Between-Batch Pharmacokinetic Variability Inflates Type I Error Rate in Conventional Bioequivalence Trials: A Randomized Advair Diskus® Clinical Trial

E Burmeister Getz¹, KJ Carroll², J Mielke³, LZ Benet⁴, B Jones³

¹Oriel Therapeutics, Inc., Berkeley, CA, USA
²KJC Statistics, Bramhall, Cheshire, UK
³Novartis Pharma AG, Basel, Switzerland
⁴University of California, San Francisco, CA, USA

Corresponding author:

Elise Burmeister Getz
Oriel Therapeutics, Inc.
Berkeley, CA 94710
EGetz@orieltherapeutics.com
ph: 510-295-5831

Keywords: fluticasone propionate, salmeterol, pharmacokinetics, bioequivalence, Advair Diskus, batch-to-batch pharmacokinetic variability, Type I error rate, consumer’s risk

Number of References: 27
Number of Figures: 4
Number of Tables: 4

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an ‘Accepted Article’, doi: 10.1002/cpt.535
Abstract

We previously demonstrated pharmacokinetic differences among manufacturing batches of an FDA-approved dry powder inhalation product (Advair Diskus® 100/50) large enough to establish between-batch bio-inequivalence. Here, we provide independent confirmation of pharmacokinetic bio-inequivalence among Advair Diskus 100/50 batches, and quantify residual and between-batch variance component magnitudes. These variance estimates are used to consider the Type I error rate of FDA’s current two-way crossover design recommendation. When between-batch pharmacokinetic variability is substantial, the conventional two-way crossover design cannot accomplish the objectives of FDA’s statistical bioequivalence test, i.e., cannot accurately estimate the Test/Reference ratio and associated confidence interval. The two-way crossover, which ignores between-batch pharmacokinetic variability, yields an artificially narrow confidence interval on the product comparison. The unavoidable consequence is Type I error rate inflation, to approximately 25%, when between-batch pharmacokinetic variability is nonzero. This risk of a false bioequivalence conclusion is substantially higher than asserted by regulators as acceptable consumer risk (5%).
**Introduction**

Advair Diskus® is an orally-inhaled dry powder product containing a corticosteroid (fluticasone propionate, FP) and a long-acting β₂-agonist (salmeterol xinafoate) to treat the two main components of asthma, inflammation and bronchoconstriction. Since its launch on the U.S. market in 2001, Advair Diskus has become a cornerstone of asthma therapy and is one of the top 10 prescription medications by sales in the U.S. International guidelines, such as those issued by Global Initiative in Asthma (GINA)¹,² and the National Heart, Lung and Blood Institute (NHLBI)³, advocate the use of inhaled long-acting β₂-adrenoceptor agonists (LABAs) in combination with inhaled corticosteroids (ICS) as maintenance therapy in asthma for subjects who remain symptomatic despite low to medium doses of ICS.

Despite its popularity, there is no generic version of Advair Diskus currently approved for the U.S. market despite patent expiry of the FP and salmeterol active ingredients in 2002 and 2008, respectively. In 2013, the U.S. Food and Drug Administration’s Office of Generic Drugs issued a draft guidance to industry⁴ that defined the bioequivalence requirements for a FP/salmeterol dry power inhaler (DPI). This draft guidance retains the conventional pharmacokinetic (PK) bioequivalence design and analysis methods, namely a two-way (i.e., two-treatment, two-period, two-sequence randomized) crossover comparison of a single manufacturing batch each of the Test (generic candidate) and Reference (Advair Diskus) products with a product ratio confidence interval derived from within-subject residual error as the assumed sole source of variability. Batch-to-batch PK variability, however, has been a consistent concern of both regulators and industry for several years in the context of orally-
inhaled drug product generic development and approval\textsuperscript{5,6,7,8}. Yet, despite this concern, FDA has not directly addressed batch-to-batch PK variability in its generic approval policies.

Recently, we demonstrated that batch-to-batch PK variability is substantial for Advair Diskus 100/50\textsuperscript{9}, in some cases demonstrating bio-inequivalence\textsuperscript{10} between batches. This variability is not necessarily surprising, given the wide acceptance range on \textit{in vitro} specifications of the FP/salmeterol dry powder product, specifically the two-fold acceptance range for respirable mass per the recent USP product-specific monograph\textsuperscript{11}, and the low systemic availability of orally-inhaled FP, cited as 5.5\% to 16.6\%\textsuperscript{12,13,14}. However, large batch-to-batch PK variability raises questions regarding the suitability of the conventional single-batch approach to PK bioequivalence testing.

Here we present a second and independent demonstration of PK bio-inequivalence among Advair Diskus 100/50 batches. Given this additional variance component (between-batch), we assess whether the recommended use of a single manufacturing batch in the PK bioequivalence study design assures the objectives of the FDA’s statistical bioequivalence test, \textit{i.e.}, provides an accurate and generalizable estimate of the Test/Reference ratio and associated confidence interval. We present the results of a clinical study designed to measure the PK from a single manufacturing batch of a generic DPI candidate (‘Test’) and three different manufacturing batches of Advair Diskus 100/50 (‘Reference’) in a four-way randomized crossover with all pairwise batch comparisons reported. Using average bioequivalence statistical methods, we assess the ability of the conventional 90\% confidence interval to quantify confidence not only in the specific \textit{batch} comparison but also in the \textit{product} comparison, recognizing that certainty in the product comparison is the true objective of bioequivalence testing. We give particular attention to the data from the
Reference product, as this is an FDA-approved product currently on the U.S. market. The goal of this work is to elucidate the increased risk of a Type I error (i.e., a false conclusion of bioequivalence) in the two-way (i.e., single batch) PK bioequivalence study design currently recommended by the FDA for Advair Diskus 100/50. Concern about Type I error rate in bioequivalence studies has been raised previously, although in a difference context.\textsuperscript{15,16}

The PK properties of single-dose 100 µg FP / 50 µg salmeterol from Advair Diskus 100/50 are not present in the literature, outside of the preceding publication.\textsuperscript{9} Here we present PK measurements from an additional four treatments with 100 µg FP / 50 µg salmeterol. These data, combined with those from the previous publication, provide a robust presentation of the PK of this widely-used drug product.

**Results**

Twenty-four subjects were allocated to six copies of a four-treatment, four-period Williams crossover design. The four treatment sequences used were A-B-C-D, B-D-A-C, C-A-D-B and D-C-B-A, where “A” was a single manufacturing batch of a generic ‘Test’ candidate, and “B”, “C” and “D” were three different manufacturing batches of Advair Diskus 100/50. Six subjects were randomly assigned to each of the four treatment sequences. All 24 randomized subjects completed all four periods of the study; no data were excluded from the analysis. First screening to last visit occurred between 20 September 2013 and 26 October 2013. Demographics of clinical study participants are given in Table 1. There were no serious adverse events or suspected unexpected serious adverse reactions on the study.
Least square geometric means and the ranges of individual subject values for the pharmacokinetics of FP and salmeterol from individual manufacturing batches are given in Table 2. Figure 1 illustrates the average blood concentration-vs-time profile for FP (first four hours after inhalation) and salmeterol (first hour after inhalation) from each batch.

At a dose of 100 µg to healthy adult subjects, FP is absorbed rapidly with a maximum plasma concentration ($C_{\text{max}}$) reached approximately 10 minutes after dosing. FP concentrations decline with an apparent terminal half-life of approximately 12 hours. At a dose of 50 µg to healthy adult subjects, salmeterol is absorbed very rapidly, with a maximum plasma concentration reached approximately 4 minutes after dosing. Distribution of salmeterol is also rapid, such that by 15 minutes post-dose the plasma concentration is less than half the peak value. Salmeterol concentrations decline with an apparent terminal half-life of approximately 14 hours.

Table 3 and Figure 2 demonstrate that differences among Reference batches were observed to be large enough to consistently fail the FDA’s PK bioequivalence requirement: all three Reference-vs-Reference pairwise batch comparisons failed the bioequivalence test required of a generic drug candidate. One comparison (batch 1 vs batch 2) demonstrated batch-to-batch PK bio-inequivalence, with 90% confidence intervals around the Test/Reference geometric mean ratio (GMR) of all PK metrics for both drug substances excluding the 80-125% bioequivalence region. For this batch pair, batch-to-batch ratios (90% confidence interval) for FP $C_{\text{max}}$, FP AUC, salmeterol $C_{\text{max}}$ and salmeterol AUC were observed to reach 151.03% (136.70-166.88%), 156.99% (136.35-180.74%), 159.79% (140.61-181.59%) and 151.39% (133.05-172.26%), respectively, independently confirming the previously...
published clinical study result of bio-inequivalence between different batches of Advair Diskus 100/50.9.

The Test-vs-Reference comparison based on individual batch pairs was highly dependent on which batch of Reference was selected. Using FP $C_{\text{max}}$ as an example metric, the Test/Reference GMR (with 90% confidence interval) for comparison of the Test batch to individual Reference batches ranged from 88.05% (79.69–97.29%) to 132.99% (120.36–146.94%). Here, these GMR differences are driven almost entirely by between-batch PK variation in the Reference product. Thus the estimated GMR demonstrated for any single comparison of the Test product to an individual Reference batch may misrepresent the true relationship between the products in the presence of batch-to-batch PK variability. This point is a straightforward conclusion from the variability of Reference presented here and in the preceding publication. Implications of between-batch PK variability on the GMR confidence interval, however, may not be immediately obvious; these implications are considered here, following variance component estimation.

Table 4 presents the estimated magnitudes of within-subject residual variance ($\sigma^2_e$) and within-subject between-batch ($\sigma^2_\beta$) variance for the Reference product, for which three different manufacturing batches were administered (thus providing two degrees of freedom on the between-batch variance component estimate). Of no surprise given the preceding results, the largest source of variability for most PK metrics was found to come from differences between batches. Again using FP $C_{\text{max}}$ as an example metric, the estimated between-batch variance (0.0598) was 1.7-fold larger than the estimated residual variance (0.0353) and was a highly-significant contributor to total variability ($p < 0.0001$). Results for salmeterol $C_{\text{max}}$ were similar, with an estimated between-batch variance (0.0718) 1.2-fold
larger than estimated residual variance (0.0617) and again a highly-significant contributor to total variability (p < 0.0001).

The within-subject residual variance identified from the Reference data suggests that under the FDA’s current definition, Advair Diskus 100/50 would not be considered a highly variable drug product. The current FDA definition of ‘highly variable’ considers only dispositional variability (i.e., variability in the body’s handling of the drug, for example, variable absorption or metabolism), without regard for manufacturing or product variability\textsuperscript{17}. Although here using a methodology (analysis of variance model with batch as a fixed effect, see Methods) that differs from the FDA’s batch-replication approach to assessing residual error\textsuperscript{18}, because no batches were replicated in the current study, the data yield within-subject residual standard deviation (WSSD, $\sqrt{\sigma^2_e}$) estimates of 0.16 to 0.25 (Table 4), consistently below the FDA’s current highly variable drug criterion of WSSD ≥ 0.294. Thus, despite substantial within-subject PK variability across batches, Advair Diskus 100/50 does not display evidence of high PK variability within a batch, and therefore is not likely to be eligible for the FDA’s current Reference-scaling methodology that widens the bioequivalence limits according to variability of the Reference product.

Variance components within the range estimated from the clinical data (Table 4) were used to construct expected Test/Reference ratio distributions for a two-way crossover design with varying amounts of between-batch PK variability; these are illustrated in Figure 3. The two-way crossover design was assumed to be two-treatment (Test, Reference), two-sequence (TR, RT) and two-period, aligned with current FDA bioequivalence study design recommendations. Here, the impact of between-batch PK variability on the shape of the...
Test/Reference ratio distribution is illustrated for distributions centered at a true Test/Reference ratio of 1.05, and with true residual error variance ($\sigma^2_e$) of 0.04 (i.e., WSSD = 0.20, equivalent to a within-subject coefficient of variation of 20.2%). When between-batch variability is zero ($\sigma^2_b = 0$), conventional sample size calculations for a two-treatment, two-period crossover design indicate that 26 subjects are needed to provide 90% power to conclude bioequivalence for true $\sigma^2_e = 0.04$ and a true Test/Reference ratio of 1.05. If this study design is repeated many, many times, with each study producing estimates of $\sigma^2_e$, the Test/Reference ratio and its confidence interval, the collection of 90% confidence intervals will demonstrate two key features: (i) 90% of the intervals will contain the true Test/Reference mean ratio, and (ii) only 5% of the intervals will lie within the bioequivalence limits (0.80 – 1.25) if the true Test/Reference ratio is 0.80 or 1.25 (i.e., when the products are not bioequivalent, bioequivalence will be concluded in only 5% of trials).

In general terms, the 90% confidence interval is often interpreted as an interval that provides a good estimate of the true ratio between the Test and Reference products.

This interpretation of the standard Test/Reference confidence interval, however, fails for a two-way (i.e., two-period, two-treatment) crossover design in the presence of between-batch variability. When the true, underlying mean and within-subject variance component values are Test/Reference = 1.05, $\sigma^2_e = 0.04$, and $\sigma^2_b = 0$, the expected 90% confidence interval for the design can be shown to be 0.955 - 1.155, indicated by the filled portion of the $\sigma^2_b = 0$’ distribution in Figure 3. Due to widening of the Test/Reference ratio distribution by the addition of between-batch variability, this (alleged) “90%” confidence interval covers the true Test/Reference mean ratio with a probability of only 46%, 29% or 21% when between-batch variance ($\sigma^2_b$) is 0.01, 0.03 or 0.06 (indicated by the red, green and gray
distributions in Figure 3, respectively). Thus the two-way crossover design confidence interval (that cannot estimate between-batch variability and therefore ignores it in confidence interval construction, \textit{i.e.}, underestimates true total within-subject variability) can no longer be interpreted as providing a good interval estimate of the Test/Reference ratio. The two-way crossover design confidence interval is, simply, incorrectly narrow.

It is this inadequate confidence interval coverage that is directly responsible for inflation of the Type I error rate. The Type I error rate from the two-way crossover design with true underlying values of \( \sigma_e^2 = 0.04 \), \( \sigma_b^2 = 0 \) and \( T/R = 1.25 \) is 5\%, the conventionally accepted upper limit on consumer risk. However, the same study design using Test and Reference products with inherent between-batch variance of 0.01, 0.03 or 0.06 (\textit{i.e.}, \( \sigma_b^2 / \sigma_e^2 \) variance ratios of 0.25, 0.75 and 1.50) leads to an inflation of the Type I error rate to 25\%, 27\% or 23\%, respectively (Figure 4). When the null hypothesis (\( H_0 \), non-equivalence) is true and the confidence interval is correctly calculated (as in the case of a two-way crossover design when between-batch variance is zero), there is only a 5\% chance of the confidence interval being fully contained within the bioequivalence limits. But, with non-zero between-batch variance, the observed Test/Reference ratio varies widely with an increased opportunity of being close to 1.0, thus increasing the chance of (incorrectly) rejecting \( H_0 \) (\textit{i.e.}, confidence interval entirely contained within 0.80-1.25) if the confidence interval remains artificially narrow, \textit{i.e.}, does not also reflect the widening of the underlying distribution. Hence the Type I error rate is increased.

The steep dependence of Type I error rate on between-batch variance as between-batch variance increases from zero emphasizes the importance of acknowledging between-batch
variability in the PK bioequivalence assessment for those drug products for which even small levels of between-batch variability are anticipated or demonstrated. In the two-way crossover study described above (26 subjects, $\sigma_e^2 = 0.04$, $T/R = 1.25$), a between-batch variance only 5% of residual error variance (i.e., $\sigma_b^2 = 0.002$) inflates the Type I error rate to 14% (Figure 4). As expected, when study designs ignore variability, the results are prone to misinterpretation.

The Type I error rate does not increase monotonically with between-batch variability. Initially, increases in variability between batches increases the probability of demonstrating bioequivalence in a two-way crossover design, due to chance batch selection, even when bioequivalence is an incorrect conclusion. Further increases of between-batch variability reduce the probability of demonstrating bioequivalence regardless of whether bioequivalence is the correct answer or not, as the probability diminishes of selecting, by chance, batches that agree. However, the region of inflated Type I error rate persists for what appears to be the full extent of clinically-realistic variance values; the probability of an incorrect bioequivalence conclusion does not fall below 5% for the two-way crossover study example considered here until between-batch variance is overwhelming (Type I error rate returns to $\leq 5\%$ for $\sigma_b^2/\sigma_e^2$ variance ratios exceeding approximately 50).

**Discussion**

For many (perhaps most) drug products, process controls limit product variability sufficiently well to allow a single manufacturing batch to represent the product for bioequivalence testing. In a conventional PK bioequivalence study, single manufacturing
batches each of the Test and Reference products are compared. Both regulators and industry
consider this design as providing a bioequivalence result that is generalizable and
representative of patients’ experience in commercial use of the products over time, all
predicated on the assumption that batch-to-batch PK variability is negligible.

However, certain products and dosage forms are more susceptible to batch-to-batch PK
variability. For example, here, and previously, we have demonstrated that PK variability
among manufacturing batches of the FP/salmeterol dry powder combination product Advair
Diskus 100/50 is unmistakably present and too substantial to ignore. In both the previous
study\(^9\) and again here, PK differences between Advair Diskus 100/50 batches are large
enough to demonstrate PK bio-
inequivalence of the product to itself, with conventional 90%
confidence intervals entirely excluding the 80-125% bioequivalence region. This additional
source of variability poses important challenges for bioequivalence testing. The European
Medicines Agency has formally acknowledged PK batch-to-batch variability, and suggested
potential approaches\(^19\). To date, however, PK batch-to-batch variability is neither addressed
nor accounted for in any FDA bioequivalence guidance.

The marked failure of the Reference product to meet the PK bioequivalence criteria when
tested against itself is not a consequence of the relatively small study size, in fact the use of
only 24 subjects makes a demonstration of bio-
inequivalence more difficult. Nor are the
between-batch PK differences attributable to use of Reference product near the end of shelf-
life; all Reference batches had more than five months remaining until expiry at the time of
dosing. The Reference product does not demonstrate high dispositional variability per the
FDA’s definition of a highly-variable drug product, but does demonstrate batch-to-batch PK
differences larger than are consistent with the statistical test of bioequivalence applied to Test-vs-Reference comparisons.

The FDA’s statistical test for bioequivalence, using average bioequivalence methodology, is a requirement on the 90% confidence interval around the geometric mean Test/Reference ratio (GMR). Here, we have shown that the confidence interval constructed from a conventional bioequivalence study design (i.e., one batch of Test versus one batch of Reference), although correct for the single specific batch of Test and Reference selected for that particular study, cannot be generalized, with adequate error rate control, to the product comparison when there is unaccounted batch-to-batch PK variability. When batch-to-batch PK variability is present but ignored, the standard error of the GMR is underestimated and the calculated confidence interval is artificially narrow, thus inflating the Type I error rate and increasing the risk of erroneous licensing decisions.

Using estimates of residual and between-batch variance that are representative of the clinical observations, we have quantified the Type I error rate inherent in the FDA-recommended two-way crossover design. The resulting broadening of the product ratio distribution with increasing between-batch variability increases the Type I error rate (“consumer’s risk”) to approximately 25%, as increased variability in the product ratio increases the probability of observing a batch ratio within the bioequivalence region even when the product ratio is non-equivalent. This Type I error rate inflation cannot be eliminated by an increase in the number of study subjects because the source of the inflation is variability between batches, not between subjects. Of note, FDA-mandated in vitro bioequivalence testing for inhalation products already requires inclusion of multiple manufacturing batches per product (minimum of three), presumably to address exactly the issue of between-batch variability.
The implications of unaccounted batch-to-batch PK variability on bioequivalence conclusions are illustrated here for the two-way crossover design because this study design is ubiquitous in bioequivalence testing, and is explicitly referenced in the Advair Diskus 100/50 product-specific FDA draft guidance\textsuperscript{4}. The extent of Type I error rate inflation could be moderated with more sophisticated study designs, although discussion of such alternate designs is beyond the scope of the current manuscript. Additionally, we have assumed, for symmetry and simplicity, that batch-to-batch PK variability affects the Test and Reference products equally. Absence of batch-to-batch PK variability in the Test product mitigates, but does not eliminate, Type I error rate inflation. The specific impact of batch-to-batch PK variability on the bioequivalence assessment will depend on the details of study design and Test and Reference product performance. The principal of incorporating clinical data from multiple batches, however, remains an essential and necessary component of bioequivalence testing following emergence of credible data demonstrating batch diversity.

Just as the FDA’s Reference-scaling method\textsuperscript{18} eliminates unnecessary human testing for products with high dispositional variability, so too could an extension of this statistical methodology reduce the regulatory burden in instances of batch-to-batch PK variability. Advair Diskus 100/50 does not display high dispositional variability for either of its active ingredients (\textit{i.e.}, WSSD < 0.294). Instead, Advair Diskus 100/50 PK variability is inherent to the product and is not the consequence of variation introduced by the body’s action on the drug. Because there is no public information to suggest significant changes to the Reference product since its approval, it is reasonable to expect that this product-based variability was present during the innovator’s safety and efficacy testing. Accordingly, as with high dispositional variability, the variability indicates a wide therapeutic index for this...
combination product. At present, however, there is no statistical method approved by regulators for handling between-batch PK variability of any magnitude. In a subsequent presentation we propose an extension of FDA’s existing Reference-scaling methodology to accommodate between-batch variability.

Between-batch variability in bioequivalence testing is analogous to heterogeneity in meta-analyses, where randomized trials of similar design, subject population and primary endpoint can give disparate results. In these instances, the treatment effect observed in any single trial does not provide a reliable estimate of the truth – only the average effect across several trials accommodates the heterogeneity to provide a reliable estimate with associated confidence interval. Importantly, a critical feature of meta-analysis in the presence of trial heterogeneity (i.e., random-effect meta-analysis\textsuperscript{20}) is that the heterogeneity across trials (here, across batches) is directly incorporated into the confidence interval to reflect the extent to which different outcomes are possible even within the constraints of the controlled clinical trial environment.

Aside from between-batch diversity, aspects of FP and salmeterol PK when administered as a single inhalation of 100 µg (FP) or 50 µg (salmeterol) merit comment. Time to peak FP plasma concentration (T\textsubscript{max}) was observed to be approximately 10 minutes, consistent with a report from single-inhalation repeated administration of Advair Diskus 100/50\textsuperscript{21}, but earlier than observed for higher doses. FP T\textsubscript{max} values in healthy subjects following single-dose dry powder (Diskus) inhalation have been reported as ranging from 25 to 66 minutes at 1,000 µg (as four 250-µg inhalations)\textsuperscript{12,22}, and 45 to 75 minutes at 400 µg (as four 100-µg inhalations)\textsuperscript{23}. It may be that FP absorption rate, and therefore T\textsubscript{max}, is sensitive to the total FP dose. The apparent terminal elimination half-life of FP reported here of 12 hours is
similar to previous reports\textsuperscript{24,25}. The PK of 50 \( \mu \)g salmeterol in healthy subjects following dry powder oral inhalation was previously reported\textsuperscript{9,26,27}. The salmeterol \( T_{\text{max}} \) of 4 minutes and terminal half-life of approximately 14 hours are consistent with these previous reports.

\textbf{Methods}

The pharmacokinetics of FP and salmeterol were observed in 24 enrolled (24 evaluable) healthy adult subjects in a clinical study performed under clinical trials authorization from the UK Medicines and Healthcare products Regulatory Agency and approval by National Research Ethics Service Committee. Written informed consent was obtained from all subjects, and all studies were conducted in accordance with the principles of the Declaration of Helsinki.

\textit{Study design}

A single dose of 100 \( \mu \)g FP with 50 \( \mu \)g salmeterol was administered by oral inhalation as either Advair Diskus 100/50 ("Reference") or the strength-matched development product ("Test"). The study used a 4-period, 4-sequence design and was conducted as a single-center, randomized, open-label, crossover, single-dose study in healthy adult males and females at Quintiles Drug Research Unit at Guy’s Hospital, London, UK. The primary objective was determination of each of the six pairwise batch comparisons (\textit{i.e.}, T/R1 T/R2, T/R3, R1/R2, R1/R3, R2/R3), using conventional average bioequivalence methodology, where ‘T’ indicates the single batch of the Test product and ‘R1’, ‘R2’ and ‘R3’ indicate the three batches of the Reference product.
Advair Diskus 100/50 (GlaxoSmithKline, Research Triangle Park, NC, USA) was purchased directly from the US market and used within labeled expiry. All Reference batches were supplied to the European clinical site in a single insulated shipment with temperature monitoring, and stored in a single temperature-controlled pharmacy to ensure that all batches were handled identically. The strength-matched Test product was manufactured for Oriel Therapeutics as powder-blend combination of micronized FP (100 µg), micronized salmeterol as the xinafoate salt (50 µg) and lactose monohydrate, an inert excipient, contained in a multi-dose dry powder inhalation device similar in size and operation to the commercially available Reference product.

All study treatments were administered under supervision and participants remained in the clinic for the duration of dosing and PK collection. The dosing procedure followed the instructions provided to patients in the Advair Diskus Medication Guide\textsuperscript{10}, namely exhalation, quick and deep inhalation with 10-second breath hold, and mouth rinse. Subjects were fasted overnight for at least 10 hours prior to dosing until 4 hours post-dose. Water was allowed \textit{ad lib} except 1 hour prior through 1 hour post-dose. Crossover treatments were separated by a washout of at least 7 days.

\textit{Clinical study participants}

Participants (≥18 years, body weight ≥50 kg, body mass index 18.8 to 29.8 kg/m\textsuperscript{2} (inclusive)) had no history of asthma, a fractional exhaled nitric oxide value ≤47 ppb and a forced expiratory volume in 1 second (FEV\textsubscript{1}) ≥90% of predicted at screening.

\textit{Pharmacokinetic samples}
Serial blood samples (6mL) were drawn for determination of FP and salmeterol concentrations prior to each dose and post-dose following inhalation at 3, 4, 5, 6, 8, 10, 15, 20, 30 and 45 minutes and 1, 2, 4, 8, 12, 16, 24, 32, 40, 48 and 56 hours.

Blood samples were centrifuged within 30 minutes of collection at approximately 2000g for 15 minutes at 4°C. The bioanalysis of FP and salmeterol was conducted by Covance (Salt Lake City, UT, USA) using a validated LC-MS/MS method with a quantitative range from 1.00 to 200 pg/mL for each analyte.

**Pharmacokinetic analysis**

Estimated PK parameters were maximum observed plasma concentration (C\text{max}) and time to C\text{max} (T\text{max}), areas under the concentration-vs-time curve to the last time of quantifiable concentration [AUC\text{0-t}] calculated using the linear trapezoidal method and extrapolated to infinity [AUC\text{0-inf}], and elimination rate constant (λ\text{Z}) and corresponding half-life (t\text{1/2}).

**Statistical analysis**

PK parameter least square geometric means were determined for individual batches using an analysis of variance (ANOVA) model with fixed effects for batch, period and sequence, and a random subject-within-sequence term, using natural logarithms of the data. Treatments were compared using average bioequivalence methods. Treatment ratio point estimates and confidence intervals were exponentiated back to the original scale for display.

Comparison of the relative magnitudes of within-subject residual error variance and within-subject between-batch variance is presented for the Reference product, for which the PK from three different manufacturing batches were measured. Variance component estimation
was based on a Type 3 analysis using SAS PROC MIXED that provided a full ANOVA table indicating sources of variation (including residual error variance), associated degrees of freedom (DF), sums of squares, mean squares and also expected mean squares, the error term and error DF for each of the expected mean squares. These outputs allowed Method of Moments (MM) estimation of the variance components for the random effect terms specified in the PROC MIXED model code. A supplementary analysis using the PROC MIXED option method=REML to provide restricted maximum likelihood (REML) estimation for the variance components was performed.

Test/Reference ratio distributions from a two-way crossover involving a single batch each of Test and Reference were determined using a standard error (SE) of the log Test/Reference ratio estimate, \( \hat{\theta} \), of \( \sqrt{\frac{\sigma_e^2}{m} + \frac{\sigma_{br}^2}{m} + \frac{\sigma_{bt}^2}{m}} \) where \( m \) is the number of subjects per sequence (2m total clinical study participants), \( \sigma_e^2 \) is within-subject estimated residual error variance, and \( \sigma_{bt}^2 \) and \( \sigma_{br}^2 \) are estimated batch-to-batch variance for Test and Reference, respectively. If batch-to-batch variance for Test and Reference are of equal magnitude, SE becomes \( \sqrt{\frac{\sigma_e^2}{m} + 2\sigma_b^2} \). With batch-to-batch variability, the correct 90% confidence interval for the log Test/Reference ratio is given by \( \hat{\theta} \pm t_{0.95,2m-2}\sqrt{\frac{\sigma_e^2}{m} + 2\sigma_b^2} \), whereas the simple two-way crossover that is ignorant of batch-to-batch variability gives a 90% confidence interval of \( \hat{\theta} \pm t_{0.95,2m-2}\sqrt{\frac{\sigma_e^2}{m}} \); in both cases \( t_{1-\alpha,df} \) represents the \( 100(1-\alpha) \) th percentile of the centralized t-distribution with \( df \) degrees of freedom. Hence, when \( \sigma_b^2 \) is non-zero, the correct confidence interval is wider than that from a two-way crossover that ignores batch-to-batch variability. It is this underestimation of the confidence interval in the two-way crossover design that gives rise to an increased Type I error rate.
The Type I error rate is easily simulated for a two-way (i.e., 2x2) crossover when there is both within-subject residual and between-batch variability. Denoting treatment sequences 1 and 2 as T→R and R→T, respectively, the basic 2x2 crossover model is given by:

\[ y_{ijk} = \mu + p_{ik} + \pi_j + t_{ij} + e_{ijk} \]

where:

- \( p_{ik} \) = effect of subject \( k \) in sequence group \( i \), \( i = 1, 2, k = 1, \ldots, m \)
- \( \pi_j \) = effect of period \( j \), \( j = 1, 2 \)
- \( t_{ij} \) = effect of treatment given in period \( j \), sequence group \( i \)
- \( e_{ijk} \) = random error associated subject \( k \), period \( j \), sequence group \( i \)

\[ p_{ik} \sim N(0, \sigma_p^2) \text{ and independently } e_{ijk} \sim N(0, \sigma_e^2) \]

Batch-to-batch variability is incorporated into each simulated trial by adding a Reference product batch effect, \( b_r \), to all subjects in period 2, sequence 1 and period 1, sequence 2 where \( b_r \sim N(0, \sigma_{br}^2) \), and, similarly, a Test product batch effect, \( b_t \), to all subjects in period 1, sequence 1 and period 2, sequence 2 where \( b_t \sim N(0, \sigma_{bt}^2) \). \( b_r, b_t, p_{ik}, e_{ijk} \) are mutually independent. Illustrative results from 10,000 trial simulations are tabulated below (Table M1) for \( \sigma_e^2 = 0.04 \) and \( \sigma_f^2 = \sigma_b^2 = 0, 0.01, 0.03 \) and 0.06.

Alternatively, Type I error rate may be arrived at by an approximate analytical approach.

Consider the usual null and alternative hypotheses:
\[ H_0: |\mu_T - \mu_R| \geq \ln(1.25) \text{ vs } H_1: |\mu_T - \mu_R| < \ln(1.25) \]

where \( \mu_T \) and \( \mu_R \) denote true Test and Reference log mean PK parameter values, with \( \mu_T - \mu_R = \theta \). \( H_0 \) is rejected if the observed difference in log means is small, i.e., if \( |\hat{\mu}_T - \hat{\mu}_R| < k \). If we consider the standard bioequivalence alternative hypothesis of \( |\mu_T - \mu_R| = 0 \), we have:

\[ \alpha = \Pr (|\hat{\mu}_T - \hat{\mu}_R| < k \mid \mu_T - \mu_R = \ln(1.25)) \]

\[ 1 - \beta = \Pr (|\hat{\mu}_T - \hat{\mu}_R| < k \mid \mu_T - \mu_R = 0). \]

Thus for any two-way crossover,

\[ k = \ln(1.25) - t_{1-\alpha,df}SE(\hat{\mu}_T - \hat{\mu}_R) = \ln(1.25) - t_{1-\alpha,df}\sqrt{\frac{\sigma^2_e}{m}}. \]

Hence,

\[ \text{Type I error} = \Pr (|\hat{\mu}_T - \hat{\mu}_R| < k \mid \mu_T - \mu_R = \ln(1.25) \text{ and correct SE} = \sqrt{\frac{\sigma^2_e}{m} + 2\sigma^2_b}). \]

We may approximate the Type I error as:

\[ \text{Approximate Type I error} = \Pr \left( \frac{-k - \ln(1.25)}{\sqrt{\frac{\sigma^2_e}{m} + 2\sigma^2_b}} < T_{2m-2} < \frac{k - \ln(1.25)}{\sqrt{\frac{\sigma^2_e}{m} + 2\sigma^2_b}} \right) \]

where \( T_{2m-2} \) represents the centralized \( t \)-distribution with \( df = 2m - 2 \). Hence,

\[ \text{Approximate Type I error} = \Pr \left( \frac{-2\ln(1.25) + t_{1-\alpha,2m-2}\sqrt{\sigma^2_e/m}}{\sqrt{\frac{\sigma^2_e}{m} + 2\sigma^2_b}} < T_{2m-2} < \frac{-t_{1-\alpha,2m-2}\sqrt{\sigma^2_e/m}}{\sqrt{\frac{\sigma^2_e}{m} + 2\sigma^2_b}} \right) \]
At the trial design stage we substitute $\sigma^*_e$ and $\sigma^*_h$ for their assumed values $\sigma^2_e$ and $\sigma^2_h$ to give

$$\approx \text{Type I error}$$

Of note, this formula reduces to the standard sample size approach employed by commercially available software (e.g., SAS, nQuery Advisor) for $\sigma^2_h = 0$. The approximate analytical results are shown alongside trial simulation results in the tabulation below (Table M1). As expected, a close match is evident.

**Study Highlights**

- **What is the current knowledge on the topic?** Pharmacokinetic bioequivalence studies conventionally compare single manufacturing batches of Test and Reference.

- **What question did this study address?** The current work investigates the reproducibility of a previous bio-inequivalent result between batches of U.S. Advair Diskus® 100/50, and assesses the ability of single-batch bioequivalence studies to accurately estimate the product ratio and associated confidence interval when batch-to-batch pharmacokinetic variability is present.
• **What does this study add to our knowledge?** The FDA-recommended two-way crossover design to assess pharmacokinetic bioequivalence fails to control Type I error rate when batch-to-batch pharmacokinetic variability is present. Batch-to-batch variability has been reproducibly demonstrated for Advair Diskus 100/50. Use of a single manufacturing batch leads to an unreliable Test/Reference estimate and a too-narrow confidence interval when between-batch pharmacokinetic variability is ignored.

• **How might this change clinical pharmacology or translational science?** To align regulation with emerging science, the bioequivalence paradigm requires revision when batch-to-batch pharmacokinetic variability is substantial and reproducible. Although between-batch pharmacokinetic variability is demonstrated here for a specific drug product, this variability source may be present for other products.

**Author Contributions**

E.B.G., K.J.C., J.M., L.Z.B., and B.J. wrote the manuscript; E.B.G. designed the research; E.B.G. performed the research; K.J.C. analyzed the data.

**Acknowledgements**

E.B.G. is an employee of Oriel Therapeutics, an indirect wholly-owned subsidiary of Novartis Pharma AG. K.J.C. and L.Z.B. are paid consultants to Oriel Therapeutics. B.J. and J.M. are employees of Novartis Pharma AG. J.M. was supported by the Swiss State.
Secretariat for Education, Research and Innovation, contract number 999754557. Opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government. J.M.’s work is part of the IDEAS European training network (http://www.ideas-itn.eu/) from the European Union’s Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 633567. The study was paid for by Oriel Therapeutics.

References


Figure Legends

**Figure 1:** Plasma concentration-vs-time profiles for fluticasone propionate (FP; 100 µg) and salmeterol (50 µg) following single-dose dry powder oral inhalation to healthy adult subjects as Advair Diskus® 100/50 (gray scale) or the Test product (red).

**Figure 2:** Pharmacokinetic comparisons between individual Advair Diskus® 100/50 (Reference) batches are shown with geometric mean ratios (GMR) and 90% confidence intervals (CI) for fluticasone propionate (FP) and salmeterol maximum observed plasma concentration ($C_{\text{max}}$) and area under the plasma concentration-vs-time curve (AUC). Individual Reference batches are indicated as ‘R1’, ‘R2’ or ‘R3’. A ratio value of 1.00 is shown via a horizontal red line. The 0.80 – 1.25 bioequivalence region is crosshatched. The ratio (y) axis is plotted on a log scale.

**Figure 3:** Distributions of the Test/Reference ratio estimate from a two-way crossover bioequivalence study design in which a single randomly selected Test batch is compared to a single randomly selected Reference batch in 26 subjects. On the logarithmic scale, the within-subject residual error variance is assumed to be 0.04. On the natural scale, the true Test/Reference ratio is assumed to be 1.05. Specific distributions are shown for between-batch variance values ($\sigma_b^2$) on the log-scale of zero (blue), 0.01 (red), 0.03 (green) and 0.06 (gray). The expected range of the 90% confidence interval of the Test/Reference ratio assuming $\sigma_b^2 = 0$ is shown as a shaded area to illustrate the coverage of a 90% confidence interval derived from a two-way crossover design. For non-zero $\sigma_b^2$, ...
the two-way crossover design 90% confidence interval provides only a fraction of the coverage provided for $\sigma_b^2 = 0$.

**Figure 4:** Type I error rate from a two-way crossover bioequivalence study design in which a single randomly selected Test batch is compared to a single randomly selected Reference batch in 26 subjects. On the logarithmic scale, the within-subject residual error variance ($\sigma^2_e$) is assumed to be 0.04. Log-scale between-batch variance ($\sigma_b^2$) is assumed to vary from zero to 0.10 (corresponding to $\sigma_b^2 / \sigma_e^2$ variance ratios ranging from zero to 2.5), with equal between-batch variance on Test and Reference products. To assess Type I error rate, the true Test/Reference ratio is assumed to be 1.25 on the natural scale. Simulation results (green circles) are compared to the approximate analytical solution (blue line).
This article is protected by copyright. All rights reserved.
This article is protected by copyright. All rights reserved.
Table 1: Subject demographics

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EudraCT Number</td>
<td>2013-003071-35</td>
<td>Population</td>
<td>healthy</td>
</tr>
<tr>
<td>FEV₁ (% predicted)</td>
<td>≥90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>39 ± 10 (22 – 56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M/F</td>
<td>20/4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>74.3 ± 10.5 (57.8 – 96.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height (cm)</td>
<td>175 ± 6 (166 – 187)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>24.2 ± 3.1 (19.7 – 29.4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data are mean ± standard deviation (minimum – maximum)
Table 2: Summary of pharmacokinetic parameters for 100 µg fluticasone propionate (FP) and 50 µg salmeterol (S) following administration to healthy subjects.

<table>
<thead>
<tr>
<th></th>
<th>Test</th>
<th>Reference, Advair Diskus® 100/50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Batch 1</td>
<td>Batch 1</td>
</tr>
<tr>
<td><strong>FP C_{max} (pg/mL)</strong></td>
<td>46.5 [25.8 – 88.5]</td>
<td>35.0 [17.2 – 52.1]</td>
</tr>
<tr>
<td><strong>S C_{max} (pg/mL)</strong></td>
<td>76.5 [29.7 – 172]</td>
<td>71.0 [24.7 – 126]</td>
</tr>
<tr>
<td><strong>S AUC_{0-1} (h·pg/mL)</strong></td>
<td>114 [50 – 301]</td>
<td>113 [41 – 333]</td>
</tr>
</tbody>
</table>

Least square geometric mean [range] except T_{max} for which the median [range] is reported.

This article is protected by copyright. All rights reserved.
Table 3: Bioequivalence assessment between manufacturing batches of Advair Diskus® 100/50.

Average bioequivalence methods were used for pairwise comparisons of three different manufacturing batches of Advair Diskus 100/50. FP: fluticasone propionate. S: salmeterol.

<table>
<thead>
<tr>
<th>Geometric Mean Ratio (%) Among Batches of Advair Diskus 100/50</th>
<th>Estimate</th>
<th>90% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch 1 – vs – Batch 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>151.03</td>
<td>136.70 – 166.88</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>147.55</td>
<td>135.21 – 161.02</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>156.99</td>
<td>136.35 – 180.74</td>
</tr>
<tr>
<td>S C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>159.79</td>
<td>140.61 – 181.59</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>140.80</td>
<td>126.61 – 156.59</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>151.39</td>
<td>133.05 – 172.26</td>
</tr>
<tr>
<td>Batch 1 – vs – Batch 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>147.64</td>
<td>133.42 – 163.37</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>131.28</td>
<td>120.14 – 143.45</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>137.95</td>
<td>121.96 – 156.03</td>
</tr>
<tr>
<td>S C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>141.33</td>
<td>124.13 – 160.92</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>128.28</td>
<td>115.16 – 142.88</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>130.28</td>
<td>116.41 – 145.81</td>
</tr>
<tr>
<td>Batch 2 – vs – Batch 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>102.30</td>
<td>92.45 – 113.20</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>112.40</td>
<td>102.86 – 122.82</td>
</tr>
<tr>
<td>FP AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>113.80</td>
<td>98.03 – 132.12</td>
</tr>
<tr>
<td>S C&lt;sub&gt;max&lt;/sub&gt;</td>
<td>113.06</td>
<td>99.30 – 128.73</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-t)&lt;/sub&gt;</td>
<td>109.76</td>
<td>98.54 – 122.27</td>
</tr>
<tr>
<td>S AUC&lt;sub&gt;(0-inf)&lt;/sub&gt;</td>
<td>116.20</td>
<td>102.12 – 132.23</td>
</tr>
</tbody>
</table>
Table 4: Variance component estimation following administration of a single dose of Advair Diskus® 100/50 from three different manufacturing batches to healthy subjects. $\sigma_b^2$: within-subject, between-batch variance estimate. $\sigma_e^2$: within-subject residual error variance estimate. FP: fluticasone propionate. S: salmeterol.

<table>
<thead>
<tr>
<th>Metric</th>
<th>DF error</th>
<th>$\sigma_e^2$ (^1)</th>
<th>DF batch</th>
<th>F-value for batch</th>
<th>p-value for batch</th>
<th>Method of Moments(^2) estimate</th>
<th>$\sigma_b^2 / \sigma_e^2$</th>
<th>REML(^3) estimate</th>
<th>$\sigma_b^2 / \sigma_e^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP C(\text{max})</td>
<td>43</td>
<td>0.03531</td>
<td>2</td>
<td>35.92</td>
<td>&lt;.0001</td>
<td>0.05984 1.69</td>
<td></td>
<td>0.06828 1.93</td>
<td></td>
</tr>
<tr>
<td>FP AUC((0-\inf))</td>
<td>15</td>
<td>0.03039</td>
<td>2</td>
<td>10.85</td>
<td>0.0012</td>
<td>0.04628 1.52</td>
<td></td>
<td>0.05808 1.91</td>
<td></td>
</tr>
<tr>
<td>FP AUC((0-t))</td>
<td>43</td>
<td>0.02578</td>
<td>2</td>
<td>33.37</td>
<td>&lt;.0001</td>
<td>0.04049 1.57</td>
<td></td>
<td>0.04541 1.76</td>
<td></td>
</tr>
<tr>
<td>S C(\text{max})</td>
<td>43</td>
<td>0.06171</td>
<td>2</td>
<td>24.97</td>
<td>&lt;.0001</td>
<td>0.07179 1.16</td>
<td></td>
<td>0.08093 1.31</td>
<td></td>
</tr>
<tr>
<td>S AUC((0-\inf))</td>
<td>26</td>
<td>0.04715</td>
<td>2</td>
<td>9.87</td>
<td>0.0006</td>
<td>0.03171 0.67</td>
<td></td>
<td>0.04231 0.90</td>
<td></td>
</tr>
<tr>
<td>S AUC((0-t))</td>
<td>43</td>
<td>0.03875</td>
<td>2</td>
<td>18.03</td>
<td>&lt;.0001</td>
<td>0.03202 0.83</td>
<td></td>
<td>0.03573 0.92</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\) Based on an ANOVA using Reference only
\(^2\) Method of Moments based on ANOVA using Reference only
\(^3\) Restricted Maximum Likelihood based on ANOVA using reference only