.68)
PBE-mCSRS	0.95	0.84 (0.68, 0.97)

Table V. Agreement of ABE Approach with Experts' Opinion for a Range of Acceptance Limits at ≥ 50% threshold[a]

Acceptance limit	Confidence level	Number of 55 PQRI scenarios that met the equivalence criteria	Agreement with experts' opinion	False pass rate	False fail rate
EMA: ±15%	EMA: 90%	4	28/55 = 50.9%	0/24 = 0%	27/31 = 87.1%
±15%	80%	4	28/55 = 50.9%	0/24 = 0%	27/31 = 87.1%
±15%	70%	4	28/55 = 50.9%	0/24 = 0%	27/31 = 87.1%
±20%	90%	7	27/55 = 49.1%	2/24 = 8.3%	26/31 = 83.9%
±25%	90%	10	28/55 = 50.9%	3/24 = 12.5%	24/31 = 77.4%
±30%	90%	14	30/55 = 54.6%	4/24 = 16.8%	21/31 = 67.7%
±40%	90%	27	31/55 = 56.4%	10/24 = 41.7%	14/31 = 45.2%

[a] A scenario was judged as equivalent if for a given scenario greater than or equal to 50% of the experts or greater than or equal to 50% of the 1000 datasets indicated equivalence between T and R profiles

non-equivalent scenarios is likely due to the use of a fixed critical value within the CSRS test, not considering reference variance or the selection of a critical value that was too relaxed (2,10). As shown by the PQRI WG, change of the critical value from 7.66 to 2.75 increased the number of scenarios not showing equivalence; however, this did not improve the agreement with the expert's judgment (10).

It was more challenging to demonstrate equivalence of T and R APSD profiles when the PBE-mCSRS approach, employing reference variance scaling, was applied to the data. The number of scenarios for which T and R products were judged to be equivalent was smaller (15 scenarios at ≥ 50% threshold level) than predicted by the PBE-CSRS method (36 scenarios) or proposed by subject matter experts (31 scenarios). Despite the lower number of equivalent scenarios, the PBE-mCSRS method showed the highest or second highest agreement with the expert opinion at ≥ 80% and ≥ 50% threshold criteria, respectively. More importantly, the PBE-mCSRS approach, with only a very few false pass decisions (4.2%, Table I; 5.3%, Table II) struck a good balance between patient's risk and manufacturer's risk (Tables I and II). It should be noted that always for the purpose of comparison with experts' opinion, within PBE-mCSRS approach, the mCSRS critical values at ± 25% acceptance limit were employed since good agreement with the experts' opinion with a reasonably false pass rate was observed at this acceptance limit (Table VI) which is in accordance with the previously reported literature (4).

To further investigate the overall accuracy of the three statistical approaches, we compared the corresponding ROC curves using the expert's opinion as the surrogate for truth and found that the mCSRS approach has significantly higher accuracy (Bonferroni-adjusted p value < 0.05) compared to the other two approaches (see Fig. 3 and Table III). ROC analysis indicated that the integration of population bioequivalence methods with the mCSRS test (mCSRS approach) improved the overall accuracy from 84% (with mCSRS test alone, data not shown) to 95% (with the combined mCSRS approach). Thus, the stepwise mCSRS approach which ensures equivalence both in terms of absolute deposition and the shape of the CI profile is valuable for making correct decisions. Moreover, unlike the ABE approach, the mCSRS test (by the design of the test statistic) puts more weight on the high deposition sites that are less variable and clinically more relevant and less weight on the low deposition sites leading to its superior performance (3).

A critical difference between CSRS and mCSRS is that the former normalizes stages to the total mass and assesses the complete profile (including non-sizing components and accessories) while the latter normalize stages to impactor-sized mass and assesses only the sized profile. Considering that experts based their judgment on the full profile, it is somewhat surprising that the mCSRS performs better in matching experts' opinion. Had the experts based their evaluation on the sized part of the profile only, the difference between the approaches would most likely have been even more impressive.

Table VI. Agreement of PBE-mCSRS Approach with Experts' Opinion for a Range of mCSRS Acceptance Limits at ≥ 50% threshold[a]

mCSRS Acceptance limit	Number of 55 PQRI scenarios that met the equivalence criteria	Agreement with experts' opinion	False pass rate	False fail rate
± 10%	3	27/55 = 49.1%	0/24 = 0%	28/31 = 90.3%
± 15%	8	32/55 = 58.2%	0/24 = 0%	23/31 = 74.2%
± 20%	12	34/55 = 61.8%	1/24 = 4.2%	20/31 = 64.5%
± 25% (previously used, (4))	15	37/55 = 67.3%	1/24 = 4.2%	17/31 = 54.8%
± 30%	22	38/55 = 69.1%	4/24 = 16.7%	13/31 = 41.9%

[a] A scenario was judged as equivalent if for a given scenario greater than or equal to 50% of the experts or greater than or equal to 50% of the 1000 data sets indicated equivalence between T and R profiles

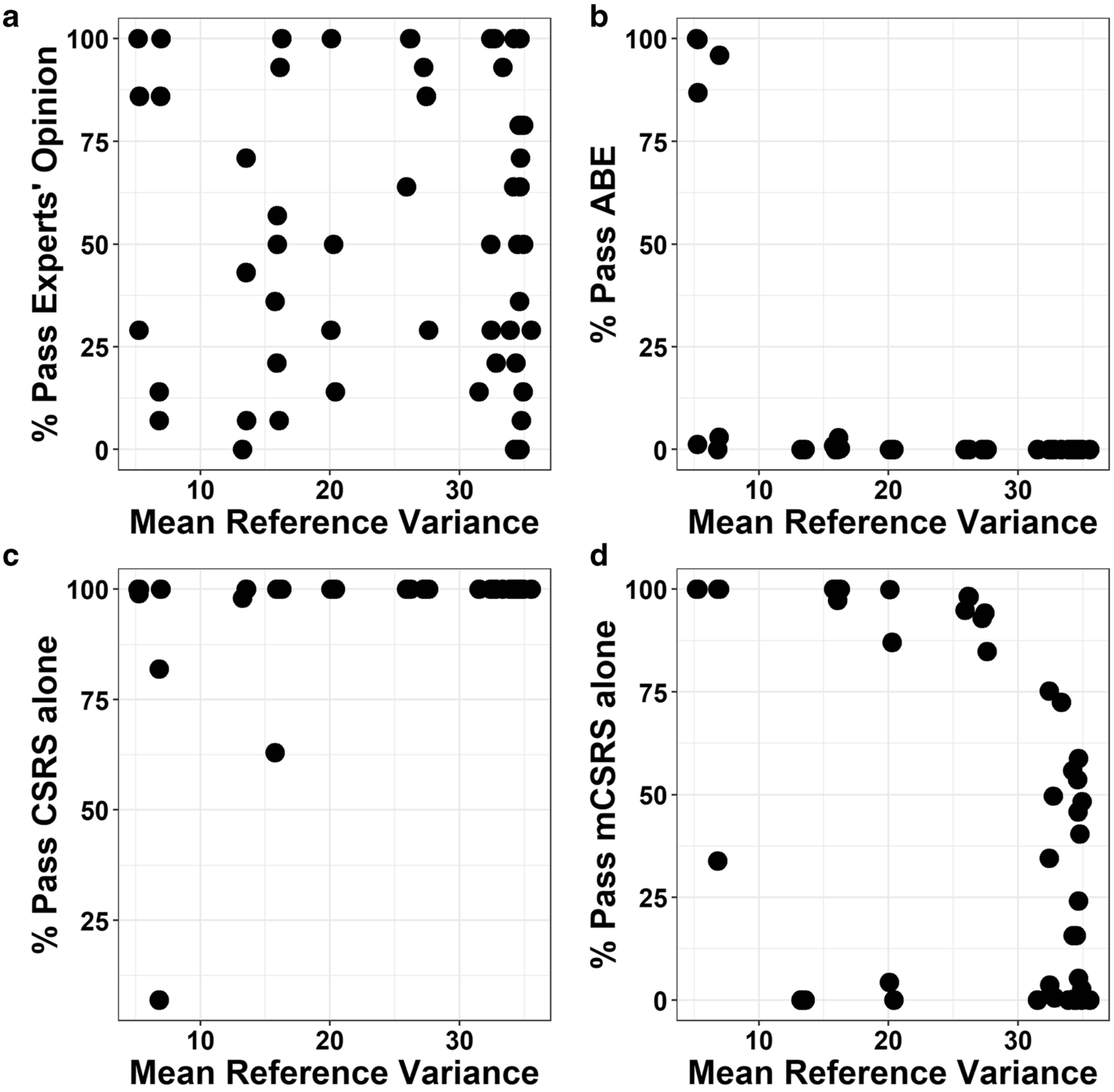

Fig. 4. Pass rate outcomes of the **a** experts' opinion, **b** ABE approach, **c** CSRS test alone (without PBE), and **d** mCSRS test alone (without PBE) as a function of mean reference variance

Considering the results obtained for the CSRS test, it was also of interest to assess the behavior of the CSRS and mCSRS tests alone (when PBE tests assessing ISM and SAC were excluded). As shown in Fig. 4, the mCSRS test alone exhibited higher discriminatory ability compared to the CSRS test alone, especially for scenarios with higher reference variability (MRV > 30). This superior performance of the mCSRS test alone might be attributed to the use of critical values that are scaled to the variability of the reference formulation as the critical value of mCSRS test decreases with increasing R formulation variability, while the critical value of CSRS remains unaltered (3,4). With reference and test variance generally being similar in the dataset of the 55 scenarios (the cumulative T/R variability ratio for the 55 PQRI scenarios was within the narrow range of 0.82–1.29, data not shown), the higher incidences of failed equivalency tests at higher variance makes sense, as it is more likely to fail the mCSRS test if variabilities of test and reference samples are high. In this study, since the cumulative T/R variability ratio for the 55 PQRI scenarios was narrow, the relationship between the pass rate outcomes and T/R variance ratio could not be evaluated. A separate simulation study evaluating the effect of changing T/R variance ratios on the outcome might be of interest.

The potential *in vivo* relevance of the ABE, PBE, and mCSRS statistical tests was further explored by comparing the CI equivalence evaluations of three experimental DPI formulations (Fig. 6) to the results obtained during PK BE studies (17). The statistical evaluation of cascade impactor data should provide, if possible, information on two relevant *in vivo* properties: (a) pulmonary deposited dose and (b) regional lung deposition. As reported elsewhere (17), $AUC_{0\text{-}inf}$ indicated that less drug was absorbed from formulation A, while formulation B and C were equivalent with respect to this parameter. Similarly, dose-adjusted C_{max} estimates indicated differences of formulation A in the regional deposition (more peripheral deposition yields faster absorption), while B and C were equivalent with respect to this parameter. The ABE test results showed lack of equivalence (Table VII) across all three formulations, further supporting that the

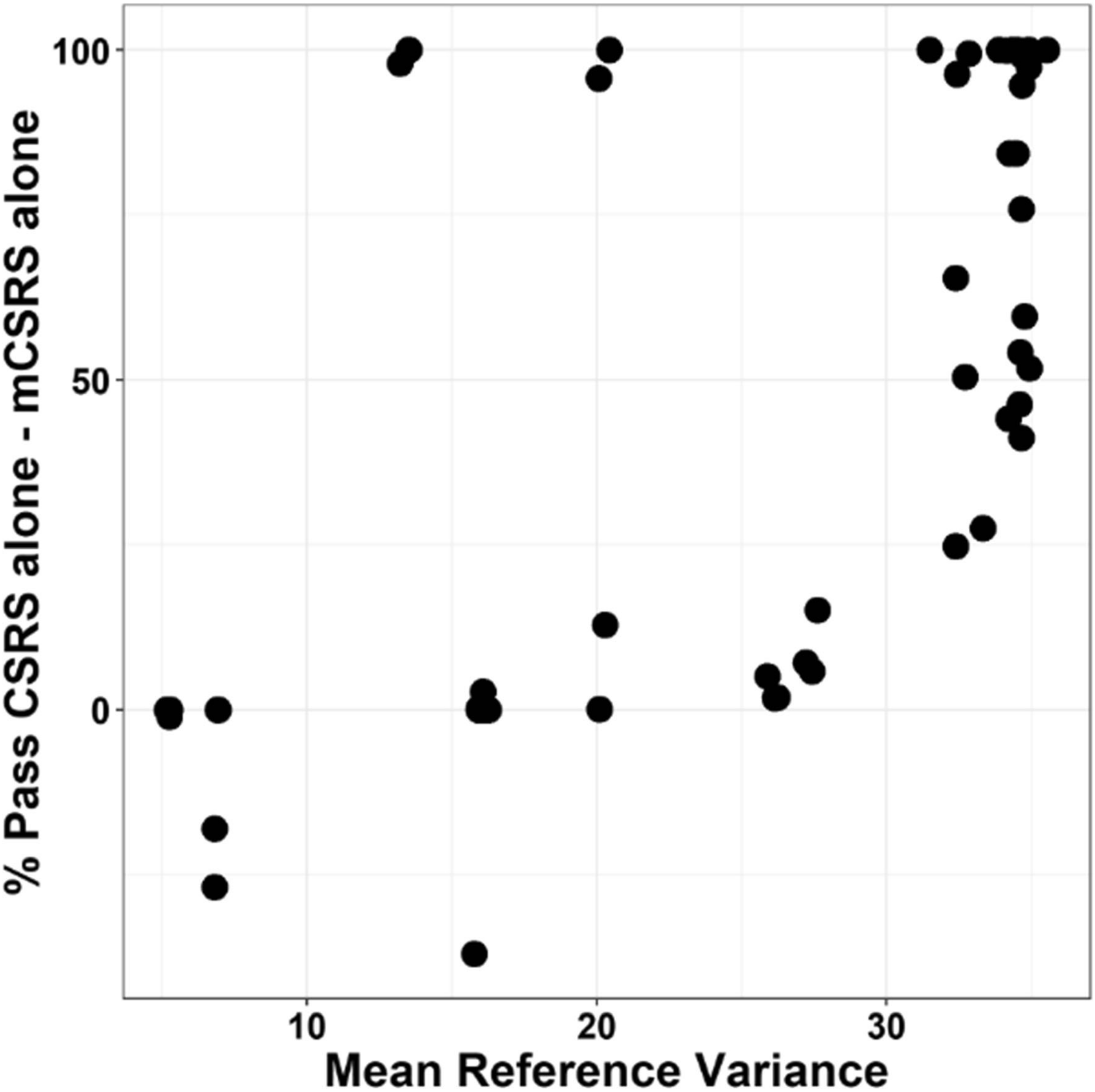

Fig. 5. Difference in the pass rate outcomes of CSRS test alone (without PBE) and mCSRS test alone (without PBE) as a function of mean reference variance

EMA-ABE method is sometimes providing false failing decisions because of the stringent criteria set by the EMA. The high ABE failure rate of the 55 PQRI scenarios is mainly due to the multiple group comparisons and EMA's stringent acceptance criteria (24).

The proposed mCSRS test is a two-step procedure (a: ISM as surrogate for the deposited lung dose; b: mCSRS test for assessing the shape of the profile). Results for the PBE-ISM assessment suggested that the deposited lung dose of formulation A was lower compared to the other two formulations (Table VII), agreeing with the PK results. However, results of PBE test conducted on ISM (NGI stage 1-filter) did not agree with the PK BE results for formulations B and C, as PBE of the ISM (NGI stage 1-filter) also indicated lack of equivalence between the two formulations. This suggests, as also indicated by Newman et al., that drug deposited on NGI stage 1-filter might not correlate with the amount of drug deposited *in vivo* and that particles smaller than 3 μm might be more relevant (25). Indeed, results of PBE analysis based on FPD less than 3 μm (NGI stage 4 through filter) agreed better with PK results, as both suggested bioequivalence for formulations B and C, while formulation A was judged to be bio-IN-equivalent to the other two formulations. It might be therefore of advantage to include such comparisons within future more detailed *in vitro/ in vivo* validations of the proposed method. Evaluating potential differences in the shape of the cascade impactor profiles, mCSRS test suggested equivalence of formulation B and C, while formulation A differed, mirroring the results of the PK study using dose-adjusted C_{max} as relevant metric for identifying differences in regional lung deposition of formulations (Table VII). A more detailed evaluation of dose-adjusted C_{max} as suitable metric for regional deposition has been recently reported (17) and will be provided elsewhere.

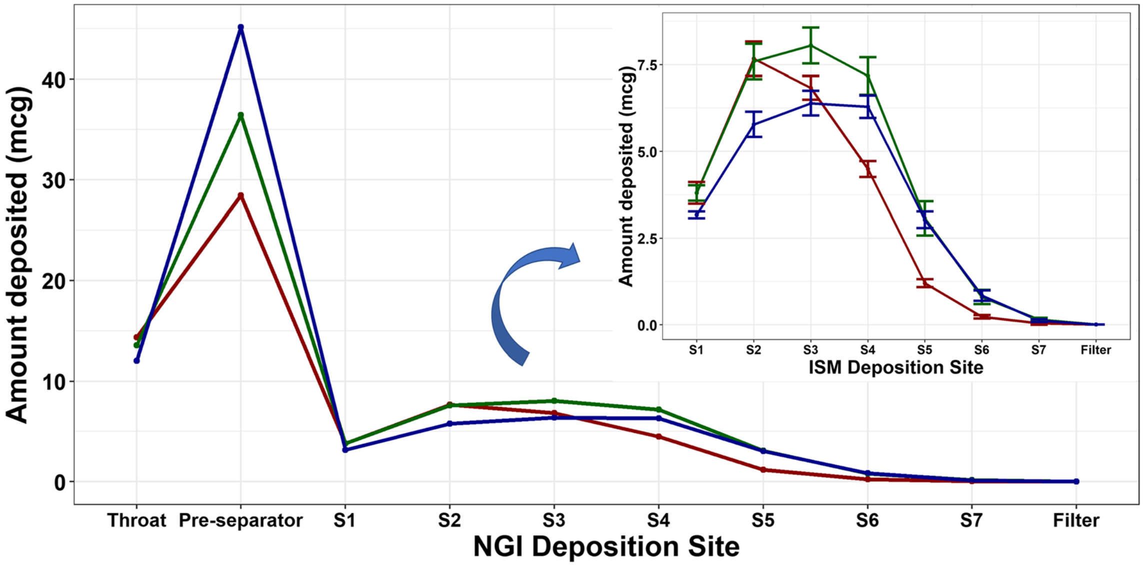

Fig. 6. Mean cascade impactor profiles (collected immediately after the preparation of formulations and after 1 month of storage at room temperature) of three fluticasone propionate dry powder formulations with different mass median aerodynamic diameter: **A** 4.5 – μm (in red), **B** – 3.8 μm (in green), and **C** (in blue). The error bars represent the standard deviation of amount on each deposition site (*N* = 10 for each formulation). The inset shows the zoomed in version of ISM deposition sites

The above discussion suggests that the mCSRS test, in addition to demonstrating high overall agreement with experts' opinion, also captures some properties of the formulations that are of *in vivo* relevance (such as regional lung deposition) reasonably well. However, caution should be exercised while interpreting the CI equivalence outcomes as it is not possible to capture some properties of the formulations (such as dissolution rate behavior of formulations with lipophilic active ingredients) through CI profile comparisons. In summary, the mCSRS test (which compares the shape of the CI profiles) results suggested that the central to peripheral lung deposition might be different for formulation A compared to the other two formulations. Overall, this study explored the *in vivo* relevance of CI profile tests using

Table VII. Cascade Impactor Equivalence Outcomes for Clinical Formulations A—4.5 μm (test), B—3.8 μm (reference), and C—3.7 μm (test)

	Test mean ± SD or Geomean [90%CI]	Reference mean ± SD or Geomean [90%CI]	Outcome[d]
Formulation C (test) *vs* B (reference)			
PBE test on total mass (μg)	82.8 ± 1.03	80.6 ± 3.56	Equivalent (− 0.0197)
PBE test on impactor-sized mass (μg)	25.6 ± 0.874	30.6 ± 2.2	Non-equivalent (0.024)
PBE test on NGI stages 4 through filter (μg)	10.3 ± 0.379	11.2 ± 1.15	Equivalent (− 0.00487)
mCSRS test[a]	N/A	N/A	Equivalent
ABE test[b]	N/A	N/A	Non-equivalent
PK: relative AUC_{inf} (% of R)[c]	94 [83.7–105.6]	100 [89.0–112.3]	Bioequivalent
PK: relative dose normalized C_{max} (% of R)[c]	104 [83.7–124.1]	100 [83.9–119.8]	Bioequivalent
Formulation A (test) *vs* B (reference)			
PBE test on total mass (μg)	67.0 ± 1.90	80.6 ± 3.56	Non-equivalent (0.0233)
PBE test on impactor-sized mass (μg)	24.3 ± 1.12	30.6 ± 2.2	Non-equivalent (0.0538)
PBE test on NGI stages 4 through filter	5.98 ± 0.307	11.2 ± 1.15	Non-equivalent (0.445)
mCSRS test[a]	N/A	N/A	Non-equivalent
ABE test[b]	N/A	N/A	Non-equivalent
PK: relative AUC_{inf} (% of R)[c]	75.0 [66.8–84.2]	100 [89.0–112.3]	Non-bioequivalent
PK: relative dose normalized C_{max} (% of R)[c]	54.0 [45.3–64.9]	100 [83.9–119.2]	Non-bioequivalent

[a] Performed at ± 25% acceptance limit (see methods section)

[b] Performed on four groups of NGI deposition sites (see methods section)

[c] Presented at DIA 2018: Pharmacokinetic Comparison of Locally Acting Dry Powder Inhalers, G. Hochhaus and J. Bulitta, DIA Workshop 2018, Generic Drug-Device Combination Complex Products. Silver Spring, Maryland October 2018. PK details to be published elsewhere

[d] Number in parenthesis represents observed PBE criterion (see "METHODS")

Table VIII. Pros and Cons of the Three Statisitcal Approaches: ABE, PBE-CSRS, and PBE-mCSRS

Method	Pros	Cons
(1) ABE approach (EMA)	• Simple algorithm • Computationally less intensive	• Stringent acceptance criteria • High false fail rate
(2) PBE-CSRS approach	• Considers fine particle mass (PBE test) • Considers shape of the CI profile (CSRS test)	• High false pass rate • Affected by no. of deposition sites • No reference scaling • Complex algorithm
(3) PBE-mCSRS approach	• Considers fine particle mass (PBE test) • Considers shape of the CI profile (mCSRS test) • Reasonably low false pass rate • Not affected by number of deposition sites • Integrates reference scaling	• Complex algorithm • Involves bootstrapping

available limited data (two sets of comparisons with common reference formulation); a detailed study with larger sample size *i.e.* a study covering the entire range of dry powder inhalation drug products present in the market might be of interest in the future.

CONCLUSION

In this paper, we compared the performance of three statistical approaches for testing the equivalence in aerodynamic particle size distribution of orally inhaled drug products. We found that the ABE approach (average bioequivalence as proposed by EMA) is conservative in conferring a pass with high false fail rate, mainly due to equal weight and limit allocated to all multiple group of stages involved in T and R equivalence testing. We also observed that relaxing the EMA acceptance criteria increased false pass decisions rather than improving the performance of the approach. On the other hand, the CSRS approach is more tolerant to differences between T and R products as indicated by the high false pass rate, mainly due to the use of fixed critical value within CSRS test and the lack of considering the reference variability. As we hypothesized, the mCSRS approach was on one hand conservative by providing less false pass decisions but still able to differentiate between equivalent and non-equivalent scenarios (contrary to the EMA approach) across the 55 scenarios with balanced number of false pass and intermediate false-fail rates, most likely due to the scaling of critical value as per the variability of the reference product and other desirable properties of mCSRS test as described above. Finally, the results of PBE-mCSRS approach based on the assessment of APSD profiles of three dry powder inhaler (DPI) formulations were found to be in full agreement with their pharmacokinetic bioequivalence outcomes.

ACKNOWLEDGMENTS

We would like to thank Dr. Benjamin Weber and Dr. Renishkumar Delvadia for their help with computations in this project.

Views expressed in this publication do not necessarily reflect the official policies of the Department of Health and Human Services, nor does any mention of trade names, commercial practices, or organization imply endorsement by the US Government.

FUNDING INFORMATION

The International Pharmaceutical Aerosol Consortium-Regulatory Science (IPAC-RS) funded part of this project. In addition, the Office of Generic Drugs (OGD) at FDA sponsored the *in vitro* CI profile testing and clinical PK study of experimental DPI formulations through GDUFA-funded contracts HHSF223201110117A, HHSF223201610099C, and HHSF223201300479A.

REFERENCES

1. Lu D, Lee SL, R a L, Choi S, Adams W, Caramenico HN, et al. International guidelines for bioequivalence of locally acting orally inhaled drug products: similarities and differences. AAPS J [Internet]. 2015;17(3):546–57 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25758352. Accessed 11 June 2019
2. Weber B, Hochhaus G, Adams W, Lionberger R, Li B, Tsong Y, et al. A stability analysis of a modified version of the chi-square ratio statistic: implications for equivalence testing of aerodynamic particle size distribution. AAPS J [Internet]. 2013 Jan [cited 2015 May 2];15(1):1–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3535092&tool=pmcentrez&rendertype=abstract. Accessed 11 June 2019
3. Weber B, Lee SL, Lionberger R, Li B V, Tsong Y, Hochhaus G. A sensitivity analysis of the modified chi-square ratio statistic for equivalence testing of aerodynamic particle size distribution. AAPS J [Internet]. 2013;15(2):465–476. Available from: https://doi.org/10.1208/s12248-013-9453-y
4. Weber B, Lee SL, Delvadia R, Lionberger R, Li BV, Tsong Y, et al. Application of the modified chi-square ratio statistic in a stepwise procedure for cascade impactor equivalence testing. AAPS J [Internet]. 2015;17(2):370–9 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4365081/. Accessed 11 June 2019
5. Nahar K, Gupta N, Gauvin R, Absar S, Patel B, Gupta V, Khademhosseini A., Ahsan F. In vitro, in vivo and ex vivo models for studying particle deposition and drug absorption of inhaled pharmaceuticals. Eur J Pharm Sci [Internet]. 2013;49(5):805–818. Available from: https://doi.org/10.1016/j.ejps.2013.06.004
6. United States Food and Drug Administration. Guidance on Fluticasone Propionate; Salmeterol Xinafoate. US Food Drug Adm [Internet]. 2016;2–8. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm367643.pdf.. Accessed 8 Mar 2018

7. FDA. Draft guidance on budesonide. 2012.
8. European Medicines Agency. Guideline on the requirements for clinical documentation for orally inhaled products (Oip) including the requirements for demonstration of therapeutic equivalence between two inhaled products for use in the treatment of asthma and chronic obstructive pulm Pdf [Internet]. 2009;(August):1–26. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003504.pdf. Accessed 11 June 2019
9. FDA. Guidance for industry studies for nasal aerosols and guidance for industry bioavailability and bioequivalence. Vol. 1999;5651:1–40.
10. Christopher D, Adams W, Amann A, Bertha C, Byron PR, Doub W, et al. Product quality research institute evaluation of cascade impactor profiles of pharmaceutical aerosols, part 3: final report on a statistical procedure for determining equivalence. AAPS PharmSciTech [Internet] 2007;8(4):E1–10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2750676/. Accessed 11 June 2019
11. Morgan B, Strickland H. Performance properties of the population bioequivalence approach for in vitro delivered dose for orally inhaled respiratory products. AAPS J [Internet]. 2014;16(1):89–100 Available from: http://www.ncbi.nlm.nih.gov/pubmed/24249218. Accessed 11 June 2019
12. Nilceia Lopes, Katherine Ruas, Cristina Helena dos Reis Serra VP. Average, population and individual bioequivalence. SapJ [Internet] 2010;77(6):46–48. Available from: http://www.sapj.co.za/index.php/SAPJ/article/view/539. Accessed 11 June 2019
13. Sandell D, Mitchell JP. Considerations for designing in vitro bioequivalence (IVBE) studies for pressurized metered dose inhalers (pMDIs) with spacer or valved holding chamber (S/VHC) add-on devices. J Aerosol Med Pulm Drug Deliv [Internet]. 2014;27(0):1–26 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25089555. Accessed 11 June 2019
14. Faraway JJ. Binary Response. In: Extending the linear model with R: generalized linear, mixed effects and non-parametric regression models, second edition. 2016. p. 25–50.
15. Shein-Chung C, Jun S. Medical Imaging. In: Statistics in drug research : methodologies and recent developments. 2002. p. 316–326.
16. Ann A, George SI. Screening in public health practice. In: epidemiology in public health; 2014. p. 417–46.
17. Pharmacokinetic Comparison of Locally Acting Dry Powder Inhalers, G. Hochhaus and J. Bulitta, DIA Workshop 2018, Generic Drug-Device Combination Complex Products. Silver Spring. Maryland: PK details to be published elsewhere; October 2018.
18. Evans C, Cipolla D, Chesworth T, Agurell E, Ahrens R, Conner D, et al. Equivalence considerations for orally inhaled products for local action—ISAM/IPAC-RS European workshop report. J Aerosol Med Pulm Drug Deliv. 2012;25(3):117–39.
19. Delong ER, Carolina N. Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach author (s): Elizabeth R . DeLong , David M . DeLong and Daniel L . Clarke-Pearson Published by : International Biometric Society Stable. Biometrics. 1988;44(3):837–45.
20. Park SH, Goo JM, Jo C-H. Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol [Internet]. 2004;5(1):11. Available from: https://kjronline.org/DOIx.php?id=10.3348/kjr.2004.5.1.11. Accessed 11 June 2019
21. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005;38(5):404–15.
22. FDA. Draft Guidance for industry: bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action.
23. Pan Z, Christopher J, Lyapustina S, Chou E. Statistical techniques used in simulation of cascade impactor particle size distribution profiles. Respir Drug Deliv IX. 2004;3:669–72.
24. Sandell D. Review of the EMEA guidelines' in-vitro equivalence criteria for Cascade impaction data [internet]. 2010 [cited 2019 Apr 30]. Available from: https://ipacrs.org/assets/uploads/outputs/Sandell.pdf. Accessed 11 June 2019
25. Newman SP, Chan H-K. In vitro/in vivo comparisons in pulmonary drug delivery. J Aerosol Med Pulm Drug Deliv. 2008;21:77–84.