Bioequivalence: Performance of Several Measures of Extent of Absorption

Frédéric Y. Bois,[1,5] Thomas N. Tozer,[2] Walter W. Hauck,[3] Mei-Ling Chen,[4] Rabindra Patnaik,[4] and Roger L. Williams[4]

Received September 14, 1993; accepted December 14, 1993

The determination of the area under the concentration–time curve (AUC) is the method most commonly used by regulatory agencies to assess extent of drug absorption after single-dose administration of oral products. Using simulations, several approaches toward measuring the actual area, in whole or part, were tested. In addition, the performance of the peak concentration (C_{max}), usually taken as a measure of the rate of absorption was assessed evaluating extent. Model scenarios for drugs with typical mean characteristics and statistical distributions were investigated. Using different kinetic models of disposition, the time course of the drug concentration in plasma was simulated. Intraindividual and interindividual variability and assay error were modeled using Monte Carlo techniques. The accuracy, precision, and ease of use of the various measures of extent were evaluated, and statistical power analyses were performed. Among the measures tested, the most reliable were the AUC computed up to the time of the last quantifiable concentration, without extrapolation, and C_{max}. However, being also sensitive to rate, C_{max} as a measure of extent is of limited potential.

KEY WORDS: pharmacokinetics; bioequivalence; extent of absorption; power analysis.

INTRODUCTION

In the United States the Food and Drug and Cosmetic Act, section 505(j)(7) of the Federal Register, mandates that the U.S. Food and Drug Administration ensure bioequivalence of new formulations to established products. Bioequivalence is demonstrated when the rate and extent of absorption of the drug are sufficiently similar for the two formulations when administered under similar experimental conditions.

The total area under the plasma concentration curve (AUC) is the most common method for assessing extent of drug absorption. In practice, estimation of area is limited by assay sensitivity, particularly when a large area exists beyond the last detectable concentration–time point. This unobservable area can be large when the peak concentration is not much above the analytical limit of quantification (LQ). Several measures may be devised to circumvent this extrapolation problem. For example, it has been proposed to use AUC_{lqc}, the area under the curve only to the time of the last quantifiable concentration (LQC) in each individual, or to use the data only to the last time (tfix) for which all subjects in the study have measurable concentrations (leading to the AUC_{tfix} variant).

One of the best-known measures of rate of absorption, C_{max}, also reflects changes in extent. We have included it in this analysis for comparison with the more traditional AUC approaches.

We examined, through simulations, the ability of various measures to assess bioequivalence. Using the conditions of a clinical bioequivalence trial, eight scenarios of absorption and disposition kinetics were investigated. The scenarios incorporate the following: ratios of absorption and elimination rate constants of 0.25 and 4, zero-order and first-order absorption kinetics, limits of quantification of 1 and 10% of the mean peak concentration, presence of a lag-time, highly variable first-pass elimination, and one- and two-compartment distribution characteristics. The time course of the concentration in the plasma was simulated with intra- and interindividual variability and assay error modeled using Monte Carlo techniques. For each scenario, the extent of absorption was assessed by each of the different measures investigated.

Computer simulations allowed us to control all of the model parameters while conducting a typical bioequivalence trial. Known differences between test and reference formulations and known levels of variability were introduced. The ability of given measures to uncover the true underlying differences in extent was thereby assessed. The reliability of the various measures of extent was evaluated and power analyses (1) were performed to examine in detail the most promising ones. A separate paper will address the issues, somewhat more complex, associated with measuring the *rate* of drug absorption.

METHODS

Simulation Framework

To evaluate the various measures of extent of absorption, it was necessary to simulate the variability typically encountered in humans. To accomplish this, Monte Carlo simulations were used to generate data sets to which the measures were applied. In Monte Carlo simulations, statistical distributions, rather than fixed values, are assigned to the various parameters of a model. The distributions of the parameter values were either truncated normal or uniform. The uniform random numbers and normal deviates were generated using the algorithms described by Park and Miller (2) and Press *et al.* (3), respectively. Therefore, under each set of conditions investigated, a distribution for the measures of extent was obtained as expected in a series of real clinical trials.

[1] Indoor Environment Program, Mail Stop 90-3058, Lawrence Berkeley Laboratory, Berkeley, California 94720.
[2] School of Pharmacy, University of California, San Francisco; currently: Division of Clinical Pharmacology, Thomas Jefferson University, Philadelphia, Pennsylvania 19107-5563.
[3] Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California 94143–0560.
[4] Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Office of Generic Drugs, 7500 Standish Place (HFD-600), Rockville, Maryland 20855–2737.
[5] To whom correspondence should be addressed.

0724-8741/94/0500-0715$07.00/0

In all scenarios, 1540 clinical trials were simulated. This number was chosen to ensure that our simulations would yield accurate power estimates. To be specific, we wanted to obtain a 95% confidence interval of width no more than ±0.025 on the value of a proportion p, such as a power estimate, assuming the binomial distribution. Each trial was a crossover design with 24 subjects and two drug formulations (test, T, and reference, R). A sequence effect, with average zero, was introduced by randomly assigning the subjects to two groups of 12.

Addition of Variability

A standard statistical model of errors and variabilities in population pharmacokinetics (4) was used to simulate the trials (Fig. 1). From population distributions, D_p, a set of pharmacokinetic parameter values P_i was sampled for subject i. At each trial period j, intraindividual variability V_j was added to the subject's baseline values, forming the new parameters P_{ij}. These parameters were assumed to remain constant over a trial period. Two periods were simulated, during which the two formulations X_j were administered. The difference between the two formulations was introduced by changing the value of the extent of absorption for each subject's test period. These changes amounted to a fixed fraction of the mean population value and were the same for each individual. Assay error, E, was added to the plasma drug concentrations, C_{ijk}, given by the model (represented by the triangle) at times T_k. The various variabilities were taken to be the same for the two formulations (i.e., there was no subject by formulation interaction).

Individual plasma concentration values at the defined sampling times were simulated using specific interindividual and intraindividual distributions for each parameter (Tables I to IV). A body weight of 70 kg was assumed in all cases. To avoid unrealistic values, all normal distributions were truncated: random deviates falling beyond ±3 standard deviations were resampled. Such truncation occurs for only 0.3% of normal deviates. Intraindividual variability was simulated by adding deviations with a mean of zero, to the parameter values previously sampled. For example, for a given individual, a k_a value of 1.31 is first sampled (out of the population distribution). During the first period (reference formulation administered) intraindividual variability (sampled from a normal distribution with mean zero, SD 0.2 × 1.31, and bounds ±3 × 0.2 × 1.31) is added to the value 1.31. For example, when −0.1 is sampled, the k_a becomes 1.21. The reference plasma concentration curve is then computed using this value. During the second period a new normal deviate is sampled from the same intraindividual distribution and is added to the value 1.31; if a 25% systematic difference in k_a exists between mean test and reference values, then 0.3275 (i.e., 0.25 × 1.31) is added to calculate the test plasma concentration curve.

Analytical assay errors were generated from truncated normal distributions with no bias (mean zero), a CV of 10%, truncation at ±3 CV, plus a fixed term equal to the product of the assay CV and the limit of quantification, LQ (the concentration below which the analytical error was 20% or more). LQ was defined as a fraction of the theoretical C_{peak} computed with the mean interindividual parameter values (Table I). The same LQ was used in all simulations for a given scenario and did not vary from individual to individual.

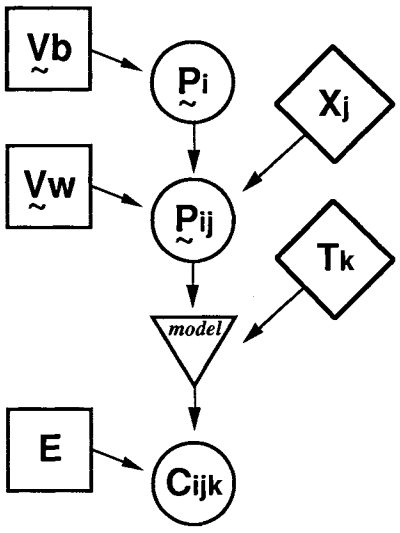

Fig. 1. Structure of the statistical model used to simulate the outcome of a given clinical trial. From the population distributions, V_b, a set of pharmacokinetic parameter values P_i is sampled for subject i (24 subjects are simulated per trial). At each period j, intraindividual variability V_w is added to the subject's baseline values, forming the new parameters P_{ij}. Two periods are simulated, during which two formulations X_j are administered. Assay error, E, is added to the plasma drug concentrations, C_{ijk}, predicted by the pharmacokinetic model (triangle) at times T_k.

Scenario Definitions

The pharmacokinetic models and parameter distributions used were grouped in eight scenarios which reflect situations commonly encountered, or of special interest, when testing bioequivalence.

Baseline Simulation Scenario

The baseline scenario had the following characteristics: one-compartment distribution kinetics; first-order absorption and elimination; a ratio of absorption to elimination rate constants of 4; no lag time; and a detection limit equal to 1% of the theoretical mean peak concentration, C_{peak}, for the reference formulation after oral administration.

Acetaminophen was the model drug for this scenario. For all studies, an arbitrary oral bolus dose of 500 mg was used (as only linear kinetics are investigated, the conclusions are not affected by dose). In this scenario, simulated sampling times were 0, 0.25, 0.5, 1, 1.5, and 2 hr and then every 2 hr up to 16 hr. This schedule provided an average of four to six observations in the upswing part of the curve, as well as near the peak and in the decline phase.

Alternative Scenarios

A set of alternative scenarios was examined. In each case only differences from the baseline scenario are described.

Table I. Distribution Type, Mean, Coefficient of Variation (CV) and Truncation for Interindividual Parameters in the One-Compartment Baseline Scenario

Parameter	Distribution	Mean	CV (%)	Truncation
Volume of distribution, V	Normal	$1\ L \cdot kg^{-1}$	10	±3 SD
Clearance, CL	Normal	$0.347\ L \cdot hr^{-1} \cdot kg^{-1}$	20	±3 SD
Absorption rate constant, k_a	Normal	$1.39\ hr^{-1}$	20	±3 SD
Bioavailability, F	Uniform	0.5	11.5[a]	0.4–0.6

[a] This value is equal to $100 \times (0.6 - 0.4)/(0.5 \times \sqrt{12})$ since the distribution is uniform.

Low Sensitivity

This corresponds to a situation in which one cannot follow the concentration over a wide range. The LQ was set at 10% (instead of 1%) of the mean interindividual C_{peak} for the reference formulation.

Zero-Order Absorption

To keep the mean input time identical between the zero-order and the first-order input cases, the duration of input (infusion time) was set to $2/k_a$, where k_a is the absorption rate constant in the baseline scenario (Table I). The coefficient of variation of the duration of input had the same CV and truncation as the baseline k_a.

Presence of a Lag Time

A random lag time of an hour, on average, was introduced for both the test and the reference formulations. A normal distribution with a CV of 50%, truncated to ±2 CV (hence a range from 0 to 2 hr), was used. Sampling times were identical to those in the baseline case.

Low Absorption/Elimination Ratio ("Flip-Flop")

The ratio of absorption/elimination rate constants was fixed at 0.25. To accommodate the slower rise to the peak, the simulated sampling times were then 0, 1, 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, and 48 hr.

Low F

This scenario corresponds to the situation of high first-pass elimination. The extent of absorption, F, was sampled following a uniform distribution with a range of 0.05 to 0.15 (mean, 0.1) for interindividual variability and a range of −0.05 to +0.05 for intraindividual variability. Consequently, the interindividual coefficient of variation was 30%, compared to 11.5% in the baseline case. LQ was maintained at the same value as in the baseline case.

Two-Compartment Models

Two two-compartment distribution kinetic models with first-order input into and elimination from the central compartment were studied. In the first model the elimination-to-distribution ratio k_{10}/k_{21} was fixed at 2.5 (model I), and in the second set at 0.4 (model II), k_{10} being equal to CL/V. The ratio k_{12}/k_{21} was 4 in both cases. Tables III and IV give the interindividual and intraindividual distributions used for the parameters of these models. In the context of multicompartment kinetics one's ability to extrapolate AUC correctly is diminished and the LQ becomes an important issue. For each model, power analysis at LQ values of 1 and 10% of the reference C_{peak} were investigated. In both cases the simulated sampling times were the same as in the baseline scenario.

Measures of Extent

The following measures of extent were evaluated.

AUC_{infl}

The total AUC was computed using the analytic formulae and parameter values for a one-compartment model fitted to all observable data points (i.e., above LQ) for a given individual. Note that the application of this method requires the explicit choice of a pharmacokinetic model. The parameters were fitted by least-squares minimization using the Fletcher–Reeves–Polack–Ribiere conjugate gradient algorithm (3).

AUC_{inf2}

In this method the total AUC was estimated by the sum of the areas of the trapezoids of successive data pairs, up to the last observable point, with the remaining area estimated by a one-compartment model fitted as above.

Table II. Distribution Type, Mean, Coefficient of Variation (CV), and Truncation for Intraindividual Parameters in the One-Compartment Baseline Scenario

Parameter	Distribution	Mean	CV (%)	Truncation
Volume of distribution, V	Normal	vps[a]	10	±3 SD
Clearance, CL	Normal	vps	20	±3 SD
Absorption rate constant, k_a	Normal	vps	20	±3 SD
Bioavailability, F	Uniform	vps	—[b]	±0.1

[a] Parameter value previously sampled (for a given individual) (see text).
[b] The CV varies between 10% and 15%, depending on vps.

Table III. Distribution Type, Mean, Coefficient of Variation (CV), and Truncation for Interindividual Parameters in the Two-Compartment Models I and II

Parameter	Distribution	Mean	CV (%)	Truncation
Volume of distribution, V_1	Normal	1 L/kg	10	±3 SD
Clearance, CL	Normal	0.5 L/(hr × kg)	20	±3 SD
Absorption rate constant, k_a	Normal	2.0 hr^{-1}	20	±3 SD
Central to peripheral distribution rate constant, k_{12}	Normal	0.2 hr^{-1} (model I) 1.25 hr^{-1} (model II)	20	±3 SD
Peripheral to central distribution rate constant, k_{21}	Normal	0.05 hr^{-1} (model I) 0.3125 hr^{-1} (model II)	20	±3 SD
Bioavailability, F	Uniform	0.5	11.5[a]	0.4–0.6

[a] This value is equal to $100 \times (0.6 - 0.4)/(0.5 \times \sqrt{12})$ since the distribution is uniform.

AUC_{inf3}

AUC_{inf3} was the sum of the areas up to the last observable point, as for AUC_{inf2}, with the remaining area estimated from a simple exponential passing through the last observable data point. The rate constant of the exponential was obtained by least-squares fitting of a straight line to the last four observable data points, after log transformation.

AUC_{inf4}

AUC_{inf4} was a variant of AUC_{inf3} in which the remaining area was obtained by dividing the estimated concentration at the last detectable point, rather than the observed value, by the exponential rate constant.

AUC_{lqc}

AUC_{lqc} was defined as the area under the curve up to the time of the last quantifiable concentration for a given individual and a given administration period. This method, which uses the trapezoidal method, does not require extrapolation.

AUC_{tfx}

This estimate was computed by the trapezoidal method up to the last common time point for which a quantifiable concentration was found for all members of the simulated study group for a given administration period. It is a variant of AUC_{lqc}.

C_{max}

C_{max} was simply the highest recorded concentration of a given concentration-time curve.

Reliability Analysis

Two sets of simulations were performed for each scenario. In the first, the same drug formulation was readministered to each individual. This set of simulations provides a "null" distribution for the difference in the extent of absorption between two formulations and tests the ability of the various measures to show bioequivalence.

In the second set, the administration of two drug formulations, differing in extent by 25% ($F_{test}/F_{reference} = 1.25$) was simulated. This difference corresponds to the statistical null hypothesis for bioequivalence, for which measures of extent of the two formulations are considered equivalent if within 80 and 125%. The ability of the various measures to show the 25% difference between formulations was then evaluated using this simulation set.

Power Analysis

Sets of simulations were performed to determine the relationship between the statistical power of the procedures examined (i.e., the probability of rejecting the null hypothesis of bioequivalence in a clinical trial) and the difference in extent between the test and the reference formulations. For a series of values of test/reference bioavailability ratio, 1540 clinical trials were generated by Monte Carlo simulations. In each trial the 24 subjects were randomly segregated into two groups to simulate sequence effects and two one-sided t tests (90% confidence interval) were performed after log transformation of the measures of extent (5,6). In the case of missing data (e.g., nonconvergence of the fitting algorithm for AUC_{inf1}), the corresponding individual was dropped from

Table IV. Distribution Type, Mean, Coefficient of Variation (CV), and Truncation for Intraindividual Parameters in the Two-Compartment Models I and II

Parameter	Distribution	Mean	CV (%)	Truncation
Volume of distribution, V_1	Normal	vps[a]	10	±3 SD
Clearance, CL	Normal	vps	20	±3 SD
Absorption rate constant, k_a	Normal	vps	20	±3 SD
Central to peripheral distribution rate constant, k_{12}	Normal	vps	20	±3 SD
Peripheral to central distribution rate constant, k_{21}	Normal	vps	20	±3 SD
Bioavailability, F	Uniform	vps	—[b]	−0.1 to 0.1

[a] Parameter value previously sampled (for a given individual) (see text).
[b] The CV varies between 10% and 15%, depending on vps.

the trial. The fraction of trials in which bioequivalence was declared was recorded. This fraction corresponds to the probability of declaring bioequivalence given a typical design of a clinical trial, human and analytical variability, data treatment procedures, and statistical analyses.

With no difference between the test and the reference formulations, one would want to conclude bioequivalence in 100% of the trials (low producer risk, which is 1 minus the power). In contrast, with a 25% increase or a 20% decrease in bioavailability, values currently used in regulatory practice, bioequivalence should be declared in no more than 5% of the trials (i.e., a 5% consumer risk). This also means that when comparing the 90% confidence intervals to an (80%, 125%) equivalence criterion, one expects the probability of declaring bioequivalence for the case of a 25% increase in F to be no more than 5% for any measure of extent proportional to F. The ability to meet these criteria depends on the quality of the measure of extent of absorption used.

The results are presented in the form of power curves, in which the x axis is the test/reference bioavailability ratio used in the simulations. The y axis, each point of which was obtained from 1540 simulated trials, is the corresponding probability of declaring bioequivalence when assessed by two one-sided t tests using a given measure of extent.

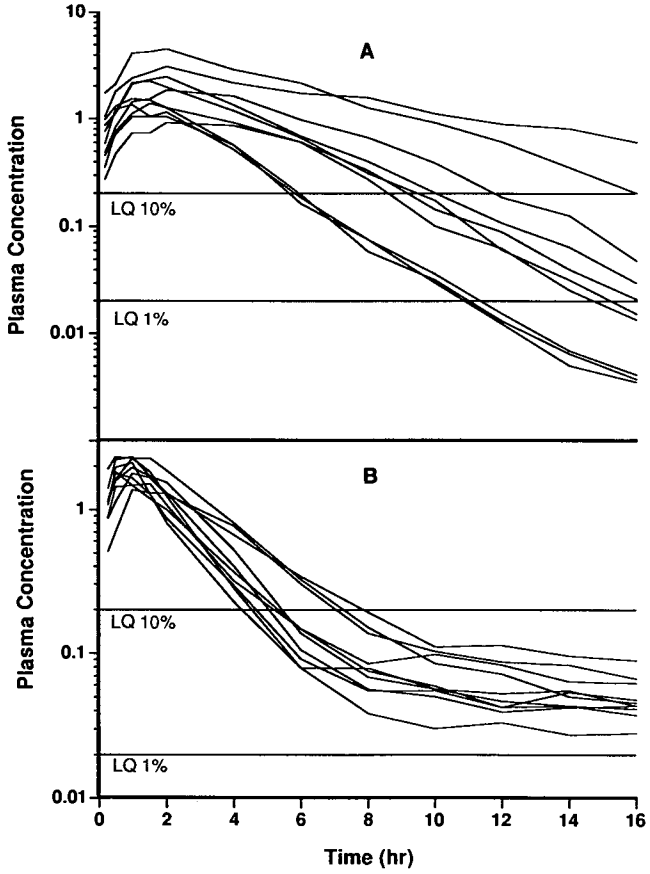

Fig. 2. Simulated plasma concentration–time curve of a hypothetical reference drug in a random sample of 10 subjects. (A) Baseline scenario (see text and Tables I and II). (B) Two-compartment model I (see alternative scenarios section). LQ is the limit of quantitation (either 1 or 10% of the theoretical peak concentration of the drug).

RESULTS AND DISCUSSION

Simulated Variability

Figure 2A illustrates the variability of individual concentration–time curves in the case of the baseline scenario, while Fig. 2B addresses the case of the two-compartment model I (elimination > distribution). In both cases the reference formulation was administered and the parameter values were sampled as given in Tables I to IV.

Because baseline values for each individual were first sampled from the interindividual parameter distributions and, for each period of the trial, intraindividual variability was introduced, the overall population variability is somewhat higher than that presented in Table I. For a sample of 500 subjects, observed twice, the CV of k_a in both cases is 28%, while the interindividual CV (i.e., the CV of the means of the two individual observations) is 20% (as prescribed in Table I), and the intraindividual CV (i.e., the average CV of the observation pairs for a given individual) is 17% (nominally 20% in Table I). For bioavailability, F, these numbers are 17, 12 (nominally 11.5%), and 9% (nominally 6%), respectively.

Figure 3 presents the distributions of C_{max} for 500 simulated random subjects. This distribution spans a factor of 4 and is skewed. Variability is therefore quite large and comparable to that observed experimentally. Product variability (e.g., random differences between dosing across tablets) was not explicitly simulated. This source of variability is included in F and k_a variances.

Reliability Analysis

Table V summarizes the distribution of percentage differences of the mean measures in the test and reference groups $100 \times (T - R)/R$ over 1540 simulated clinical trials of size $N = 24$. The test and reference formulations have, on average, the same rate and extent characteristics. Thus, the relative differences should be centered on zero. This is almost always the case, except when more than one compart-

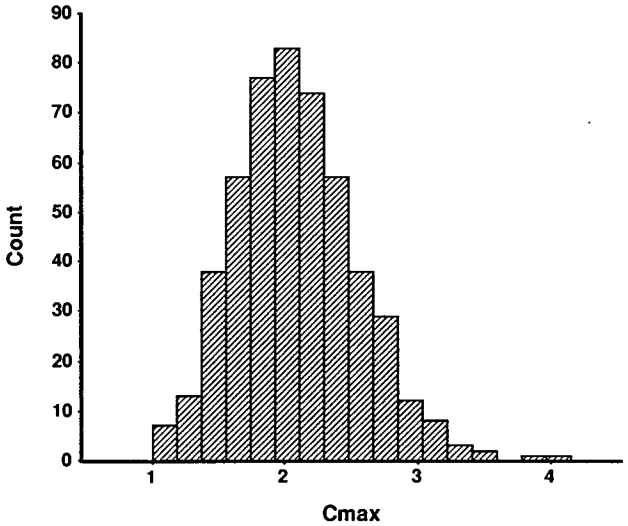

Fig. 3. Frequency distribution of simulated C_{max} values for 500 random subjects given the reference formulation in the conditions of the baseline scenario (see text).

Table V. Mean ± SD of the Percentage Difference Between the Average of Each Extent Measure for Test and Reference Groups, When Formulations Are Bioequivalent[a]

Scenario	Measure						
	AUC_{inf1}	AUC_{inf2}	AUC_{inf3}	AUC_{inf4}	AUC_{lqc}	AUC_{tfix}	C_{max}
Baseline	0.039 ± 6.7	0.013 ± 6.6	−0.0026 ± 6.7	−0.0087 ± 6.7	−0.0038 ± 6.1	0.043 ± 8.1	−0.10 ± 4.2
Low sensitivity	0.53 ± 6.9	0.53 ± 6.8	0.51 ± 6.7	0.49 ± 6.6	0.53 ± 6.8	1.2 ± 18	0.16 ± 4.1
Zero-order absorption	0.13 ± 6.5	0.11 ± 6.6	0.13 ± 6.8	0.13 ± 6.8	0.10 ± 6.2	0.26 ± 8.7	0.080 ± 4.3
Lag time	0.14 ± 6.7	0.14 ± 6.5	0.16 ± 6.9	0.16 ± 6.9	0.089 ± 6.2	0.12 ± 8.9	0.022 ± 4.9
"Flip-flop"	0.30 ± 6.9	0.29 ± 6.8	0.27 ± 6.9	0.28 ± 7.1	0.24 ± 6.8	0.43 ± 7.6	0.073 ± 6.8
Low F	0.48 ± 7.1	0.48 ± 7.0	0.52 ± 7.1	0.51 ± 7.1	0.46 ± 6.8	1.4 ± 14	0.23 ± 4.0
Two-compartment I (elim. > distr.)	26 ± 130	26 ± 130	4.9 ± 87	5.0 ± 86	−0.0014 ± 5.3	−0.0043 ± 9.1	0.022 ± 4.4
Two-compartment II (elim. < distr.)	260 ± 790	260 ± 780	11 ± 190	11 ± 200	0.14 ± 5.3	0.14 ± 5.4	0.31 ± 5.3

[a] Percentage differences were obtained from 1540 simulated clinical trials. For a given measure, the percentage difference between test and reference is $100 \times (X_T - X_R)/X_R$, X being the average value of the measured extent of absorption across individuals.

ment is involved. In this case, AUC_{inf1} and AUC_{inf2} show poor behavior, which can be expected since they are based on the assumption of a one-compartment model (i.e., the wrong model). AUC_{inf3} and AUC_{inf4}, which use a log-linear extrapolation to estimate the unobservable area also perform poorly in case II (elimination < distribution): A 10% difference in extent between test and reference is the average result for a clinical trial. Trials of size 24 can even lead to very erroneous conclusions; a 200% difference would not be rare (it corresponds to one SD), when in fact test and reference are identical. The probability of a correct conclusion could be improved by increasing the number of subjects but a more reliable measure is an obvious advantage. AUC_{lqc} and AUC_{tfix}, which use only the quantifiable concentration points and do not incorporate extrapolation procedures, perform quite well in all cases. However, AUC_{lqc} is always better than AUC_{tfix}, in terms of both mean and SD, particularly when the assay is not sensitive (10% LQ). This behavior is expected since AUC_{tfix} uses less information by truncating all curves at an earlier time point. In terms of bias, C_{max} is practically equivalent to AUC_{lqc} (only small deviations from zero occur). Furthermore, the SDs associated with C_{max} are always the lowest.

In Table VI similar data are presented for the case in which the extent of absorption for the test formulation is systematically 25% higher than for the reference. Good measures of extent should differ by 25% in most clinical trials. This is true in most cases, except again when AUC_{inf1}, AUC_{inf2}, AUC_{inf3}, and AUC_{inf4} are applied to multicompartment kinetic data. SDs are even larger than in Table V. AUC_{lqc} is still consistently better than AUC_{tfix}. In the case of the one-compartment model, AUC_{tfix} was the worst of the area measures tested. C_{max} is the best measure in terms of both bias and dispersion. However, in additional simulations in which extent was 25% lower and rate 25% higher for the test than for the reference formulation, C_{max} differed by only 20% on average between the two groups. The AUC measures did not exhibit such a drop in sensitivity.

Power Analyses

Power analyses are presented for AUC_{inf4}, AUC_{lqc}, and C_{max}, the best-performing measures, for all scenarios. The analyses were computed only for F_T/F_R ratios greater than 1.

In the baseline scenario (Fig. 4A), it appears that C_{max} has the best behavior. In the case of perfect bioequivalence ($F_T/F_R = 1$), two one-sided t tests applied to C_{max} data almost always lead to the correct conclusion, i.e., the producer risk (risk for an actually bioequivalent test formulation to be rejected) is nearly zero. For a 25% difference in extent ($F_T/F_R = 1.25$) there is only a 5% chance of declaring bioequivalence for all three measures. This probability of declaring bioequivalent two formulations which actually differ by the maximally permitted extent of absorption is the consumer risk, the only one regulated. The consumer risk, when F_T/F_R is 1.25, should be exactly 5% if the assumptions (for example, normality of the distributions, direct proportionality between the measure and the extent of absorption) made by the t tests are satisfied. AUC_{inf4} and AUC_{lqc} have less satisfying behaviors: they lead to a 5 to 10% producer risk. Overall, AUC_{lqc} has a slightly better behavior than AUC_{inf4}. The converse is true in the case of low assay sensitivity (Fig. 4B). AUC_{lqc} has less power, probably because fewer data points are available to establish it. Both AUC_{lqc} and AUC_{inf4} have decreased power in the case of low sensitivity. The nominal level of the t test (5%) is not exactly

Table VI. Mean ± SD of the Percentage Difference Between the Average of Each Extent Measure for Test and Reference Groups, When Test and Reference Differ in Extent by 25%[a]

Scenario	Measure						
	AUC_{inf1}	AUC_{inf2}	AUC_{inf3}	AUC_{inf4}	AUC_{lqc}	AUC_{tfix}	C_{max}
Baseline	25 ± 8.7	25 ± 8.6	25 ± 8.6	25 ± 8.6	25 ± 7.8	27 ± 10	25 ± 5.3
Low sensitivity	25 ± 8.9	25 ± 8.8	25 ± 8.8	24 ± 8.6	27 ± 9.0	35 ± 23	25 ± 5.2
Zero-order absorption	25 ± 8.0	25 ± 8.2	25 ± 8.5	25 ± 8.5	25 ± 7.9	27 ± 11	25 ± 5.4
Lag time	24 ± 8.3	25 ± 9.6	25 ± 8.9	25 ± 8.9	25 ± 8.0	27 ± 11	25 ± 6.2
"Flip-flop"	25 ± 8.9	25 ± 8.8	25 ± 8.8	25 ± 9.1	25 ± 8.6	27 ± 9.8	25 ± 8.4
Low F	25 ± 8.7	25 ± 8.6	25 ± 8.6	25 ± 8.6	27 ± 8.3	32 ± 17	25 ± 5.3
Two-compartment I (elim. > distr.)	59 ± 160	59 ± 160	38 ± 130	38 ± 130	25 ± 6.8	30 ± 12	25 ± 5.5
Two-compartment II (elim. < distr.)	66 ± 880	66 ± 880	45 ± 400	45 ± 400	25 ± 6.6	25 ± 6.6	25 ± 6.5

[a] Percentage differences were obtained from 1540 simulated clinical trials. For a given measure, the percentage difference between test and reference is $100 \times (X_T - X_R)/X_R$, X being the average value of the measured extent of absorption across individuals.

respected: The actual estimate is 4% when AUC_{lqc} is used and 7% for AUC_{inf4}, but only 7% is statistically different from 5%. Statisticians prefer the 4% of AUC_{lqc} (a conservative test) to the 7% of AUC_{inf4}, as the latter does not protect the consumer at the nominal 5% level. This result could indicate that even after log transformation, the assumptions made by the t test are not satisfied by the distributions of these measures (there is, however, some uncertainty in these results since only 1540 trials were simulated). C_{max} has the highest power and is not affected by the change in LQ.

In the presence of zero-order absorption, lag time or low bioavailability, conclusions are similar to those in the baseline scenario. C_{max} has a higher power than AUC_{inf4} or AUC_{lqc}, which behave very similarly (data not shown).

Figure 5 shows that in the case of "flip-flop," C_{max} may have a slightly worse behavior than AUC_{inf4} or AUC_{lqc}, which behave similarly. For the three measures, the producer risk is about 10%, while the 5% nominal level is maintained when F_T/F_R is 1.25. In the case of flip-flop, the peak occurs later and is less precisely measured (although the sampling schedule was adjusted), which explains the poorer performance of C_{max}.

Figure 6A shows the statistical power of C_{max}, AUC_{inf4}, or AUC_{lqc} when applied to two-compartment kinetic data for which elimination dominates over distribution (model I). C_{max} and AUC_{lqc} behave well, while AUC_{inf4} has very low power, as expected from the results presented in Tables V and VI. In the case of low sensitivity (Fig. 6B), AUC_{inf4} regains some power, while assessment of bioequivalence based on AUC_{lqc} worsens, as expected. From Fig. 2B the improved behavior of AUC_{inf4} can be understood; a lower sensitivity (higher LQ) hides the most variable segment of the curves. Although a worse estimate of the AUC is obtained (in absolute terms), its stability is improved. For an

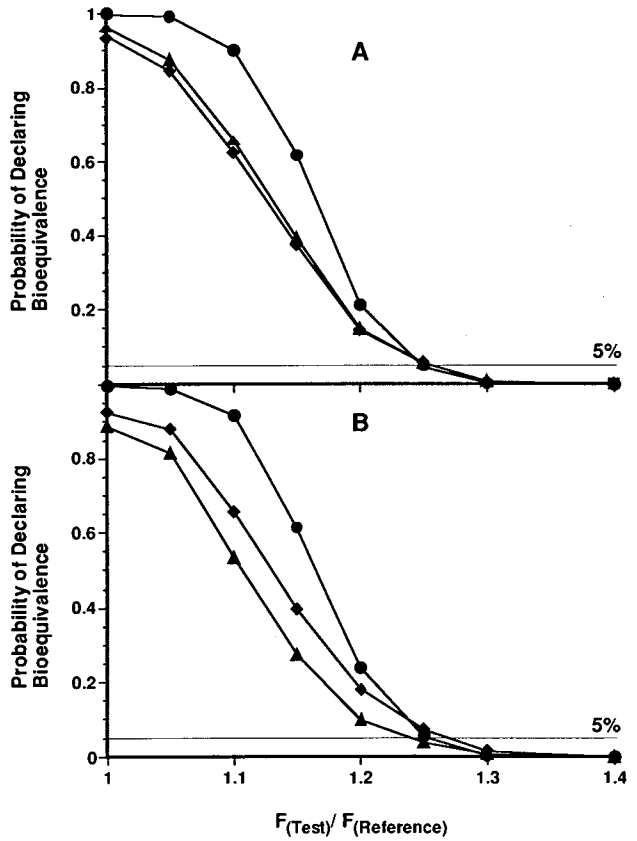

Fig. 4. Statistical power curves for three measures of extent of absorption: C_{max} (circles), AUC_{inf4} (diamonds), and AUC_{lqc} (triangles). Scenarios: (A) baseline; (B) low assay sensitivity. The probability of declaring bioequivalence using two one-sided t tests is given as a function of the ratio of bioavailabilities for test and reference formulations.

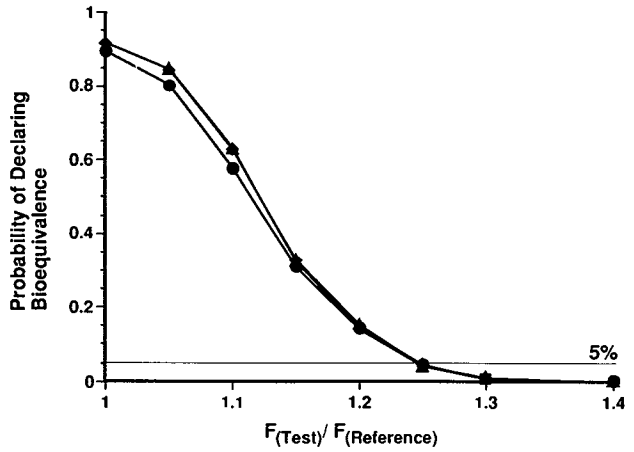

Fig. 5. Statistical power curves for three measures of extent of absorption: C_{max} (circles), AUC_{inf4} (diamonds), and AUC_{lqc} (triangles). Scenario: "flip-flop." The probability of declaring bioequivalence using two one-sided t tests is given as a function of the ratio of bioavailabilities for test and reference formulations. For most of the points, triangles and diamonds overlap.

F_T/F_R of 1.25 the consumer risk with AUC_{lqc} data is well below its nominal 5% level.

In the second two-compartment model (model II, elimination < distribution; Figs. 6C and D), AUC_{inf4} consistently has a low power and correctly identifies bioequivalence in no more than 55 and 35% of crossover clinical trails of 24 subjects for the high- and low-sensitivity cases, respectively.

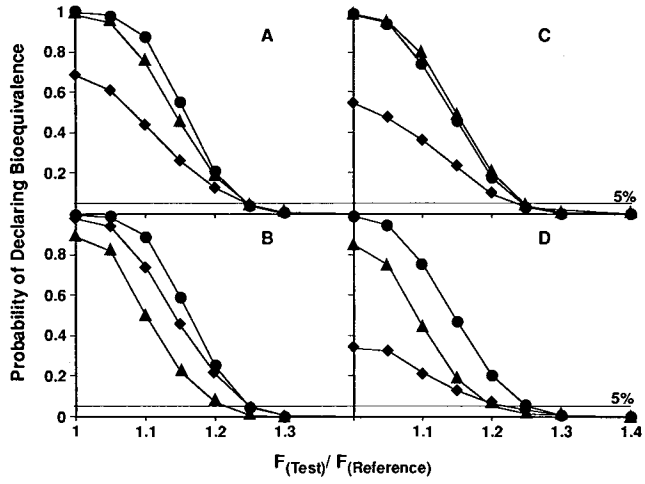

Fig. 6. Statistical power curves for three measures of extent of absorption: C_{max} (circles), AUC_{inf4} (diamonds), and AUC_{lqc} (triangles). Scenarios: two-compartment model I [(A) with 1% LQ and (B) with 10% LQ] and two-compartment model II [(C) with 1% LQ and (D) with 10% LQ]. The probability of declaring bioequivalence using two one-sided t tests is given as a function of the ratio of bioavailabilities for test and reference formulations.

C_{max} always has good power, while AUC_{lqc} may be even slightly better than C_{max} when LQ is low. In the case of low assay sensitivity, AUC_{lqc} is less reliable in that its power for an F_T/F_R ratio of 1.25 becomes far below the nominal 5% level.

CONCLUSION

Overall, when testing bioequivalence in terms of extent, both consistency and accuracy need to be considered. For example, AUC_{lqc} performs well in most cases, even though it underestimates the actual area. There appears to be no advantage in using AUC_{tfix} instead of AUC_{lqc}. Thus, among the classical measures of extent, our preference would then go toward the readily computed AUC_{lqc}. Model-based estimates of AUC, such as AUC_{infl} and AUC_{inf2}, which attempt to extrapolate the observed portion of the curve, may perform very poorly when two-compartment kinetics underlie what is treated as one-compartmental data. AUC_{inf3} and AUC_{inf4}, which rely more simply on linear extrapolation, may also lead to considerable error.

C_{max}, not usually used to assess extent, is the best of the measures tested in that it consistently has a low producer risk and correct consumer risk. C_{max}, a biased estimate of C_{peak}, is not as accurate a measure of extent of absorption as AUC can be. However, the consistency of C_{max} values for an individual, its relative protection from experimental error (assay error is small at high concentrations), and its statistical distribution contribute to its success. However, C_{max} is also sensitive to rate of absorption. In some cases, rate and extent influences on C_{max} could cancel each other. However, combining C_{max} with other measures of rate and extent may give a good overall assessment of bioequivalence.

ACKNOWLEDGMENTS

This work was supported by interagency agreement No. 224-91-3007 between the U.S. Food and Drug Administration and the U.S. Department of Energy.

REFERENCES

1. T. Colton. *Statistics in Medicine,* Little, Brown, Boston, 1974.
2. S. K. Park and K. W. Miller. Random number generators: Good ones are hard to find. *Commun. ACM* 31:1192–1201 (1988).
3. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C,* Cambridge University Press, Cambridge, MA, 1988.
4. A. Racine-Poon and A. F. Smith. Population models. In D. A. Berry (ed.), *Statistical Methodology in the Pharmaceutical Sciences,* Marcel Dekker, New York, 1990, pp. 139–162.
5. D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* 15:657–680 (1987).
6. Food and Drug Administration—Office of Generic Drugs—Division of Bioequivalence. Guidance on statistical procedures for bioequivalence studies using a standard two-treatment crossover design, 1992.