The AAPS Journal (2020) 22:139
DOI: 10.1208/s12248-020-00520-6

Research Article

Simulation Informed Design and Performance of *In Vitro* Bioequivalence Trials for Particle Size Distributions

William J. Ganley,[1,2] Jagdeep Shur,[1] and Robert Price[1]

Received 30 June 2020; accepted 6 October 2020

Abstract. This study used statistical simulations to investigate the performance of the population bioequivalence test applied to image-based particle size measurements (such as morphologically directed Raman spectroscopy) and methods for designing *in vitro* bioequivalence trials using prior information. Simulations of *in vitro* population bioequivalence trials were conducted across a range of representative D_{50} (number-weighted median particle diameter from a log-normal particle size distribution) and span (which is defined as $\frac{D_{90}-D_{10}}{D_{50}}$ where D_{90} and D_{10} are the number-weighted 90th and 10th percentiles in particle diameters sampled from a log-normal particle size distribution) values respectively. The performance of the population bioequivalence test in the simulations was driven by an interplay between overall test variability and the widening or narrowing of the bioequivalence region due to variance terms in the test statistic definition. These findings were dependent upon differences in the variability of D_{50} and span and may generalise to a wider range of *in vitro* metrics. Trial design optimisation using power and assurance approaches followed patterns consistent with these findings. As more novel scientific methods are applied to the development of complex generic drug products, the procedures outlined in this study may be used at the inception stage of future *in vitro* bioequivalence trials to reduce the risk of conducting costly trials with low probabilities of success.

KEY WORDS: orally inhaled and nasal drug products; particle size distribution; population bioequivalence; simulation.

INTRODUCTION

The determination of bioequivalence between complex locally acting drug products is challenging and continues to hinder generic entry to the market for many orally inhaled and nasal drug products (OINDPs). Bioequivalence is defined as equivalence in the rate and extent at which the active pharmaceutical ingredient (API) becomes available at the site of action (1). For locally acting drug products, such as OINDPs, bioequivalence is currently demonstrated through a combination of *in vitro*, pharmacokinetic, and comparative clinical endpoint tests (2,3). In 2017, FDA pledged to reduce the "hurdles" for generic product development (4) which has involved the publication of several revised product-specific guidance (PSG) documents for OINDPs, some of which include alternatives to comparative clinical endpoint tests which could reduce the testing burden on generic drug product development whilst maintaining accuracy in bioequivalence determinations. In 2016, measurements of API particle size distribution by morphologically directed Raman spectroscopy (MDRS) provided a measure of *in vitro bioequivalence* in the approval of a generic mometasone furoate nasal spray for marketing by FDA where *in vitro* bioequivalence eliminated the need for a comparative clinical endpoint study (5).

Any measure of bioequivalence, be it *in vivo* such as maximum plasma concentration or *in vitro* such as median particle diameter, must be accompanied by a statistical approach which can confirm equivalence with a controlled error rate. The FDA guidance Statistical Approaches to Establishing Bioequivalence (6) details average bioequivalence (ABE, which considers only the difference in test and reference product means), population bioequivalence (PBE, which considers the difference in test and reference product means and the between-subject or between-unit variances), and individual bioequivalence (IBE, which considers the difference in test and reference product means and the within-subject or within-unit variances). Multiple comparisons of ABE and PBE have been conducted (7–9) generally showing that the tests perform similarly apart from when

Electronic supplementary material The online version of this article (https://doi.org/10.1208/s12248-020-00520-6) contains supplementary material, which is available to authorized users.

[1] Nanopharm Ltd, an Aptar Pharma Company, Cavendish House, Hazell Drive, Newport, NP10 8FY, UK.
[2] To whom correspondence should be addressed. (e–mail: w.ganley@nanopharm.co.uk)

product variances are relatively large as this dominates the PBE result. PBE is recommended in the Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action (10), and multiple PSGs for metered dose inhalers and dry powder inhalers (11,12) so will be the focus of this study.

It was reported almost a decade ago that the design of *in vitro* bioequivalence trials should be tailored to the predicted product properties rather than prescribed (13). More recently a series of thorough investigations into the performance of the PBE test, including the analyses of large databases containing delivered dose and impactor stage mass data from many inhaled products (8,14,15), have uncovered a number of key properties. The first is that OINDP performance metrics such as impactor stage mass (15) and delivered dose (14,15) are normally distributed which breaks the assumption of log-normality in the PBE test. The consequence of performing the PBE test on normally distributed data is an asymmetry in PBE power where test and reference product pairs for which the test product mean is higher than the reference product mean have a higher probability of determining equivalence than the opposing case. The addition of between-batch (8) variance and within-container variance (arising from replicate measurements) (13) to the calculation method suggested in the FDA PSG on Budesonide (16) has also been investigated. It was shown that including both between- and within-batch variance in the PBE procedure increased the true positive and decreased the false negative rates when between-batch variability was present (8); however, the authors were keen to state that this did not address the issue of the asymmetry in the PBE test when applied to normally distributed measures (8). The application of PBE has recently been extended to other metrics such as distance measures for the comparison of non-monomodal size distributions (17).

Following the precedent set by the approval of a generic mometasone furoate (5), it is expected that future similar generic approvals will be based in some part on a wider range of *in vitro* techniques. Techniques that could emerge are included in the "Alternative approach to the comparative clinical endpoint BE study" section of multiple recently revised FDA PSG documents (11,12,18). It is pertinent that statistical tests for bioequivalence applied to these novel methods are effective, well understood, and appropriately applied. Once adopted, bioequivalence trials using novel performance metrics will require careful design to ensure an adequate chance of determining BE when it is true and avoiding costly underpowered trials. A number of PSGs for OINDPs quote a minimum of 3 batches and 10 containers per product per trial (12,16); however, designs that are case specific can be more effective (13).

In this paper, we use statistical simulations to explore the performance and design of PBE trials comparing particle size distributions using median particle size (D_{50}) and span metrics which are the typical outcomes of API-specific particle size distributions measured using MDRS and were included in the approval of a generic mometasone furoate nasal spray (5). The simulation approach can be generalised to other *in vitro* metrics and builds on previous work (19) to design trials from prior knowledge where uncertainty in product performance is considered.

METHODS

Population Bioequivalence Test

The PBE test is defined as (16):

$$\frac{(\mu_T-\mu_R)^2+(\sigma_T-\sigma_R)^2}{\max\left(\sigma_R^2,\sigma_{T0}^2\right)}\le\theta \tag{1}$$

or in the linearised form:

$$(\mu_T-\mu_R)^2+\left(\sigma_T^2-\sigma_R^2\right)-\theta\max\left(\sigma_R^2,\sigma_{T0}^2\right)<0 \tag{2}$$

where μ_T and μ_R are the means of the log-transformed measured test and reference variables respectively, σ_T and σ_R are their corresponding standard deviations, $\sigma_{T0}=0.1$, and θ is the PBE criterion defined as 2.089 which approximately captures $\mu_T=90\%$ and $\mu_R=100\%$, $\sigma_T^2=0.02$ and $\sigma_R^2=0.01$. A test and reference product is deemed bioequivalent in a PBE trial when the 95% confidence interval of the linearised criterion (calculated as shown in reference (16)) is below 0.

Statistical Model

The *in vitro* measurements investigated in this study were the number-weighted median particle diameter from a log-normal particle size distribution (D_{50}) and the span which is defined as $\frac{D_{90}-D_{10}}{D_{50}}$ where D_{90} and D_{10} are the number-weighted 90th and 10th percentiles in particle diameters sampled from a log-normal particle size distribution. Individual measurements were simulated by assuming that one measurement samples individual diameters from a population log-normal size distribution with characteristic geometric mean (μ_{D50}) and geometric standard deviation (defined as $\log(\Sigma)$) which are both subject to batch, container, lifestage, and unexplained variation.

Sampling distributions for the D_{50} and span metrics were used in the model for computational efficiency, rather than generating individual particle diameters from the parent population. The D_{50} sampling distribution was estimated as shown below (20). The span sampling distribution was estimated by fitting polynomial expressions to empirical values obtained by summarising randomly generated samples of diameters and is detailed in the supplementary information.

In each simulated trial, a D_{50} or span was generated for each product, batch, container, and lifestage combination using the following multi-level model:

$$D_{50}\sim N\left(\mu_{D50},\sigma_{D50}^2\right) \tag{3}$$

$$\mu_{D50}=M_i+B_{M,j}+C_{M,k(j)}+L_{M,ijkl}+\epsilon_{M,ijkl} \tag{4}$$

$$\sigma_{D50}=\frac{1}{4nf_{LN}\left(M,\Sigma^2\right)^2} \tag{5}$$

$$\text{span} \sim LN\left(\mu_{\text{span}}(n,\Sigma), \sigma^2_{\text{span}}(n,\Sigma)\right) \tag{6}$$

$$\log(\Sigma) = \log(\Sigma_i) + B_{\Sigma,j} + C_{\Sigma,k(j)} + L_{\Sigma,ijkl} + \epsilon_{\Sigma,ijk} \tag{7}$$

where n is the number of particles sampled, f_{LN} is the log-normal probability density function, M_i is the logarithm of the average geometric mean of product i (test or reference), B_{M} is the normally distributed mean zero batch effect, C_{M} is the normally distributed zero mean container (nested in batch) effect, L_{M} is the normally distributed zero mean lifestage effect, and ϵ_{M} is the unexplained variation which is normally distributed with mean zero. Σ is the logarithm of the geometric standard deviation of the particle size distribution, Σ_i is the average effect of product i (test or reference), and B_Σ, C_Σ, L_Σ, and ϵ_Σ are defined as for M. $ijkl$ refers to the ith product, jth batch, kth container, and lth lifestage.

Simulation Strategy

PBE Test Performance

Multiple scenarios were simulated capturing the sensitivity of the PBE test performance to different aspects of the true product characteristics. The impact of total variance, variance distribution (amongst batch effect, container effect, and unexplained variance), lifestage effects, and number of particles were studied.

Unless otherwise specified, the median particle diameter was fixed at 2.5 μm and the geometric standard deviation was fixed at 1.5, the total variance was defined by fixing the relative standard deviation (RSD) at 12% for the M and 4% for $\log(\Sigma)$, the distribution of the total variance amongst batches, containers, and unexplained variation was 0.2, 0.4, and 0.4 respectively, and no lifestage effect was assumed.

In each simulated trial, 5000 sets of test and reference D_{50} and span values were generated using the model described by Eqs. 3–7 and tested for PBE using the method described in reference (16). The fraction of passing trials were then reported for each combination of parameters. For example, in a trial with 3 batches, 10 containers, and 3 lifestages, 90 D_{50} values would be generated for the test product and 90 for the reference product, these values would be tested for population bioequivalence, and the process repeated a total of 5000 times. Unless otherwise specified, the number of containers was fixed at 10, the number of lifestages was fixed at 3, and the number of batches was fixed at 3.

Power and Assurance Calculations

Power and assurance were calculated to represent the progression from a small-scale feasibility study to a full *in vitro* bioequivalence trial. First, a small set of test and reference product D_{50} data were simulated from the model described by Eqs. 3–7. The test product average D_{50} was 2.55 μm (102% of the reference average which for identical variance is within definition of population bioequivalence), and the distribution of the total variance amongst batches, containers, and unexplained variation was 0.4, 0.4, and 0.2 respectively. All other parameters were fixed as described above. Two batches, 2 containers, and 3 lifestages were simulated for a total of 12 simulated measurements per product.

The simulated feasibility scale data were then fitted to a series of simplified models, capturing either the product effect alone, the product effect alongside the batch effects, or the product effect, batch effect, and container effect. Lifestage effects were not considered. The simulated data was fitted by Bayesian inference using Stan 2.19.1 (21) via the R package rstan 2.19.3 (22) and the convenience functions in the rethinking 2.0 package (23) and the following prior distributions describe the most detailed model (other models were constructed by removing terms).

$$D_{50} \sim N\left(\mu_{D50}, \sigma^2_{D50}\right) \tag{8}$$

$$\mu_{D50} = M + B_i + C_j \tag{9}$$

$$M \sim N\left(3, 1.2^2\right) \tag{10}$$

$$B_i \sim N\left(0, \sigma^2_{\text{B}}\right) \tag{11}$$

$$C_j \sim N\left(0, \sigma^2_{\text{C}}\right) \tag{12}$$

$$\sigma_{D50}, \sigma_{\text{B}}, \sigma_{\text{C}} \sim \text{Exponential}(1) \tag{13}$$

where D_{50} is the observed variable, M is the product mean of the observed variable, B_i and C_j are the ith and jth batch and container effects respectively, σ_{D50} is the residual standard deviation in the observed variable, and σ_{B} and σ_{c} are the standard deviations of the batch effect and container effects respectively.

The models were compared using the widely applicable information criterion (WAIC), and a model was selected for the test and reference products separately. Power was calculated by simulating pairs of test and reference datasets from the selected model where the values of M, σ_{D50}, σ_B, and σ_C were fixed at the means of their posterior distributions. A total of 10,000 datasets were simulated for 2–20 batches, 2–20 containers, and 1–3 lifestages, the PBE test performed, and the power for each sampling configuration was equal to the fraction of passing trials. The assurance was calculated in a

similar manner; M, σ_{D50}, σ_B, and σ_C were treated as random variables and sampled from their posterior distributions at each simulation run, therefore capturing uncertainty in the parameters. It should be noted that the large variance in these simulations resulted in some simulated datasets producing negative diameters which were rejected. The minimum number of runs for a single configuration was 8048 for assurance and 9980 for power.

All data were simulated using R 3.6.1 (24) and visualised using ggplot2 3.2.1 (25), and metR 0.7.0 (26) was used to add labels to the contour plots.

RESULTS

Statistical simulations following the method described above were conducted for a variety of parameter combinations. In each case, the reference product D_{50} and GSD were fixed at 2.5 μm and 1.5 respectively. The test product D_{50} varied from 0.6 to 1.4 times the reference product value and the test product GSD from 0.8 to 1.2 times the reference product value.

Figure 1 shows an example D_{50} simulation result for RSDs of 10% and 14%. The simulation results show a previously reported asymmetry in the fraction of passing trials about a test/reference ratio of 1 (8,14) where ratios below this have a lower fraction of passing trials than the equivalent ratios greater than 1. To concisely summarise the results of multiple simulations, the average power (AP) was calculated for each, which is the mean of the fraction of passing trials in the bioequivalence region across all test product D_{50} or span values simulated (the points within the corresponding boxes in Fig. 1).

Total Product Variance

The AP for simulations of different test/reference RSD pairs and different numbers of batches for both D_{50} and span are shown in Fig. 2. Increasing the RSD reduces the AP for both metrics. In cases where the test and reference RSDs are equal, an increase in the total number of batches tested can counteract the low AP for variable products. For asymmetric test and reference RSDs (the rightmost two points in Fig. 2), the power increases where the test product RSD is greater than that of the reference product and the reverse is true for the opposing case.

Product Variance Pattern

The distribution of the variance amongst batch effects, container effects, and unexplained variance also impacts the results of PBE tests. The AP for 4 different variance patterns are shown in Fig. 3. Note that where the batch and container fractions do not total 1.0, the remaining variance is allocated to the unexplained fraction; for example, the leftmost points have patterns of 0.1 batch, 0.1 container, and 0.8 unexplained variance. Figure 3 shows that the D_{50} AP decreases as the unexplained variance contribution decreases (from left to right) and for the central two points, where the unexplained variance contribution is equal, the AP is lower for a higher contribution of between-batch variance. Similarly, to the AP results for total variance (shown in Fig. 2), increasing the

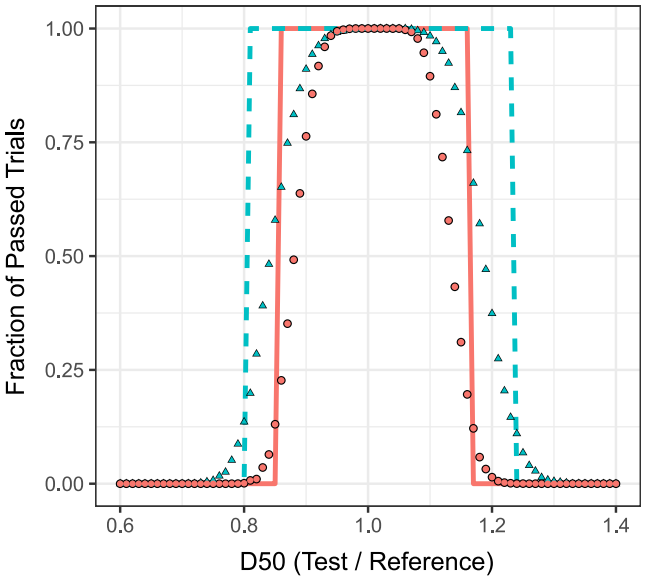

Fig. 1. Fraction of passing simulated D_{50} trials for 9 batches, 10 containers, and 3 lifestages. Points show simulation results and lines show the bioequivalence region. Total RSDs for both test and reference products were fixed at 10% (circles and solid line) and 14% (triangles and dashed line). All other parameters were fixed as detailed in the "METHODS" section

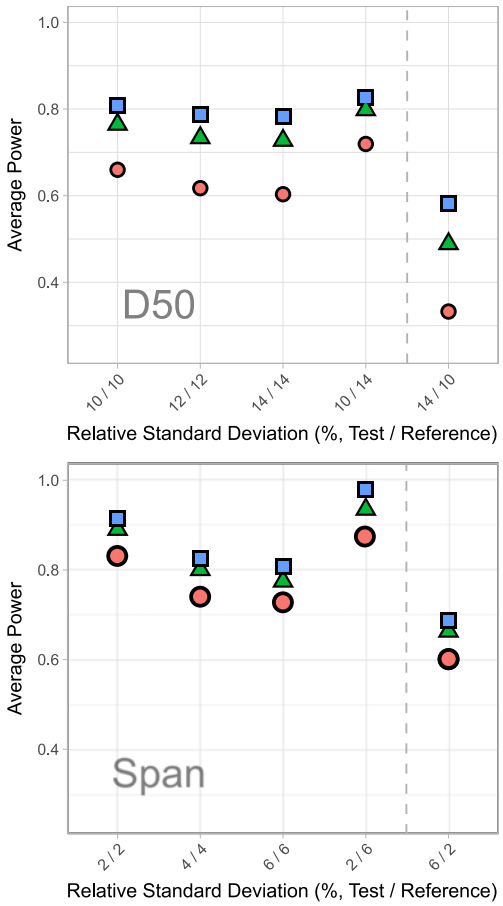

Fig. 2. Average power of PBE trial simulations for different test and reference product RSDs. 3 (circles), 6 (triangles), and 9 (squares) batches

number of batches tested can counteract low AP due to unfavourable variance patterns. The low AP for 0.4 batch and 0.4 container variance contributions increases to a level higher than the 3 batch, 0.1 batch, and 0.1 container variance contribution AP by increasing the number of batches tested from 3 to 6. The AP of span is insensitive to variance pattern over the range of variance contributions tested.

Lifestage Effect

Lifestage effects have been observed in measures commonly used in *in vitro* OINDP bioequivalence testing (8,15). It is important to measure products across their entire lives to eliminate biases in test/reference ratios and product variance estimations which can affect PBE test results.

The three lifestage regimes investigated were contained a beginning, middle, and end stage and were captured as systematic shifts in M and $\log(\Sigma)$ shown in Eqs. 4 and 7. The beginning, middle, and end of life regimes tested were [0%, 0%, 0%], [− 5%, 0%, 5%], and [− 10%, 0%, 10%] and will be referred to as the 0%, 5%, and 10% lifestage effects respectively. The 0% and 5% regimes are small and of similar magnitude to those investigated in previous similar studies (8,15). The 10% regime is exceptionally large and was included to test the sensitivity of the PBE test to large lifestage effects.

The AP of both D_{50} and span shown in Fig. 4 show little change for the 0% and 5% regimes and, as for other parameters investigated, are less significant than the impact of increasing the number of batches tested. For D_{50}, the 10% regime results in a small increase in AP.

Number of Particles

The final parameter investigated in the simulation study was the number of particles tested in each trial which is related to the variance of the D_{50} and span sampling distributions through Eqs. 5 and 6 and is distinct from the RSD and the product variance patterns which impact M and $\log(\Sigma)$ through Eqs. 4 and 7. The AP for D_{50} reduced marginally as the number of particles increased. The asymmetric samples (which would only result from poor experimental control or deliberate biasing) produce the same behaviour as observed for total variance in Fig. 2 where a low test product n and high reference product n result in high test and low reference product variances respectively. The opposite is true when the test product n is high and reference product n is low. The lower plot in Fig. 5 shows that span AP increases with increasing number of particles sampled, decreases for when test product n is much lower than reference product n, and is consistently close to zero (off axis scale) for the opposite case.

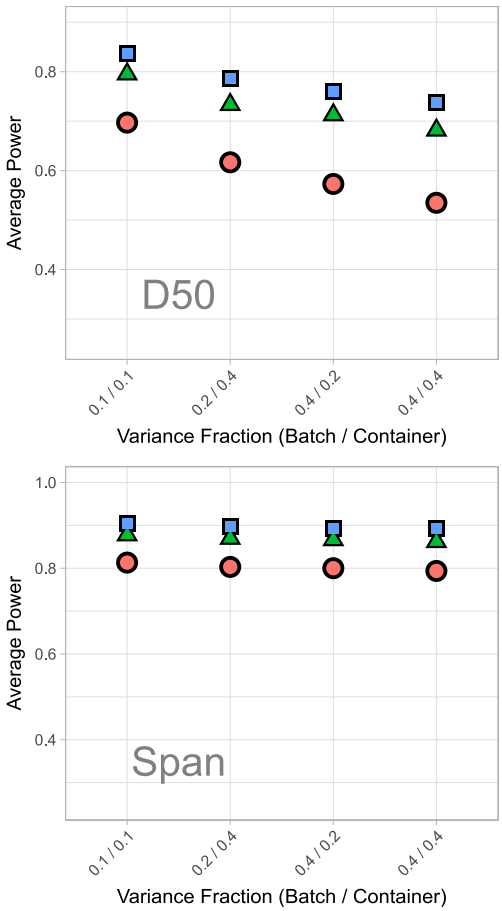

Fig. 3. Average power of PBE trial simulations for different test and reference product variance patterns. Horizontal axis labels show batch and container effect variance fractions; unexplained variance comprises the remainder. 3 (circles), 6 (triangles), and 9 (squares) batches

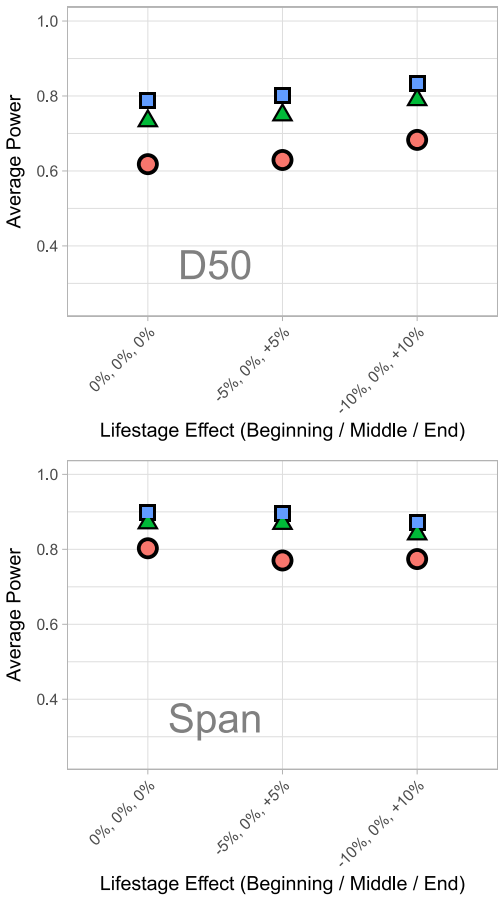

Fig. 4. Average power of PBE trial simulations for different test and reference product lifestage effects. 3 (circles), 6 (triangles), and 9 (squares) batches

Power and Assurance

The description of AP of a statistical test can develop intuition in the design of trials; however in real cases, the parameters discussed above are not known with certainty. This section describes methods for incorporating prior knowledge into *in vitro* bioequivalence trial design. The term design here refers to the number of batches, containers, and lifestages to be tested in an *in vitro* bioequivalence trial.

Current FDA guidance for *in vitro* bioequivalence testing suggests a trial design of 3 batches, 10 containers, and 3 lifestages per product (16). This design is expected to sufficiently sample the variance of a typical product where the main source of variability comes from within a batch (27). With an increasing number of *in vitro* metrics being used to show bioequivalence, it is important to explore how small deviations from the well-established design might help or hinder the probability of success of a trial. Such methods are explored in this section.

To represent the process of designing a full PBE trial from a feasibility scale trial, a small set of test and reference product data were simulated using the model described by Eqs. 3–8 where the test product mean D_{50} was equal to 102% of the reference product value (other parameter values are detailed in the "METHODS" section). This simulated dataset is shown in Table I and represents a pilot or feasibility scale study.

Bayesian inference was used to estimate the test and reference product parameters used in the power calculations.

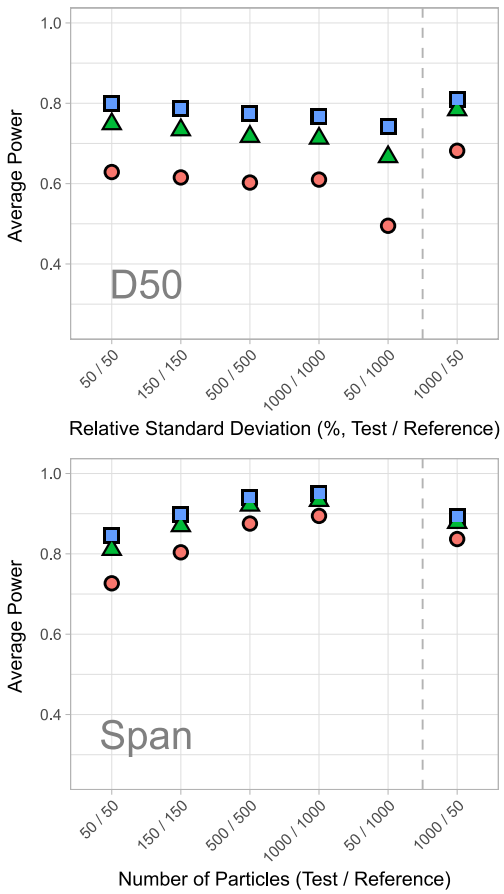

Fig. 5. Average power of PBE trial simulations for different numbers of particles. 3 (circles), 6 (triangles), and 9 (squares) batches

Three different hierarchical model structures (detailed in the "METHODS" section) were compared as shown in Table II. The test product data was best described by a model including batch and container effects whereas the reference product data was best described by a model including just only batch effects. The obtained parameter estimates for the final models are shown in Table III.

The power of different trial designs calculated with the model parameters shown in Table III is summarised as a contour plot in Fig. 6 where the red point indicates the design shown in the FDA PSG for Budesonide example calculations (16) of 3 batches, 10 containers, and 3 lifestages, here giving a power of 0.510. The power was then used to explore the impact of small adjustments to the design on the probability of success of the trial. For example, the gradient is steeper parallel to the batches' axis than the containers' axis. Incrementing the number of containers by 1 leads to a power of 0.508 which is a negligible change. Incrementing the number of batches by 1 leads to a more substantial increase in power to 0.596. Increasing the number of containers per batch to 14 (maintaining a similar number of measurements to the 4-batch, 10-container design) resulted in a power of 0.531. Such changes in power can be balanced against the costs of obtaining further batches or units, and those that would change the outcome negligibly, such as increasing the number of containers by 1, can be discounted.

The unconditional analogue of power is known as assurance and is defined in Eqs. 14–16 (notation as in reference (28)).

Table I. Simulated Small-scale Feasibility Dataset Used for Power and Assurance Calculations. R and T Represent Reference and Test Products Respectively

Product/batch/container/lifestage identifier	D_{50} (μm)
R/1/1/1	2.7570
R/1/1/2	2.7342
R/1/1/3	2.7895
R/1/2/1	2.5799
R/1/2/2	2.9341
R/1/2/3	2.6412
R/2/3/1	2.6875
R/2/3/2	2.6981
R/2/3/3	2.2010
R/2/4/1	2.4813
R/2/4/2	2.4267
R/2/4/3	2.8144
T/1/1/1	2.3873
T/1/1/2	2.6598
T/1/1/3	2.7659
T/1/2/1	2.8045
T/1/2/2	2.9116
T/1/2/3	2.9156
T/2/3/1	2.5507
T/2/3/2	3.0620
T/2/3/3	2.9092
T/2/4/1	2.3225
T/2/4/2	2.3766
T/2/4/3	2.6106

Table II. Model Comparisons for Simulated Test and Reference Products. Asterisks Denote Selected Models

Model	Test product WAIC	Reference product WAIC
All effects	0.2*	−0.8
Product and batch effects	4.4	−2.7*
Only product effect	2.8	−1.5

$$\pi(\theta) = P(R|\theta) \tag{14}$$

$$\gamma(\theta) = E(\pi(\theta)) \tag{15}$$

$$E(\pi(\theta)) = \int P(R|\theta)P(\theta)\mathrm{d}\theta \tag{16}$$

where π is the power, R is the trial outcome, θ is a vector of assumed parameter values, γ is the assurance function, E is the expected value, and $P(\theta)$ is a probability density function describing the prior knowledge of θ. In this study, the posterior distributions summarised in Table III were used as $P(\theta)$. The calculated assurance curves are shown as a contour plot in Fig. 7.

Upon visual inspection, Fig. 7 shows a much steeper rise in the regions of lower containers and batches, diminishing returns as the sampling requirement increases and a plateau assurance of around 0.6. The design shown in the FDA PSG for Budesonide example calculations (16) of 3 batches, 10 containers, and 3 lifestages gives an assurance of 0.403. Incrementing the number of containers by 1 gives an assurance of 0.419, and incrementing the number of batches by 1 gives an assurance of 0.465. Increasing the number of containers to 14 gives an assurance of 0.428. Inspection of the assurance of these designs gave a similar result to that of the power. Increasing the number of batches raised the assurance more than increasing the number of containers, even when the total number of measurements was held constant.

In a practical setting, the results of this and the previous section can be utilised to set expectations and guide necessary adjustments in trial design within reasonable limits. For example, if the analysis of a feasibility study suggests that the test product has a much higher variance than the reference product, then the results of Fig. 2 show that the expected probability of success will be diminished. More subtle is the insight that a substantially lower residual variance component can hinder the probability of success in trials of a highly variable parameter such as D_{50} in this study (see Fig. 3). In that case, different study designs can be explored using the power and assurance approaches, optimising adjustments made (if possible), and expectations of the trial outcomes can be set based on prior knowledge of the products being tested (the feasibility trial) and the known behaviour of the PBE approach for the given measure.

DISCUSSION

The previous section highlighted two main effects that impact the performance of PBE trials of inherently high variance (D_{50} in this study) and low variance (span in this study) metrics. The first was variability in the PBE test results due to sampling variance. For higher variability products, the distribution of PBE test results is expected to be wider and may result in a greater proportion of false negative outcomes. The second factor is the width of the bioequivalence region which is dictated by the $\left(\sigma_T^2 - \sigma_R^2\right)$ and max $\left(\sigma_R^2, \sigma_0^2\right)$ terms in the definition of the PBE test statistics (Eqs. 1 and 2) and shown by the lines in Fig. 1. By the definition of the test statistic, a higher reference product variance widens the range of test/reference mean ratios that fall under the definition of population bioequivalence and a higher test product variance reduces the range.

The interplay between the widening of the bioequivalence region and the variability in test results was observed across the different variance patterns shown in Fig. 3. The overall test and reference product variances, and therefore the width of the bioequivalence region, in these simulations were fixed, yet the AP of D_{50} and span behaved differently. The span AP did not change across the different variance patterns suggesting that the PBE test result was dominated by the width of the bioequivalence region and change variability of the test result between the variance patterns was insignificant. The D_{50} AP decreased as residual variance fraction decreased. The width of the bioequivalence region is constant; therefore, decreasing residual variance must have increased the variability of the test result. This could be explained by a larger impact of outlying batch or container effects in biasing the results of individual tests which would be reduced if a greater fraction of the variance was residual. Additionally it has been reported that the absence of a batch effect in the method used to conduct the PBE test described in the FDA PSG for Budesonide (16) can result in small decreases in average power and increases in false equivalence rates with increasing batch effects (8). The conclusions from the study detailed in reference (8) were that incorporation of

Table III. Posterior Parameter Means and Standard Deviations (Shown in Brackets) for Simulated Test and Reference Product data

Parameter	Test product	Reference product
M	1.03 (0.12)	1.01 (0.10)
σ_{D50}	0.22 (0.07)	0.20 (0.05)
σ_B	0.22 (0.31)	0.21 (0.28)
σ_C	0.14 (0.13)	-

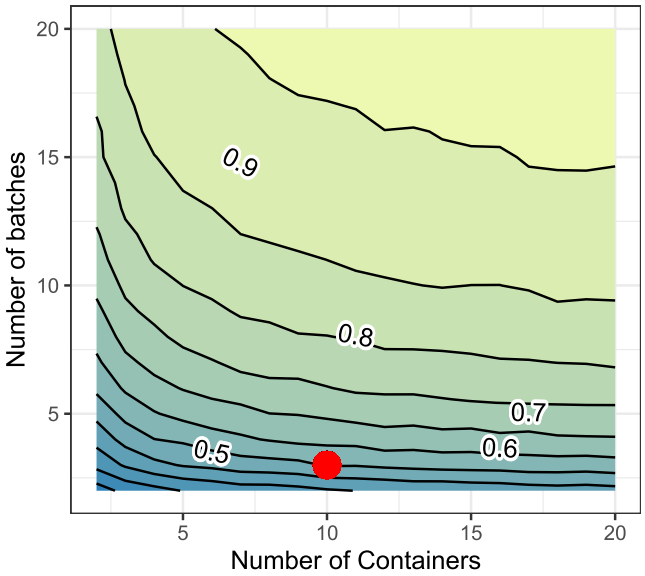

Fig. 6. Predicted power of the PBE test from statistical models calculated using simulated feasibility scale trials data. Power for 3 lifestages is shown. Point shows 3 batches and 10 containers

the batch effect and a larger number of tested batches are required to improve PBE testing. The effect of testing a larger number of batches is shown in Fig. 3 to offset the drop in AP due to greater between-batch variability. The trial design methods shown in "Product Variance Pattern" suggest that many more batches than the minimum number suggested in the FDA PSG for Budesonide (16) may be required to achieve trials with sufficient probabilities of success. For example, the assurance increased from 0.403 to 0.465 upon increasing the number of batches by 1 and only increased to 0.428 upon increasing the number of containers to 14, maintaining a similar number of total measurements as the addition of a single batch.

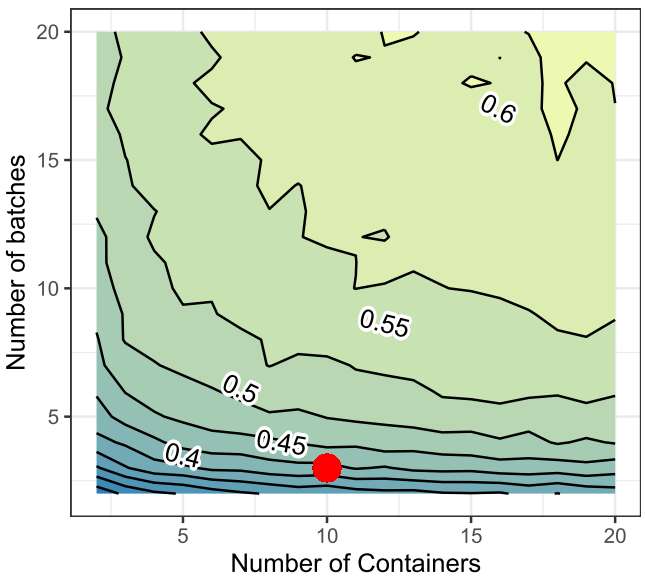

Fig. 7. Predicted assurance of the PBE test from statistical models calculated using simulated feasibility scale trials data. Assurance for 3 lifestages is shown. Point shows 3 batches and 10 containers

The response of AP to changes in the number of particles tested (n) also highlighted the interplay between the width of the bioequivalence region and variability in PBE test results. D_{50} AP reduced slightly with increasing n as the reference product variance decreased, narrowing the bioequivalence region, and must therefore have not been counteracted by a fall in PBE test result variability. The span AP showed the opposite effect, increasing with increased n due to a fall in the PBE test result variability that was more significant than the narrowing of the bioequivalence region. For the asymmetric n simulations, the span AP was close to zero for all test/reference mean ratios examined when the test product n was much lower than the reference product n due to very large $\left(\sigma_T^2-\sigma_R^2\right)$ value, penalising the result. It should be noted that the asymmetric n calculations were merely illustrative and well-controlled trials should always analyse similar numbers of particles.

The results described in the initial phase of the simulation study provided a general insight of the performance of the PBE test under different circumstances. The section that followed investigated the design of *in vitro* bioequivalence trials using statistical power and assurance. Power and assurance both require some prior information of the products to be tested. Power is conditional on exact quantities, which are not known at the time of calculation and must be assumed, and assurance treats the quantities as random variables accounting for their uncertainty. Power and assurance were calculated based on results from a simulated feasibility scale trial, and the results differed in two main ways. The first difference is that assurance does not tend towards certainty even at impractically large numbers of total measurements. The plateau assurance reflects the probability distribution of the product parameters ($P(\theta)$ is Eq. 16), which in this case is around 0.6 as it is not known with certainty whether the products are truly bioequivalent. The second difference is that the gradients of the assurance curves diminish at smaller numbers of batches and containers than those in the power curves which can be seen as wider gaps between contour lines in Fig. 7 than in Fig. 6 above assurance values of around 0.5. This is a well-documented behaviour of assurance when compared to power in the context of clinical trials (29) and can lead to counterintuitive results, such as assurance values of 0.5 regardless of trial design if prior variance is high (30). The results of assurance calculations should therefore be communicated with care. It is notable in this example that in the region of practical trial designs, centred around 3 batches, 10 containers, and 3 lifestages, the assurance is still far from the plateau region meaning that increases in the scale of the trial will appreciably improve the probability of success of the trial.

In a practical setting power and assurance, calculations can be used to explore the impacts of possible changes from the minimum trial design recommended in many FDA PSGs (16) and can be used to decide on whether any increase in the probability of success of the trial justifies the additional cost. Deviations from the 3-batch, 10-container, 3-lifestage design in both power and assurance (shown as red points in Figs. 6 and 7) show steeper gradients for increases in the number of batches than the number of containers. The procurement of additional batches for a trial may be impractical; therefore, the impact of increasing the number of containers or

lifestages can be assessed. For example, fixing the number of batches and containers at 3 and 10 respectively and changing the number of lifestages from 1 to 3 give an increase in assurance from 0.348 to 0.392 to 0.403. Increasing the number of containers from the FDA-suggested design to 14 increases the assurance from 0.403 to 0.428 whilst increasing the total number of measurements from 90 to 126 (where adding a 4th batch would require a total of 120 measurements). It would therefore be up to those designing the trial to use prior knowledge of the products alongside calculations of power and assurance to determine whether the minimum trial design suggested by the FDA results in a sufficient probability of success and what possible cost- and resource-efficient adjustments can be made to maximise the chance of correctly determining bioequivalence.

The results of the trial design simulations are consistent with the simulations of different overall variance and variance patterns shown in Figs. 2 and 3 where increasing total variance and increasing fractional batch effect reduced the AP. Both figures show that low AP resulting from the inherent variance in the trialled products and metrics can be counteracted by increasing the number of total measurements.

One variable that was not explored in the trial design section of this work was the effect of the number of sampled particles which is unique to image-based particle size measurements. The simulations at different sample sizes shown in Fig. 5 suggest that the effect of the number of particles on high variance parameters is negligible, provided that it is balanced between test and reference products, and is an avenue for further study of bioequivalence tests using related measurements.

CONCLUSIONS

This study has shown the use of statistical simulations to investigate the performance and design of *in vitro* population bioequivalence trials of two metrics commonly used to summarise particle size distributions from image-based methods such as morphologically directed Raman spectroscopy. The average power of the population bioequivalence test for the higher variability D_{50} was sensitive to the total variance, variance pattern, and lifestage effect but not the number of particles tested. The average power of test for the lower variability span was sensitive to the total variance, and the total number of particles tested but not the variance pattern or lifestage effects. These differences were due to the interaction between variation in PBE test results and the width of the bioequivalence region determined by the test and reference product variances. This finding may generalise to other metrics tested using the PBE approach as more *in vitro* techniques are used as evidence for bioequivalence of complex drug products.

A trial design method was also investigated where the statistical power (which is conditional on unknown product characteristics) and assurance (which can be thought of as the unconditional expected power) were estimated from simulated feasibility scale data. The results of both the power and assurance calculations were well explained by the PBE test performance investigations. Calculations of power and assurance were used to show how prior knowledge of test and reference products could be used to assess the probability of success upon practical changes from a commonly used trial design. Increasing the number of batches gave the most efficient gain; however, increasing the number of containers or lifestages also improved the probability of success. It would therefore be up to those designing a trial to assess the costs and benefits associated with changing the design. Using this more informed approach would prevent arbitrary increases in testing burden which do not sufficiently improve the probability of success and the execution of trials with unacceptably low probabilities of success.

In the future, it is hoped that further real-world data will be generated for an increasing number of *in vitro* metrics (like D_{50} and span) that can provide more informed simulations, similar to those conducted for impactor stage mass and emitted dose (8,15), and that additional experimental variables such as number of particles tested will be further understood. Ultimately work in this area should provide further tools for the generics industry to bring more complex locally acting drug products to the market.

REFERENCES

1. 21 CFR 320.1(e).
2. Lee SL, Adams WP, Li BV, Conner DP, Chowdhury BA, Yu LX. In vitro considerations to support bioequivalence of locally acting drugs in dry powder inhalers for lung diseases. AAPS J. 2009;11(3):414–23.
3. Lu Y, Chow SC, Zhu S. In vivo and in vitro bioequivalence testing. J Bioequivalence Bioavailab. 2014;6(2):67–74.
4. Gottlieb S. Reducing the hurdles for complex generic drug development [Internet]. 2017 [cited 2020 Jan 7]. Available from: https://www.fda.gov/news-events/fda-voices-perspectives-fda-leadership-and-experts/reducing-hurdles-complex-generic-drug-development. Accessed 7 Jan 2020.
5. Liu Q, Absar M, Saluja B, Guo C, Chowdhury B, Lionberger R, *et al.* Scientific considerations for the review and approval of first generic mometasone furoate nasal suspension spray in the United States from the bioequivalence perspective. AAPS J. 2019;21(2):1–6.
6. FDA. Statistical approaches to establishing bioequivalence [Internet]. 2001. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-approaches-establishing-bioequivalence. Accessed 7 Jan 2020.
7. Sandell D. A real case comparison of average and population bioequivalence for evaluation of APSD data. In: IPAC-RS/UF Orlando Inhalation Conference [Internet]. 2014. Available from: https://ipacrs.org/assets/uploads/outputs/08-Day_2_OIC_2014_Sandell.pdf. Accessed 7 Jan 2020.
8. Chen S, Morgan B, Beresford H, Burmeister Getz E, Christopher D, Långström G, *et al.* Performance of the population bioequivalence (PBE) statistical test with impactor sized mass data. AAPS PharmSciTech. 2019;20(7).
9. Grmaš J, Lužar-Stiffler V, Dreu R, Injac R. A novel simulation-based approach for comparing the population against average bioequivalence statistical test for the evaluation of nasal spray products on spray pattern and droplet size distribution parameters. AAPS PharmSciTech. 2019;20(1):1–14.
10. FDA. Bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action [Internet]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioavailability-and-bioequivalence-studies-nasal-aerosols-and-nasal-sprays-local-action. Accessed 7 Jan 2020.
11. FDA. Draft guidance on beclomethasone dipropionate [Internet]. 2019. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/psg/BeclomethasondipropionateInhalationAerosolMeteredNDA207921PSGPageRCMay2019.pdf. Accessed 7 Jan 2020.

12. FDA. Draft guidance on fluticasone propionate [Internet]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/psg/PSG_020121.pdf. Accessed 7 Jan 2020.
13. Chow SC, Shao J, Wang H. In vitro bioequivalence testing. Stat Med. 2003;22(1):55–68.
14. Morgan B, Strickland H. Performance properties of the population bioequivalence approach for in vitro delivered dose for orally inhaled respiratory products. AAPS J. 2014;16(1):89–100.
15. Morgan B, Chen S, Christopher D, Långström G, Wiggenhorn C, Burmeister Getz E, *et al*. Performance of the population bioequivalence (PBE) statistical test using an IPAC-RS database of delivered dose from metered dose inhalers. AAPS PharmSciTech. 2018;19(3):1410–25.
16. FDA. Draft guidance on budesonide [Internet]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/psg/Budesonide_Inhalation_Sus_20929_RC_09-12.pdf. Accessed 7 Jan 2020.
17. Hu M, Jiang X, Absar M, Choi S, Kozak D, Shen M, *et al*. Equivalence testing of complex particle size distribution profiles based on earth mover's distance. AAPS J. 2018;20(3):1–10.
18. FDA. Draft guidance on azelastine hydrochloride; fluticasone propionate [Internet]. 2020. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/psg/PSG_202236.PDF. Accessed 7 Jan 2020.
19. Ganley WJ, Shur J, Price R. Model informed design of in vitro bioequivalence trials. Respir Drug Deliv 2020. 2020;2:385–8.
20. Price RM, Bonett DG. Estimating the variance of the sample median. J Stat Comput Simul. 2001;68(3):295–305.
21. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, *et al*. Stan: a probabilistic programming language. J Stat Softw. 2017;76(1).
22. Stan Development Team. RStan: the R interface to Stan [Internet]. 2020. Available from: http://mc-stan.org/. Accessed 1 Feb 2020.
23. McElreath R. rethinking: statistical rethinking book package. 2020.
24. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria; 2020. Available from: https://www.r-project.org/. Accessed 1 Feb 2020.
25. Wickham H. ggplot2: elegant graphics for data analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org. Accessed 1 Feb 2020.
26. Campitelli E. metR: tools for easier analysis of meteorological fields [Internet]. 2020. Available from: https://cran.r-project.org/package=metR. Accessed 1 Feb 2020.
27. The authors thank reviewer 2 for suggestions on how to better relate this section to current guidance.
28. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharm Stat. 2005;4(3):187–201.
29. Chen DG (Din), Ho S. From statistical power to statistical assurance: it's time for a paradigm change in clinical trial design. Commun Stat Simul Comput 2017;46(10):7957–7971.
30. Carroll KJ. Decision making from phase II to phase III and the probability of success: reassured by assurance? J Biopharm Stat. 2013;23(5):1188–200.