Evaluation of a Scenario in Which Estimates of Bioequivalence Are Biased and a Proposed Solution: t_{last} (Common)

Dennis Fisher, MD[1], William Kramer, PhD[2], and Elise Burmeister Getz, PhD[3]

Abstract

In bioequivalence (BE) testing, it is the convention to identify t_{last} separately for each concentration-vs-time profile. Within-subject differences in t_{last} between treatments can arise when assay sensitivity is reached during washout, causing profiles to fall below the limit of quantitation (LOQ) at different sampling times. The resulting t_{last} difference may be systematic, due to true differences in exposure, and/or random, due to measurement noise. The conventional profile-specific t_{last} approach assumes that concentrations in the terminal phase are sufficiently low that use of different t_{last} values between treatments within a subject causes negligible bias in the AUC_{0-t} geometric mean ratio (GMR). Here we investigate the validity of this assumption. Using concentration-vs-time data following oral inhalation of 50 μg salmeterol as an example data set, we conducted simulations to evaluate whether use of different test/reference AUC timeframes arising from a systematic difference in exposure causes sufficient AUC_{0-t} GMR bias to influence the determination of BE. To ensure that results would be relevant to BE testing, we considered only test/reference relative systemic exposures within the BE window (80.00%–125.00%). We show that use of conventional profile-specific t_{last} exaggerates true differences in systemic exposure; the resulting AUC_{0-t} ratios are biased from true relative exposure by an amount large enough to impact the conclusion of BE. Thus, drugs whose concentrations fall below LOQ during washout may fail BE inappropriately using conventional methods. AUC_{0-t} calculated over a common timeframe within each subject (t_{last}[common]) minimizes this bias and harmonizes the statistical analysis of BE.

Average bioequivalence (BE) of drugs requires that the 90% confidence interval (CI) for the geometric mean ratio (GMR) of each of 3 metrics (C_{max}, AUC_{0-t}, and $AUC_{0-\infty}$) for the test product (Test) vs reference product (Reference) be fully contained within 80.00%–125.00%. It is the convention to identify the final time point for the AUC_{0-t} calculation, t_{last}, separately for each concentration-vs-time profile. However, if the bioanalytical assay limit of quantitation (LOQ) is reached during the washout phase, profiles for any given subject may fall below LOQ at different sampling times, in which case the conventional approach will calculate AUC_{0-t} using a different time window for Test than for Reference. In turn, AUC_{0-t} for the treatment with the earlier t_{last} will be relatively underestimated (Table 1, Figure 1), thereby biasing the GMR for that metric (and, because $AUC_{0-\infty}$ is a function of AUC_{0-t}, also potentially biasing the GMR for $AUC_{0-\infty}$). The t_{last} differences may be systematic, due to true differences in exposure, and/or random, due to measurement noise.

Recently, Allen et al[1] proposed that limiting the timeframe for calculation of AUC_{0-t} for each subject to the earliest t_{last} obtained for that subject across all periods of a crossover trial, but permitting different timeframes for each subject, provides the most robust estimate of AUC_{0-t} for the purpose of a BE assessment. The inhaled fluticasone furoate BE data set reported by Allen et al provides an extreme example of this bias because relative systemic exposures among the treatments (129% to 160%), as judged by $AUC_{0-\infty}$, were outside the BE acceptance zone, thus causing large systematic differences in t_{last} between treatments. Further, terminal AUC trapezoids each contributed substantially to the AUC_{0-t} estimate because of infrequent terminal sampling in the context of a relatively high LOQ (only 5- to 7-fold below C_{max}). Inclusion/exclusion of such terminal AUC trapezoids due to within-subject differences in t_{last} would bias the AUC_{0-t} treatment

[1]P Less Than, San Francisco, CA, USA
[2]Kramer Consulting LLC, North Potomac, MD, USA
[3]Oriel Therapeutics, Berkeley, CA, USA

Corresponding Author:
Dennis Fisher, MD, P Less Than, San Francisco, CA 94116
Email: fisher@plessthan.com

Table 1. Reference and Test Cp Profiles for the Same Subject as Figure 1[†]

| Time (hours) | Cp (pg/mL) | |
	Reference	Test
0	0	0
0.05	106.73	96.057
0.067	111.10	99.990
0.083	103.74	93.366
0.117	86.683	78.0147
0.167	68.219	61.3971
0.33	41.215	37.0935
0.5	33.34	30.006
0.67	30.584	27.5256
0.83	29.165	26.2485
1.0	28.003	25.2027
1.5	25.141	22.6269
2	22.699	20.4291
4	15.812	14.2308
8	9.4365	8.49285
12	6.7864	6.10776
16	5.3161	4.78449
24	3.5183	3.16647
28[*]	2.8892	2.60028
32	2.3756	2.13804
36[*]	1.9541	1.75869
40	1.6076	1.44684
44[*]	1.3226	1.19034
48	1.0882	**0.97938**
52[*]	**0.89531**	**0.805779**
56	**0.73661**	**0.662949**

[†] Test values were obtained by multiplying Reference values by a factor (0.90 in this example). Boldfaced entries are samples <LOQ (1.00 pg/mL) that were excluded from calculations of AUC_{0-t} and half-life. At 48 hours (shaded row), the samples straddled the LOQ; thus, this subject was "unmatched."

[*] These samples were added to the data set to evaluate the impact of increased frequency of sampling. They were included in certain calculations, as indicated in the text.

ratios, as observed by Allen et al. Under these extreme circumstances, the need to harmonize the AUC_{0-t} time window across treatments within a subject was evident. However, it was not the objective of Allen et al to explore the implications of a bias in AUC_{0-t} GMR due to use of different time windows under conditions more common to BE studies, specifically exposure ratios within the BE window (80.00%–125.00%) and terminal AUC trapezoids that represent a small percentage of total AUC.

Because blood concentrations during washout are typically low relative to C_{max}, the conventional assumption is that bias caused by use of different t_{last} values across treatments within a subject is negligible, especially if true relative exposure between treatments is not substantially different from unity. Here, we investigate the validity of this assumption using an example drug, salmeterol administered by oral inhalation. To generate results relevant to most BE determinations, we simulated only situations in which true Test-vs-Reference differences in systemic exposure are within the BE requirement (80.00%–125.00%), and terminal AUC trapezoids are a small percentage of total AUC. We extend the analysis to consider the magnitude of error introduced to $AUC_{0-\infty}$. Finally, we explore the sensitivity of our results to the blood-sampling schedule.

Methods

Pharmacokinetic data from a clinical trial were obtained from the Novartis clinical trials library. Salmeterol (50 μg) was administered on 4 occasions to 60 healthy adult male and female subjects as a single dose by oral inhalation. The interval between treatments was
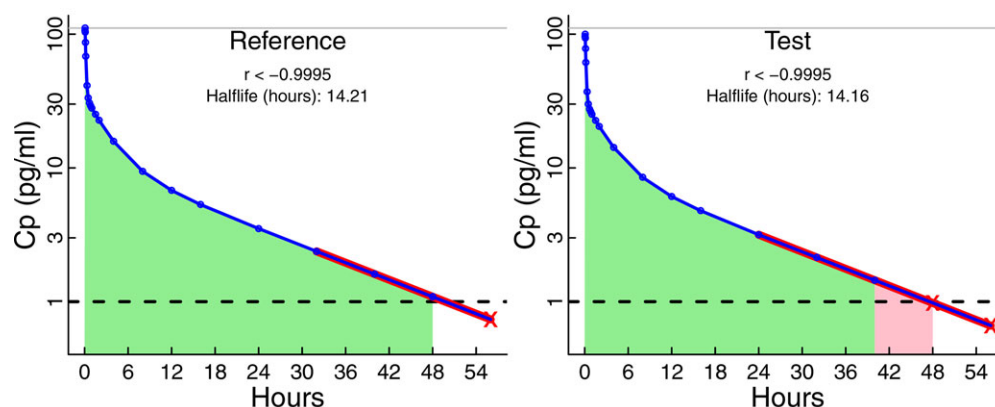
Figure 1. Hypothetical plasma concentration (Cp) profiles for Reference (left) and Test (right) products are displayed on semi-log coordinates. Both profiles are based on predicted values for 1 subject; values for Test are exactly 90% of the values for Reference. A thin horizontal line appears in both panels at C_{max} of Reference. With an LOQ of 1.00 pg/mL (dashed line), blue circles represent "observed" samples and X indicates samples reported as <LOQ. AUC_{0-t} calculations for Reference include entire area through 48 hours, whereas calculations for Test are truncated at 40 hours. As a result, the pink shaded region (hours 40–48) is included in the calculations of AUC_{0-t} for Reference but not for Test. Half-life (calculated from linear regression of the log of the final 3 samples >LOQ) was slightly longer for Reference than Test (r for the linear regression is displayed). The fit of the linear regression (extrapolated through hour 56) is shown as a thick red line.

≥7 days. Plasma was sampled at frequent intervals for 56 hours for each session; after 24 hours, the interval was 8 hours (24, 32, 40, 48, and 56 hours). Plasma concentration (Cp) was determined using a validated high-performance liquid chromatography/tandem mass spectrometry assay with an LOQ of 1.00 pg/mL. A data set suitable for a mixed-effects population analysis with NONMEM (Version 7.3.0, ICON Development Solutions, Hanover, Maryland) was assembled using R (Version 3.2.0, www.R-project.org). The data set included the first session for each subject. Actual sample times were used in the NONMEM analysis; predictions were obtained at all nominal times (and at 4 additional timepoints: 28, 36, 44, and 52 hours). Linear 2- or 3-compartment models with first-order absorption from a depot were fit to the data.

To determine the impact of true relative bioavailability deviating from unity on calculations of GMR and 90% CI for each of AUC_{0-t} and $AUC_{0-\infty}$, we performed the following steps for bioavailability factors (F) ranging from 80% to 95% in increments of 1%:

1. For each subject, the predicted Cp profile (excluding measurement error; through the sample at 56 hours) was reviewed; concentrations <LOQ were discarded. These concentrations, the corresponding sample times, and subject IDs were included in the "Reference" data set. Next, each Cp value was multiplied by F. The adjusted Cp values >LOQ, corresponding sample times, and subject IDs were included in the "Test" data set. For each subject, 2 scenarios existed. In one, in which the subject was identified as "unmatched," multiplying Reference concentrations by F caused 1 or more Test Cp values to become <LOQ (eg, if a Reference sample had a value of 1.02 pg/mL, F = 0.9 would cause the corresponding Test sample to have Cp less than the LOQ of 1.00 pg/mL). If none of the Test Cp values was <LOQ, the subject was identified as "matched" to indicate that the timeframe over which there was quantifiable data was identical for Test and Reference. Data from a representative subject are shown in Table 1.
2. For each subject, for each of the Reference and Test profiles, AUC_{0-t} (using the linear trapezoidal method), terminal rate constant (based on the final 3 time points), AUC_{tail}, and $AUC_{0-\infty}$ (the sum of AUC_{tail} and AUC_{0-t}) were calculated using standard methods. If a subject was matched, the same time points were used in the linear regression; in turn, terminal rate constant and terminal half-life were identical for Test and Reference. Because linear regressions were based on predicted Cp (without residual error), r^2 for the linear regression was always >0.95. Median

Table 2. Analyses Based on the Nominal Sampling Regimen and LOQ = 1.00 pg/mL*

True Value (%)	GMR (%) for AUC_{0-t}		GMR (%) for $AUC_{0-\infty}$	
	All	Unmatched	All	Unmatched
80	76.70	74.27	79.65	79.38
81	78.00	75.30	80.67	80.37
82	79.32	76.35	81.72	81.40
83	80.61	77.59	82.74	82.41
84	81.75	78.49	83.75	83.37
85	82.84	79.47	84.75	84.34
86	83.95	80.53	85.75	85.32
87	85.00	81.41	86.76	86.31
88	86.13	82.24	87.76	87.24
89	87.20	83.14	88.76	88.21
90	88.35	84.30	89.83	89.39
91	89.33	85.24	90.82	90.38
92	90.51	86.18	91.86	91.45
93	91.67	87.02	92.87	92.41
94	93.18	88.64	93.93	93.50
95	94.26	89.61	94.93	94.50

*GMR for AUC_{0-t} and $AUC_{0-\infty}$ for all subjects and unmatched subjects as a function of relative F. For matched subjects, GMR was identical to the true value in all instances.

values for AUC_{tail} (across F values) were typically 11% of $AUC_{0-\infty}$.
3. For each of AUC_{0-t} and $AUC_{0-\infty}$, GMR and 90%CI were calculated based on log-transformed values using a standard ANOVA model. Because data were simulated, period and sequence were not used as terms in the ANOVA model. Three sets of analyses were performed:
   a. All subjects
   b. Only matched subjects
   c. Only unmatched subjects
4. Analyses were performed with and without the additional sampling times (28, 36, 44, and 52 hours) to examine whether the findings were robust with a different sampling schedule.
5. Graphics were prepared to show:
   a. Estimated relative F vs the true value used in the simulation.
   b. Lower bounds of the 90%CI for each of AUC_{0-t} and $AUC_{0-\infty}$ vs the true value used in the simulation.

All calculations (except the initial fitting performed with NONMEM) were conducted using R.

Results

Analyses Based on the Nominal Sampling Regimen

GMR for both metrics were further from unity relative to the true value (Table 2, Figure 2); the magnitude of the difference between estimated and true
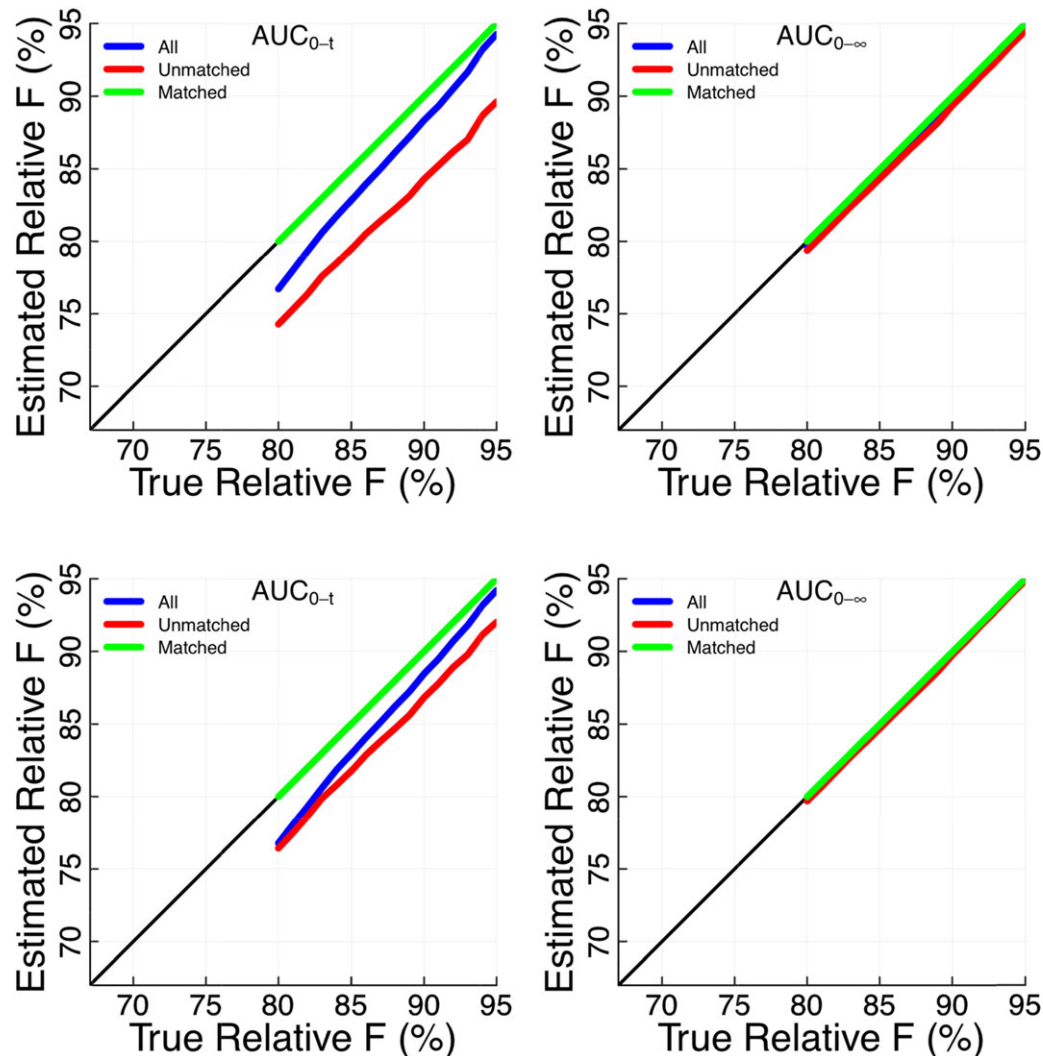
Figure 2. Values for GMR for AUC_{0-t} (left) and $AUC_{0-\infty}$ (right) are displayed against true relative bioavailability (top, nominal sampling regimen; bottom, supplemented sampling regimen). Colors distinguish all, unmatched, and matched subjects; the line of identity is displayed in black.

Table 3. Results (Expressed as %) for Unmatched Subjects and All Subjects From the Analysis Based on the Nominal Sampling Regimen*

Population	N	AUC_{0-t}			$AUC_{0-\infty}$		
		GMR	Lower Bound	Upper Bound	GMR	Lower Bound	Upper Bound
All	60	79.32	**78.61**	80.02	81.72	**81.58**	81.86
Unmatched	28	76.35	75.62	77.09	81.40	81.12	81.67

*With a relative F of 0.82 (82%) and LOQ = 1.00 pg/mL. For matched subjects (N = 32), all individual T/R ratios were identical to the true value (82.00%); therefore, the GMR was identical to the true value, and the confidence interval had zero width. The lower bounds for AUC_{0-t} and $AUC_{0-\infty}$ for all subjects (boldfaced) straddled the 80.00% acceptance value.

AUC ratios was always larger for AUC_{0-t} than for $AUC_{0-\infty}$. In subset analyses for unmatched subjects, deviations for both metrics were larger compared to the analysis that pooled matched and unmatched subjects ("all subjects"). In contrast, in subset analyses for matched subjects, GMR for both metrics matched the true value. The CIs for these metrics, although narrow

(a result of both the absence of within-subject variability and the identical true relative exposure value applied to every subject with these simulated data), reflect the larger bias from true relative bioavailability for AUC_{0-t} relative to $AUC_{0-\infty}$. With a true relative bioavailability of 82%, the lower bound of the confidence interval for all subjects was 78.61% for AUC_{0-t} and 81.58%
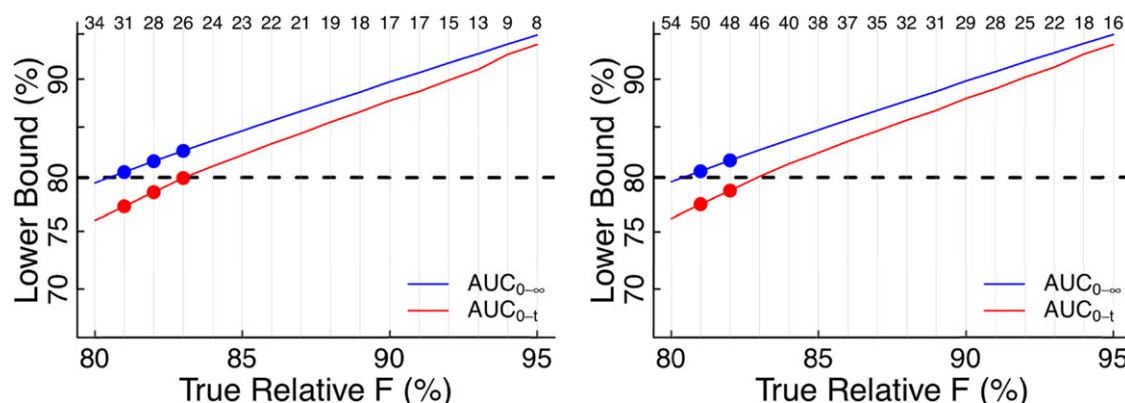
Figure 3. Panels (left, nominal sampling regimen; right, supplemented sampling regimen) display lower bounds for CI for each metric vs the true value; circles mark values for which the lower bounds straddle the acceptance criteria. The number of unmatched subjects (out of the 60 total subjects) in each analysis is shown across the top of the panel as an integer.

for $AUC_{0-\infty}$ (Table 3, Figure 3); thus, BE acceptance criteria were met for 1 metric but not for the other despite a common true value and the absence of measurement noise. This result is entirely a consequence of within-subject differences in the timeframe used for the AUC_{0-t} calculation. Similar straddling occurred with true F values of 81% and 83%.

The difference between the lower confidence bounds for AUC_{0-t} relative to $AUC_{0-\infty}$ in the "all subjects" population increased as the true relative bioavailability value deviated further from unity. This same situation applied for the subset analysis for unmatched subjects (not shown). However, consistent with estimates for matched subjects matching the true value (and with no intrasubject variability), the lower and upper bounds of the confidence interval were identical to the GMR in the "matched subjects" population (not shown).

Analyses Based on the Supplemented Sampling Regimen
Denser sampling during late washout increased the number of unmatched subjects (compare counts in the panels in Figure 3). However, deviations of GMR from true relative F for both metrics for unmatched subjects were slightly smaller compared to the nominal sampling regimen (Table S1, Table S2). Thus, because of the balance between the greater number of unmatched subjects and the smaller bias among these unmatched subjects, there was little change in the magnitude of GMR bias in the "all subjects" population with the denser sampling regimen. Lower confidence bounds were similar to those from the nominal sampling regimen. Two pairs of lower bounds straddled the 80.00% value.

Discussion

We demonstrate that the magnitude of bias incurred by use of different t_{last} values between treatments within a subject for AUC_{0-t} calculation is not negligible and

might lead to an incorrect conclusion with respect to bioequivalence. If systemic drug concentrations fall below LOQ during the washout phase, AUC_{0-t} GMR estimates of relative F exaggerate true deviations of relative F from unity. In turn, BE analyses for drugs with terminal phase concentrations near assay sensitivity may fail inappropriately. Separate evaluations of matched vs unmatched subjects identify the source of the problem: as illustrated in Figure 1, the problem with AUC_{0-t} results from its calculation over different time windows when there is systematic truncation of Cp data <LOQ at different t_{last} values. When t_{last} is matched within a subject [ie, using the earliest t_{last} obtained across all periods of the crossover trial, t_{last}(common)], there is no longer a significant bias in the GMR (or confidence interval) for AUC_{0-t}.

We agree with the Allen et al proposal that when blood/plasma concentrations fall below LOQ during the terminal phase, calculation of AUC_{0-t} to assess bioequivalence should use a single t_{last}(common) for each subject instead of the t_{last} of each individual profile. This approach should yield consistent results across pharmacokinetic BE studies, minimizing the opportunity for the BE conclusions to depend on the interaction among LOQ, sampling schedule, and active ingredient kinetics. This is particularly relevant for products for which systemic exposure is intended to be low (and, therefore, terminal phase concentrations fall below LOQ) such as orally inhaled salmeterol, used as an example here, or orally inhaled fluticasone furoate, the example offered by Allen et al, or for products with higher systemic exposure but for which current technology cannot improve the assay LOQ. We propose that the AUC BE requirement should be harmonized across all BE studies and that the t_{last}(common) approach, as presented here, meets that objective.

Although the opportunity for bias from the true exposure ratio when different time windows are used in the AUC_{0-t} calculation may be self-evident, Allen

et al provided no assessment of whether the magnitude of this bias could account for the observed difference in GMR between AUC_{0-t} and $AUC_{(0-tcommon)}$, nor did they assess the impact of this bias on a bioequivalence interpretation. We present a quantitative assessment of the magnitude of the bias, identifying the root cause—different time windows in the AUC calculations—and demonstrating that the magnitude of this bias can be sufficiently large to affect the bioequivalence outcome, always exaggerating the true exposure difference between Test and Reference.

One implication of truncating the Cp profile at t_{last}(common), compared to t_{last}, is that data are removed from the analysis, generally an anathema to regulatory bodies. However, the data that would be removed from the analysis under this approach are those closest to the assay LOQ, for which variability is often largest. Further, $AUC_{0-\infty}$ provides a logical companion metric to $AUC_{(0-tcommon)}$ in that a common time window for the AUC calculation is achieved by each of truncation ($AUC_{(0-tcommon)}$) and extrapolation ($AUC_{0-\infty}$).

Our analyses show a sharp distinction between subjects designated as matched vs unmatched. For subjects designated as matched, GMR is identical to the true value. In contrast, for subjects designated as unmatched, GMR and lower bounds of the CI deviate from the true value; however, the magnitude of deviation is markedly larger for AUC_{0-t} compared to $AUC_{0-\infty}$ because the mismatch in the time window in the AUC_{0-t} calculation is counterbalanced by the extrapolation to infinite time in $AUC_{0-\infty}$. In turn, if the number of unmatched subjects is large, AUC_{0-t} may be sufficiently biased relative to the true value to cause a BE study to fail erroneously. In contrast, the error in $AUC_{0-\infty}$ is relatively small and less likely to cause a BE study to fail.

The critical element in our calculations is the AUC in the pink region in Figure 1. Analyses in which we simulated additional samples during washout examined 2 issues:

1. Denser sampling might decrease the width of that pink region.
2. Some subjects previously designated as matched might become unmatched.

On balance, adding samples during washout had minimal impact on the deviations between the lower bounds of the 2 metrics.

Having established that the conventional AUC_{0-t} calculation leads to a biased GMR estimate, we note that the GMR and CI for both metrics were identical to the true values for matched subjects. This supports our claim that calculating both metrics over the same time interval mitigates biases introduced when AUC_{0-t} is calculated over different timeframes for each of 2 treatments. Although removing data from the analysis can introduce bias to the metrics themselves, in this instance the imbalance in data in unmatched subjects creates a bias in the treatment ratio that is large enough to alter a BE determination. Had we applied a common t_{last} [which we term t_{last}(common)] to each subject, the estimates of AUC_{0-t} would each be biased, but to the same extent; in turn, their ratio (the value of interest in a BE study) would be unbiased under the circumstances of the present analysis. In a real-world situation that includes both measurement error and within-subject variability in both relative F and the shape of the concentration profile, factors in addition to differences in timeframe also influence calculation of the AUC. In these circumstances, the t_{last}(common) approach mitigates bias introduced by differences in timeframe and should be considered an appropriate approach.

Two alternative approaches might minimize the opportunity for subjects to be mismatched, thus minimizing the bias between true and estimated AUC GMRs. The first involves truncating sampling at a time when all samples are likely to be >LOQ. Although this might assure that all subjects are matched, it might provide biased estimates for the AUC metrics if absorption and distribution are incomplete. The second approach is to administer a higher-than-marketed dose, yielding higher Cp values and increasing the duration over which the Cp profile is > LOQ. If this decreased the number of unmatched subjects, it might eliminate the problem that Allen et al identified. However, accepting the results of such an analysis requires the assumption that data from a higher-than-marketed dose are relevant to the marketed dose.

Several issues of our analysis plan warrant comment. First, we used predicted values obtained from a compartmental model (and obtained predictions at nominal times to ensure that each subject's sample times were identical). Use of predicted values eliminated noise prevalent with observed concentrations; this assured that assay and/or physiologic variability did not impact the last quantifiable value or estimation of terminal half-life. The linear regression used for calculation of the terminal rate constant (and, in turn, AUC_{tail} and $AUC_{0-\infty}$) yielded coefficients of determination close to unity (with the nominal sampling regimen, r^2 was always ≥ 0.97). In addition, this approach assured that the AUC_{tail} was a small fraction of $AUC_{0-\infty}$ (with the nominal sampling regimen, median values were typically 11%).

We calculated the Test Cp profile by multiplying the Reference Cp profile by a factor. This resulted in the 2 profiles having identical shapes (particularly evident in

graphics in the log domain). This approach minimized intrasubject variability, as evidenced by the complete lack of intrasubject variability in matched subjects. In turn, the confidence intervals were unrealistically small, their magnitude depending on the fraction of unmatched vs matched subjects. One consequence of these artificially small CIs is that the GMR associated with the CI for the 2 AUC metrics straddling the 80.00% BE criteria was close to 80.00%. With "real" data, it is possible that a GMR close to, or even exceeding, 90% could be associated with CIs for the 2 AUC metrics that straddle 80.00%. Had our analyses incorporated more within-subject variability, we could have increased the width of the CI to that typically seen with experimental data. Regardless, the present approach demonstrates the impact on both the GMR and CI for each metric.

Finally, we present results for scenarios with relative $F < 100\%$. Additional simulations (not reported) confirm a similar deviation with relative $F > 100\%$.* This is expected because BE analyses (which are conducted in the log domain) are symmetric with respect to 100% in the log domain (eg, relative F of 90.00% for Test vs Reference is identical to relative F of 111.11% for Reference vs Test).

Several investigators have examined the impact of duration of sampling on noncompartmental estimation of $AUC_{0-\infty}$. El-Tahtawy et al[2] concluded that sampling should not extend excessively. Colucci et al[3] concluded that the sampling regimen should cover at least 2-4 half-lives. However, neither of these investigators examined the discrepancy between AUC_{0-t} and $AUC_{0-\infty}$.

Our findings support the decision of Allen et al to apply a t_{last}(common) approach for the calculation of AUC_{0-t} geometric mean ratios and confirm that this approach is both necessary and sufficient to address the otherwise inherent bias in the AUC_{0-t} GMR for drugs that fall below LOQ during washout. The t_{last}(common) approach serves to harmonize the metrics used for the statistical analysis of pharmacokinetic BE across all BE studies. It remains to be seen if regulatory bodies will consider this approach.

Declaration of Conflicting Interests

References

1. Allen A, Bal J, Moore A, Stone S, Tombs L. Bioequivalence and dose proportionality of inhaled fluticasone furoate. *J Bioequiv Availab.* 2014;6:24–32.
2. El-Tahtawy A, Harrison F, Zirkelbach JF, Jackson AJ. Bioequivalence of long half-life drugs—informative sampling determination—parallel designed studies. *J Bioequiv Availab.* 2011;3:56–61.
3. Colucci P, Turgeon J, Ducharme MP. How critical is the duration of the sampling scheme for the determination of half-life, characterization of exposure and assessment of bioequivalence. *J Pharm Pharmaceut Sci.* 2011;14:217–226.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website.

---

*This requires that the Reference values for at least 1 subject reach LOQ before the 56-hour sample. If the 56-hour Reference samples for all subjects were >LOQ and F is >100%, all subjects would be matched, and timeframe would be identical for all subjects.