dx.doi.org/10.14227/DT230116P14

In Vitro Dissolution Curve Comparisons: A Critique of Current Practice

Dave LeBlond[1,*], Stan Altan[2], Steven Novick[3], John Peterson[4], Yan Shen[2], and Harry Yang[5]

e-mail: David.LeBlond@sbcglobal.net

[1] *CMC Statistics Consultant, Wadsworth, IL 60083*
[2] *Janssen Research & Development LLC, Raritan, NJ 08869*
[3] *GlaxoSmithKline Pharmaceuticals, Research Triangle Park, NC 27709*
[4] *GlaxoSmithKline Pharmaceuticals, Collegeville, PA 19426*
[5] *MedImmune LLC, One MedImmune Way, Gaithersburg, MD 20878*

ABSTRACT

Many pharmacologically active molecules are formulated as solid dosage form drug products. Following oral administration, the diffusion of an active molecule from the gastrointestinal tract into systemic distribution requires the disintegration of the dosage form followed by the dissolution of the molecule in the stomach lumen. Its dissolution properties may have a direct impact on its bioavailability and subsequent therapeutic effect. Consequently, dissolution (or in vitro release) testing has been the subject of intense scientific and regulatory interest over the past several decades. Much interest has focused on models describing in vitro release profiles over a time scale, and a number of methods have been proposed for testing similarity of profiles. In this article, we review previously published work on dissolution profile similarity testing and provide a detailed critique of current methods in order to set the stage for a Bayesian approach.

KEY WORDS: Dissolution profile similarity; in vitro release; f_2 statistic; Bayesian model.

INTRODUCTION

Many pharmacologically active molecules are formulated as immediate-release (IR) solid dosage form drug products. Following oral administration, the diffusion of an active molecule from the gastrointestinal tract into systemic distribution requires the disintegration of the dosage form followed by the dissolution of the molecule in the stomach lumen. Its dissolution properties may have a direct impact on its bioavailability and subsequent therapeutic effect. Consequently, dissolution (or in vitro release) testing methods have been studied as possible surrogates for human bioavailability studies in addition to product quality control.

In vitro dissolution testing as an analytical methodology measures drug release in liquid media. The method requires specialized laboratory equipment, following a well-defined protocol such as described in the United States Pharmacopeial convention reference (*1*). In vitro dissolution testing is frequently used as a surrogate measure of bioavailability. This avoids the risk and expense of human trials and facilitates the implementation of improvements in processes and products. Authorization to market a generic compound also makes use of dissolution profile comparisons. Consequently, in vitro dissolution testing has played an increasingly important role in drug development.

The 1997 FDA guidance on dissolution testing (*2*) describes three important uses for it: (1) assess lot-to-lot quality of a drug product, (2) guide development of new formulations, and (3) ensure continuing product quality and performance after certain events, such as changes in formulation, the manufacturing process, the site of manufacture, or scale-up of the manufacturing process. In addition, for certain drugs, in vitro dissolution results might be sufficient to gain regulatory approval for post-marketing changes and waiver of bioequivalence requirements for lower-strength dosage forms as described by Moore and Flanner (*3*). A formal similarity evaluation is a regulatory requirement for this purpose as described in FDA guidances (*2, 4–8*) and the 2008 EMA guidelines (*9*). Thus, the question of demonstrating similarity or equivalence between the reference and test drug dissolution curves is of both scientific and regulatory importance.

Various methods comparing drug dissolution profiles have been proposed. In general, they can be classified into three categories: model-independent approaches based on a similarity factor; model-independent methods using

*Corresponding author.

multivariate statistical distance (MSD) test; and model-dependent methods using parametric curves to describe dissolution profiles.

Model-Independent Approach Based on the f_2 Similarity Statistic

Moore and Flanner (*3*) proposed a model-independent approach to measure the similarity between drug dissolution profiles of a test and reference formulation. The method includes calculations of two mathematical indices:

$$f_1 = \frac{\sum_{i=1}^{k} |R_i - T_i|}{\sum_{i=1}^{k} R_i} \times 100$$

and

$$f_2 = 50 \log_{10} \left\{ 100 \left(1 + \frac{1}{k} \sum_{i=1}^{k} w_i (R_i - T_i)^2 \right)^{-1/2} \right\}$$

where R_i and T_i are reference and test results, respectively, at time t_i, i = 1, ..., k, and w_i is an optional weight factor. Subsequently, a similarity test procedure based on the difference factor f_1 and similarity factor f_2 with w_i = 1 was recommended in the FDA guidances for dissolution profile comparison (*2, 5–8*). Note that f_2 = 100 when the two dissolution curves are identical. When $f_2 \geq 50$ (e.g., when $|R_i - T_i| \leq 10$ for all i), the two dissolution profiles are deemed to be similar according to the FDA guidance for immediate-release (IR) solid oral dosage forms (*2*). While not explicitly stated, inference based on f_2 is seemingly intended to test the null hypothesis: H_0: $f_2^0 < 50$ versus the alternative: H_1: $f_2^0 \geq 50$, where

$$f_2^0 = 50 \log_{10} \left\{ 100 \left(1 + \frac{1}{k} \sum_{i=1}^{k} (E[R_i] - E[T_i])^2 \right)^{-1/2} \right\}$$

and $E(R_i)$ and $E(T_i)$ are the expected values of the ith dissolution measurements at time t_i, i = 1, ..., k, of reference and test batches across the population of dosage units, respectively. Since $\log_{10}(x)$ is a concave function, applying Jenson's inequality, it can be shown that $E(f_2) < f_2^0$. The bias of f_2 was also empirically evaluated by Ma et al. (*10*). Since f_2 underestimates the similarity measure f_2^0, a test based on f_2 is less likely to conclude similarity than one based on an unbiased estimator. A method based on bias correction was suggested by Shah et al. (*11*). The latter authors also explored the use of a bootstrap approach to construct a confidence interval on f_2^0 since the sampling distribution of f_2 is difficult to derive analytically as shown by Ma et al. (*10*). Other deficiencies of f_2 as a similarity measure will be discussed further in later sections. A few other model-independent similarity factors have also been suggested in the literature such as the one given by Moore and Flanner (*3*) based on the index of Rescigno (*12*) originally used to compare plasma drug concentration curves and mean dissolution times. As with f_2, these methods do not explicitly define the hypothesis of similarity, nor do they have the ability to demonstrate operating characteristics of the tests such as Type I and II errors.

Model-Independent Multivariate Approach

Another class of model-independent methods hinges on the normality assumption underlying the in vitro release values observed at different time points and constructs a measure of distance between two sets of multivariate random variables. Consider a dissolution study with times t_i, i = 1, ..., k. Let $\bar{R}$ and $\bar{T}$ denote the vectors of sample means, respectively, for reference and test dissolution measurements across k time points and let S_p be the pooled sample covariance matrix. Thus

$$T^2 = n_1 n_2 (n_1 + n_2)^{-1} [(\bar{R} - \bar{T}) - (\mu_R - \mu_T)]' S_p^{-1} [(\bar{R} - \bar{T}) - (\mu_R - \mu_T)]$$

is the Hotelling's T^2 statistic. It can be shown that $\frac{n_1 + n_2 - k - 1}{(n_1 + n_2 - 2)k} T^2$ follows an F distribution with k and $n_1 + n_2 - k - 1$ degrees of freedom. A 95% confidence region of $\mu_R - \mu_T$ is obtained as follows:

$$C_{0.05}(\bar{R}, \bar{T}, S_p) = \left\{ T^2 \leq \frac{(n_1 + n_2 - 2)k}{n_1 + n_2 - k - 1} F_{k, n_1 + n_2 - k - 1; 0.05} \right\}$$

where $F_{k, n_1 + n_2 - k - 1, 0.05}$ denotes the 5% quantile of an F-distribution with the given degrees of freedom.

In cases where excessive variance is present, the FDA guidance (*2*) recommends an alternative multivariate statistical distance (MSD) equivalence test in place of using the f_2 statistic. Tsong et al. (*13*) proposed a method that compares the maximum MSD,

$$D_{M,\alpha} = \max \sqrt{(\mu_R - \mu_T)' S_p^{-1} (\mu_R - \mu_T)}$$

obtained over the (1 - α)% confidence region $C_\alpha(\bar{R}, \bar{T}, S_p)$ with the tolerance limit

$$TL = \sqrt{\delta^2 J' S_p^{-1} J},$$

where δ is the maximum allowable similarity limit at each sampling time, and J denotes the $k \times 1$ vector of ones. Similarity is claimed if $D_{M,\alpha} \leq TL$. Although $|(\mu_R - \mu_T)_i| \leq \delta$, for i = 1, 2, ..., k, implies $D_{M,\alpha} \leq TL$, the converse is not true. Therefore, it is conceivable that the two dissolution curves may have a meaningful difference at particular sampling times, whereas the maximum MSD remains acceptable. Notice also that TL, as defined above, depends on the observed pooled sample variance S_p. Thus the definition of similarity will vary from data set to data set.

Saranadasa (*14*) suggested a model-independent method under the assumption that the reference and test dissolution curves are parallel—one overlaps with the other after an upward or downward shift of *d*. A maximum shift d_M is determined as the solution to the equation

$$[(\bar{R}-\bar{T})-dJ]' S_p^{-1}[(\bar{R}-\bar{T})-dJ] = \frac{n_1+n_2}{n_1 n_2}\frac{nk}{n-k+1}F_{k,n-k+1;\alpha}$$

where $n = n_1 + n_2 - 2$. The global similarity is achieved if $d_M \le \delta$. As the reference and test dissolution curves are seldom parallel, the method is of limited utility in practice.

One issue associated with the above multivariate distance methods is the difficulty in assessing their operating characteristics. Saranadasa and Krishnamoorthy (*15*), in an effort to address these shortcomings, developed a multivariate test of size α for assessing the similarity of two dissolution profiles assuming parallelism. Chow and Ki (*16*) used a time-series approach to account for the correlation between consecutive time points.

Model-Dependent Approach

Model-dependent methods are also listed in the regulatory guidances as alternative procedures for dissolution similarity testing. These rely on describing dissolution profiles through mathematical functions, based on an understanding of the dissolution processes. Because biological systems and related physiological–chemical processes concerning drug dissolution are exceedingly complex, both mechanistic models (*17–21*) and empirical models such as the Weibull (*22*) have been proposed to describe dissolution concentration–time profiles. Other functions that can be used to fit the dissolution profiles include exponential, probit, Gompertz, and logistic (*23*). In general, the Weibull was considered to be the most flexible (*24, 25*) in describing a wide variety of shapes. More recently, linear mixed effects and nonlinear mixed effects models have been suggested (*25, 26*). One advantage of these methods is that they permit an explicit specification of the covariance structure of the data. One potential pitfall of these methods is that the model parameters may be biased or not estimable if the sampling points are not appropriately chosen.

Yuksel et al. (*27*) provided an empirical study of observed dissolution profiles of IR tablets of naproxen sodium comparing an analysis of variance (ANOVA) standard hypothesis testing approach with model-dependent approaches through t-tests and comparing these with the f_2 calculation. It was restricted to the comparison of a test batch with a reference batch. No discussion was given of the more general case of making inferences to multiple batches from two different populations, which might happen in a post-approval change of manufacturing process or site. In addition, no discussion was given of an equivalence approach in place of hypothesis testing where a criterion for similarity would be specified in the case of the ANOVA or model-based approaches. The conclusions given in the paper cannot be generalized beyond the test case discussed.

Non-Parametric Comparison of Dissolution Curves

Bartoszynski et al. (*28*) provided a statistically sound approach to dissolution profile comparisons based on three candidate statistics. These were statistics related to (1) the rank-score method as an extension of the Mann–Whitney test, (2) an extension of the Kolmogorov–Smirnov *D*-statistic comparing three empirical cumulative distributions, and (3) an adaptation of the chi-squared test. A proximity measure related to a Minkowski metric was defined to compare all possible pairs of curves. A conventional hypothesis test of equality methodological approach formed the basis for drawing inferences in all three cases. It was limited to the case of testing two batches, test versus reference. No discussion of an equivalence approach in which the method would accommodate a meaningful difference criterion was given, nor is it obvious how one would extend this method to the case of multiple batches from two populations. The ranking method ignores actual differences so it is also conceivable that varying magnitudes of difference would be considered different depending on the pattern of the data rather than any given practically meaningful overall difference criterion.

NORMATIVE REQUIREMENTS FOR A DISSOLUTION PROFILE SIMILARITY TEST

Eaton et al. (*29*) have criticized the approaches discussed in the previous section for in vitro dissolution profile similarity testing. The authors defined two normative requirements (NRs) for a sound statistical methodology for testing dissolution profile similarity:

NR1: A specified function of population parameters (not involving data or experimental design) should be used to define dissolution profile similarity.

NR2: No matter what testing procedure is used, there needs to be sufficiently detailed knowledge about the power function to allow an assessment of the probabilities of Type I and Type II errors.

As indicated by Eaton et al. (*29*), the f_2 statistic and the MSD equivalence test fall short of these two NRs. These will be discussed further in the next section.

In addition to the two NRs suggested by Eaton et al., a sound statistical testing methodology should also satisfy the following additional requirements:

NR3: The parametric definition of similarity should be independent of the statistical methodology used to test for similarity.

NR4: The test for similarity should make clear the inference space for the conclusion. For instance, does the conclusion apply to the populations of test and reference batches or only to those batches providing data for the comparison?

NR5: A minimum confidence level should be specified that properly accounts for estimation uncertainty.

In the following sections, we review various published methods for dissolution profile similarity testing from the perspective of the above set of basic requirements for a statistical procedure. We then describe some of the differences in motivation between scientific and regulatory stakeholders that should be considered in designing a statistically appropriate dissolution profile similarity approach. Finally we suggest that a Bayesian approach to the question of similarity may offer many conceptual and practical advantages that will be attractive to both scientific and regulatory stakeholders.

DETAILED CRITIQUE OF CURRENT APPROACHES TO SIMILARITY TESTING OF DISSOLUTION PROFILES

The f_2 Similarity Statistic

The f_2 similarity statistic, introduced by Moore and Flanner (*3*), was immediately incorporated into an FDA guidance (*2*). It is very simple for scientists to calculate and use. It does not consider statistical or modeling details and does not require (or take advantage of) any theoretical model of profile shape. The f_2 statistic has been adopted by regulatory bodies worldwide, although implementation details and limits in specific guidances differ somewhat. Despite its widespread use, it has serious technical shortcomings. In addition, it fails to meet the normative requirements for an appropriate statistical procedure discussed previously. Specific criticisms of the f_2 statistic are the following:

I. Current regulatory guidances (*2, 5–7, 9, 30, 31*) fail to define the population characteristic that f_2 estimates. These guidances provide no underlying statistical model that allows f_2 to be defined parametrically, thus violating NR1. Therefore, it is not possible to construct a statistical hypothesis or equivalence test based on this statistic. Should one consider the parameter f_2^0 given previously, f_2 is a biased estimator and possesses a complex and intractable sampling distribution. Bootstrapping has been investigated in the literature, but the coverage is not nominal and the bias requires approximations and corrections to obtain a bootstrap interval estimate.

II. Without a definition of similarity in terms of model parameters, the operating characteristics of any hypothesis or equivalence test based on the f_2 statistic cannot be determined, thus violating NR2.

III. Current regulatory guidances do not state a minimal confidence level at which a similarity determination based on f_2 should be made, as indicated by NR5. The recommended approach is based solely on a point estimate of f_2 (i.e., whether or not the observed f_2 statistic is greater than 50). Thus if the median sampling distribution of f_2, for a particular comparison, equals 50, the Type I error of the determination is 50% (equivalent to a decision based on a fair-coin flip). The difficulties of developing an equivalence test based on the f_2 statistic are described well by Eaton et al. (*29*) who conclude that "…an analytic description of the sampling distribution of f_2 seems quite hopeless," and that "It is straightforward to show that f_2 underestimates [its underlying parameter] and this may adversely affect any bootstrap confidence interval/testing argument."

IV. Current regulatory guidances provide insufficient direction regarding experimental design of the comparison trial.

A. Twelve dosage units are required of both test and reference, but no justification is given for this sample size. None of the regulatory guidances state how to proceed if more units of either test or reference are available.

B. It is not clear whether these 12 should come from single or multiple batches. FDA (*2*) states that the reference should be either the most recent manufactured lot prechange or the last two or more consecutive lots prechange. FDA (*6*) states that the reference should be three consecutive recent lots, prechange. The EMA *(9)*, Japanese (*30*) and WHO (*31*) guidances give no information on the selection of test or reference batches.

C. When multiple batches of either test or reference (or both) are available, it is not clear whether to "pool" batches into test and reference superbatches and conduct a single comparison, or to conduct multiple comparisons of all possible combinations

of test and reference batches. The first approach would seem to require some pooling test, and the second approach would seem to demand a multiple comparison correction of some kind. None of the regulatory guidances address these issues.

D. EMA (*9*), FDA (*2*), and WHO (*31*) state that no more than a single time point past 85% dissolution should be used for either test or reference batches. However, it is not clear whether this requirement applies to the observed dissolution mean or to that of individual dose units.

E. FDA (*7*) states the number of time points to be used should be a "sufficient number of points." EMA (*9*), FDA (*2*), and WHO (*31*) recommend at least three. The Japanese guideline (*30*) specifies four according to a formula. As shown in criticism V of this section, the conservatism of the f_2 similarity requirement depends on the number of time points used to calculate the f_2 statistic. For immediate-release products, it may be difficult to obtain more than three time points, but many more time points can be available for modified- or extended-release products.

F. Four of the guidances (*2, 7, 9, 31*) disallow the use of the f_2 statistic in cases of excess variation in the determination of percent dissolution. Two of the guidances (*9, 31*) state an RSD strictly less than 20% for the first time point and strictly less than 10% at later times, whereas FDA (*2*) allows up to and including 20% and 10%, respectively, for the corresponding time points. FDA (*6*) also imposes an additional constraint that no difference between reference and test means across any of the time points can exceed 15%. The guidances do not make clear whether the RSD requirement is based on intrabatch RSD, interbatch RSD, or some other (e.g., total RSD) measure.

V. As the number of time points included in the comparison is increased, the f_2 criterion becomes more liberal in that larger deviations can be accommodated. The f_2 similarity region is based on a hypersphere having a dimension equal to the number of time points. The radius of the sphere increases with the number of time points, being 29% larger for five time points than for three. This leads to an opportunity to increase the likelihood of claiming similarity by increasing the number of time points used to calculate the f_2 statistic or by choosing an excess of early or late time points where test and reference percent dissolution may be physically constrained to be similar (NR1 violation).

VI. To use the f_2 statistic, the test and reference must use exactly the same time points. However, there are cases where this requirement may not be met. For instance, it is sometimes necessary to make similarity comparisons against historical batch release tests or early development tests that use a different dissolution protocol than that for a newer process. This also may be the case for comparisons made using test results obtained in different laboratories. Such comparisons would require the use of a theoretical model for the shape of the profile. EMA (*9*) mentions briefly the use of a Weibull model, and FDA (*2*) permits the use of a model-dependent approach. The other guidances provide no explicit provision for use of such a model. Thus it is possible that conflicting results could arise since regulatory agencies typically demand the use of the f_2 statistic for comparison purposes.

VII. Current regulatory guidances avoid the inconvenient yet critical considerations of the intended inference space for similarity comparisons (NR4 violation). The apparent intent of the similarity comparison is to assure that units produced in the future by a new (test) process can provide equivalent dissolution characteristics to those formerly produced by an approved (reference) process. However these guidances seem to imply that the similarity decision is based on the particular batches available. They do not discuss how similarity conclusions with respect to the test and reference population of batches should be drawn. Drawing inferences about the processes (rather than the batches) requires consideration of sampling design, relative magnitude of intrabatch and interbatch variances, and justification of an appropriate hierarchical statistical model. If the true objects of comparison are test and reference populations or the manufacturing processes overall, then avoiding such considerations is unscientific.

VIII. The f_2 statistic does not account for differences in variance among the time points. In other words, the statistic is "unweighted." Yet it is commonly observed that the variance is larger near 50% dissolution than near 0% or 100% dissolution. This raises the concern of whether statistical weighting should be incorporated into the calculation of the statistic, as would be required to provide a proper statistical hypothesis or equivalence test. Such concerns beg the question of exactly what is meant by similarity.

IX. The f_2 statistic is location invariant. Its value depends solely on the observed mean differences between test

and reference units at the various time points. There is no consideration of the magnitude of these differences relative to the overall change in percent dissolution over the time interval studied or of the time-order in which these differences occur. Thus the f_2 statistic is more of a distance metric than a profile-shape metric. It would not distinguish between non-similarity of the test units due to dose dumping from that due to failure to deliver the labeled dose at longer times.

X. The ability of the f_2 statistic to signal a difference in mean percent dissolution between test and reference units may depend critically on the choice of time points at which measurements are taken. Because it is based on a limited set of discrete time points, it is possible for f_2 to exceed 50, while the two dissolution profiles are not equivalently close enough between two observed time points. Clarity in the current guidances is lacking on the number and spacing of time points, and their dependence on dissolution profile shape of the reference.

Multivariate Statistical Distance (MSD) Equivalence Test

The switch to use of the MSD test represents a fundamental change in the definition of similarity. Whereas the f_2 statistic depends only on mean differences, the MSD depends on the pooled variances at each time point and the correlations among time points. While the f_2 test of similarity is based on an observed f_2, the MSD test of similarity is based on the outcome of a formal multivariate statistical equivalence test with a specified maximum Type I error level (0.05). Despite an apparent advantage in relation to known statistical properties, we feel the MSD test as implemented in the guidance and literature also fails to provide a satisfactory approach to a test for similarity. The reasons are summarized as follows:

I. The decision whether to use f_2 or MSD, and thus on the fundamental definition of similarity, depends on observed data (i.e., sample variances or %CVs). Further, the tolerance limit, *TL*, is a function of the observed pooled sample variance, S_p (NR1 violation).

II. The MSD similarity limit implementation is unclear. Whereas the decision limit for the f_2 statistic ($f_2 > 50$) is provided in the guidances, the MSD limit is not provided but must be determined on a case-by-case basis and is dependent on differences among reference batches. No direction is provided on the amount of historical data required or the degree of conservatism to be used in setting this limit. A greater interbatch variance among reference batches implies a wider acceptance limit for MSD. Thus there is no clear understanding of the meaning of similarity across products.

III. The MSD test is not based on a hierarchical model that accounts for the relative magnitudes of intrabatch and interbatch variance components. Whereas the f_2 statistic does not distinguish between intrabatch and interbatch variance, the decision to switch to the MSD test depends on intrabatch variance, and the MSD limit itself depends on interbatch variance (of the reference formulation). Yet there is no requirement that the observed MSD take into account the relative magnitudes of these variances. As with the f_2 statistic, it is not clear whether the conclusion regarding similarity is meant to apply to the test and reference populations of batches or merely to the batches used to conduct the MSD equivalence test (NR4 violation).

IV. The MSD metric is not a measure of profile shape similarity. It is essentially a ratio of squared dissolution differences divided by their pooled measurement variances. Processes that exhibit larger variances can accommodate greater mean differences for a given fixed value of the MSD metric. For this reason, the decision criterion for the MSD test involves the construction of a multivariate confidence region. The complexity of the MSD test requires greater statistical expertise and specialized statistical software. The need to use the MSD approach thus leads to a considerable "culture shock" for the dissolution scientists who may not have anticipated the need for statistical support prior to collection of the data.

V. Current regulatory guidances do not mention the construction of a multivariate confidence region test based on the MSD. Berger and Hsu (*32*) have shown that this test has a nominal size, given the usual assumption of normality. However, literature examples have used an inversion of Hotelling's T^2 to obtain a test based on an estimated confidence region. Such a test can be very conservative and increasingly so as the number of test points (or dimensions) increases as shown by Eaton et al. (*29*).

VI. Multivariate confidence region estimates are not unique. While the minimum coverage of the MSD-based multivariate confidence region (0.90) is specified (*2*), other aspects of the region are not defined. Confidence regions for a given coverage can take on arbitrary shape (hyper-elliptical, hyper-rectangular, one- or two-sided in some or all dimensions, etc.) and, in general, will not conform to the shape of the defined similarity region. The location and extent of the confidence region

extrema thus depend on these shape choices and the way in which risk is allocated. The MSD equivalence test requires that the entire confidence region (including extrema) be contained within whatever similarity region is determined. Thus the MSD equivalence test can be made more or less conservative, depending on the shape choices made by those conducting and reporting the equivalence test results.

VII. Literature implementations of the MSD test have been controversial. For instance, the implementations reported by Tsong et al. (*13*) and Sathe et al. (*33*) illustrate the subjective impact of similarity and confidence region shape choices. They illustrate rectangular similarity neighborhoods, whereas the f_2 statistic utilizes a spherical similarity neighborhood. Their stated measure of similarity (the pivotal MSD statistic) is independent of model parameters and thus seems to violate NR1. The implementation reported by Saranadasa and Krishnamoorthy (*15*) assumes that the curve shapes of test and reference are parallel, which is likely never exactly true and therefore cannot be a consistently tenable approximation.

SCIENTIFIC AND QUALITY STAKEHOLDER PERSPECTIVES

Approval of a submitted request to change an existing medical intervention product involves scientific and quality considerations related to risk. One could pose the following questions in this regard:

Scientific Perspective: What is the probability that this particular change is unsafe or ineffective?

Quality Perspective: What is the probability (over many submissions) that ineffective or unsafe changes will be approved?

The former would be a primary concern to the sponsor of the change, whereas the latter would be a primary concern to the regulator. Both questions can be framed in terms of probabilities, although the natures of these probabilities are very different.

The scientific perspective demands development of a body of knowledge—a knowledge base specific to the test and reference processes being compared. Such knowledge may come from established theoretical or statistical models, appropriately designed experiments, relevant historical development experience, and the prior experience of subject matter experts, but in general contains substantial uncertainty. Good scientific decision-making utilizes the entire body of accumulated knowledge to make an informed decision about the proposed change in the face of this uncertainty. This is an inevitable feature of such comparisons and calls for an appropriate risk assessment methodology that provides a clear characterization of the uncertainty contained in the knowledge base.

The quality perspective is more concerned with the calibration of risk decision tools that are used across many sponsors and many processes and is the basis for NR1 described earlier. We submit that the maintenance of a high level of overall quality of biopharmaceutical products and their impact on public health over time requires decision-making tools whose operating characteristics are documented (as indicated by NR2). This is a policy consideration that deserves wider recognition and discussion.

Clearly, both perspectives are important to sponsors and regulators alike. If powerful and informative decision-making tools are not utilized, or the operating characteristics of the statistical tools are not well documented and understood, this could lead to inconsistencies in regulatory risk management and quality level across sponsors and processes, especially if only the scientific concerns are addressed. On the other hand, focusing solely on quality concerns may not fully utilize the available knowledge resulting in an unknown performance of acceptable decision-making. In considering improved tools for the process of decision-making, the impact on both scientific and quality risk must be considered by both sponsors and regulators.

Below, we discuss the possibility of Bayesian approaches to in vitro dissolution similarity comparisons which we feel offer many benefits and improvements from both scientific and quality perspectives as well as satisfying the normative requirements listed in the introduction.

A BAYESIAN PERSPECTIVE TO THE IN VITRO DISSOLUTION COMPARISON PROBLEM

Traditional statistical approaches to similarity and equivalence testing involve hypothesis tests or confidence-region arguments based on Type I level or repeated sampling coverage arguments. These methodologies involve deductive reasoning (i.e., inferring the probability of observed data, given a hypothesis). The probabilities (*p*-values, confidence coefficients) associated with such approaches are attractive from a quality perspective because they are sometimes (but not always) nominal values equal to the probability coefficients used in the theoretical derivation of the method. However, these probabilities are more concerned with the long-term

performance of a given decision tool than with the likely values of the uncertain parameters that are the subject of a given scientific investigation.

Goodman (*34, 35*) has argued that measuring the strength of scientific evidence requires both deductive and inductive reasoning (i.e., inferring the probability of a hypothesis, given observed data) and that this is facilitated by a Bayesian perspective. According to Bayes' rule, all inferences about the values of uncertain parameters are obtained by combining information from data with prior knowledge to obtain a multivariate posterior distribution of model parameters. Since NR1 states that similarity be defined parametrically, a Bayesian approach to in vitro dissolution comparisons seems most appropriate. Sometimes there is a close relationship between frequentist and Bayesian probabilities, but the events associated with these two probabilities are different. More often, frequentist and Bayesian probabilities will be different. We feel that Bayesian probabilities offer the following advantages as a profile comparison tool:

- Unlike confidence region methods, the integrated posterior corresponds exactly to the region of similarity. There is no conservatism.
- A Bayesian approach makes possible an intuitive statement such as "The probability of similarity is *p*." Such a probability statement takes into account the uncertainty of the unknown model parameters. In contrast, the results of a frequentist approach to inference through a confidence interval or hypothesis test cannot be interpreted in the same intuitive manner. Frequentist inference may also require the use of large-sample asymptotics, which further complicates the interpretation of the resulting inferential statements.
- Using modern Bayesian sampling approaches, it is possible to handle in an exact way (i.e., estimation accuracy limited only by the size of the posterior sample drawn) complex situations, which require conservative, asymptotic, or large-sample approximations using traditional approaches. This includes the following types of models that can be usefully employed in profile comparisons:
  - ♦ Multivariate.
  - ♦ Nonlinear, including theoretical models of dissolution profile shape such as Weibull.
  - ♦ Hierarchical models that include both intrabatch and interbatch variance.
  - ♦ "Mixed" models in which some parameters are considered fixed and others random.
  - ♦ Predictive models in which we are concerned with behavior of future batches.
  - ♦ Missing data.
- It is straightforward to calculate the posterior distribution of any fixed or random function of these parameters (such as a given similarity metric). This circumvents the need to derive a sampling distribution for such functions.
- The availability of modern Bayesian methods makes it possible to calculate various probabilities by simple counting. This circumvents the need to analytically or numerically integrate the distributions.
- Bayesian computations can be efficiently carried out by a number of freely available software such as R (*36*), WinBUGS (*37*), Stan (*38*), and JAGS (*39*). WinBUGS in particular represents over 20 years of experience by thousands of users worldwide and is a very mature and stable computing environment. A plethora of computing examples are available with the software and in many text books, including the kind of models appropriate for dissolution similarity comparisons.
- The Bayesian approach permits prior knowledge to be incorporated into the decision process in a quantitative, objective way. When there is no relevant prior knowledge available, or where the data must "stand on its own," non-informative or vague priors are well known and their impact on the decision can be determined. The incorporation of prior knowledge can be of great utility in community decision-making.
- As noted above, the sampling distribution of f_2 is intractable, which inhibits development of a statistical equivalence test based on f_2. However, it is trivial to obtain the posterior distribution of f_2 (given an appropriate definition of f_2 in terms of model parameters). Thus the Bayesian approach can provide a link to the established metric.

We see two principal barriers to taking a Bayesian approach. First, while the results of Bayesian analyses will be intuitive and understandable to scientists, statisticians, and regulators alike, the use of Bayesian technology is probably unfamiliar to many. Consequently, statistical support is recommended. Certain aspects such as model parameterization, prior elicitation and choice, and convergence verification require care. However,

our experiences with Bayesian methodology have been positive. We feel this barrier can be overcome with training. Second, mindful of the quality perspective, there is a need to quantify the risks associated with a decision rule, particularly when an analytical power calculation is not possible. This can be computationally intensive, but in principle, it is straightforward to accomplish using a Bayesian simulation approach. Such methodology is becoming more commonplace, and we see this as part of best statistical practice. For these reasons, we believe Bayesian methodology has the potential of overcoming many of the issues with f_2 and MSD while maintaining a link to established criteria. Such methodologies are the subject of a separate article by the authors of this review (*40*). Other papers expanding on the topic are under discussion.

CONFLICT OF INTEREST

The authors certify that they have received no funding from any company that may be affected by the research reported in this paper.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Mandy Bergquist and Raymond Buck for their careful reading and comments on an earlier draft of this paper.

REFERENCES

1. *The United States Pharmacopeia and National Formulary USP 35–NF 30*; The United States Pharmacopeial Convention, Inc.: Rockville, MD, 2012.
2. *Dissolution Testing of Immediate Release Solid Oral Dosage Forms;* Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 1997).
3. Moore, J. W.; Flanner, H. H. Mathematical Comparison of Curves with an Emphasis on In-Vitro Dissolution Profiles. *Pharm. Technol.* **1996**, *20* (6), 64–74.
4. *Immediate Release Solid Oral Dosage Forms, Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation*; Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 1995.
5. *Extended Release Oral Dosage Forms: Development, Evaluation, and Application of In Vitro/In Vivo Correlations*; Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 1997.
6. *SUPAC-MR: Modified Release Solid Oral Dosage Forms, Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation*; Guidance for Industry; Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 1997.
7. *Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System*; Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 2000.
8. *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations;* Guidance for Industry; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), U.S. Government Printing Office: Washington, DC, 2003.
9. Guideline on the Investigation of Bioequivalence; CPMP/EWP/QWP/1401/98 Rev. 1; Committee for Medicinal Products for Human Use (CHMP), European Medicines Agency: London, 2008.
10. Ma, M.-C.; Wang, B. B. C.; Liu, J.-P.; Tsong, Y. ASSESSMENT OF SIMILARITY BETWEEN DISSOLUTION PROFILES. *J. Biopharm. Stat.* **2000**, *10* (2), 229–249. DOI: 10.1081/BIP-100101024.
11. Shah, V. P.; Tsong, Y.; Sathe, P.; Liu, J.-P. In Vitro Dissolution Profile Comparison–Statistics and Analysis of the Similarity Factor, f_2. *Pharm. Res.* **1998**, *15* (6), 889–896. DOI: 10.1023/A:1011976615750.
12. Rescigno, A. Bioequivalence. *Pharm. Res.* **1992**, *9* (7), 925–928. DOI: 10.1023/A:1015809201503.
13. Tsong, Y.; Hammerstrom, T.; Sathe, P.; Shah, V. P. Statistical Assessment of Mean Differences between Two Dissolution Data Sets. *Ther. Innovation Reg. Sci.* **1996**, *30* (4), 1105–1112. DOI: 10.1177/009286159603000427.
14. Saranadasa, H. Defining the Similarity of Dissolution Profiles through Hotelling's T^2 Statistic. *Pharm. Tech.* **2001**, *25* (2), 46–54.
15. Saranadasa, H.; Krishnamoorthy, K. A MULTIVARIATE TEST FOR SIMILARITY OF TWO DISSOLUTION PROFILES. *J. Biopharm. Stat.* **2005**, *15* (2), 265–278. DOI: 10.1081/BIP-200049832.
16. Chow, S.-C.; Ki, F. Y. C. Statistical Comparison Between

Dissolution Profiles of Drug Products. *J. Biopharm. Stat.* **1997**, *7* (2), 241–258. DOI: 10.1080/10543409708835184.
17. Gibaldi, M.; Feldman, S. Establishment of sink conditions in dissolution rate determinations. Theoretical considerations and application to nondisintegrating dosage forms. *J. Pharm. Sci*. **1967**, *56* (10), 1238–1242. DOI: 10.1002/jps.2600561005.
18. Wagner, J. G. Interpretation of percent dissolved–time plots derived from in vitro testing of conventional tablets and capsules. *J. Pharm. Sci.* **1969**, *58* (10), 1253–1257. DOI: 10.1002/jps.2600581021.
19. Higuchi, T. Rate of release of medicaments from ointment bases containing drugs in suspension. *J. Pharm. Sci.* **1961**, *50* (10), 874–875. DOI: 10.1002/jps.2600501018.
20. Higuchi, T. Mechanism of sustained-action medication. Theoretical analysis of rate of release of solid drugs dispersed in solid matrices. *J. Pharm. Sci.* **1963**, *52* (12), 1145–1149. DOI: 10.1002/jps.2600521210.
21. Hixson, A. W.; Crowell, J. H. Dependence of Reaction Velocity upon Surface and Agitation. I-Theoretical Consideration. *Ind. Eng. Chem.* **1931**, *23* (8), 923–931. DOI: 10.1021/ie50260a018.
22. Langenbucher, F. Letters to the Editor: Linearization of dissolution rate curves by the Weibull distribution. *J. Pharm. Pharmacol*. **1972**, *24* (12), 979–981. DOI: 10.1111/j.2042-7158.1972.tb08930.x.
23. Tsong, Y.; Sathe, P. M.; Shah, V. P. In Vitro Dissolution Profile Comparison. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd ed.; Chow, S.-C., Ed.; CRC Press: Boca Raton, FL, 2003; pp 456–462. DOI: 10.1201/b14760-68.
24. Polli, J. E.; Rekhi, G. S.; Augsburger L. L.; Shah, V. P. Methods to Compare Dissolution Profiles and a Rationale for Wide Dissolution Specifications for Metoprolol Tartrate Tablets. *J. Pharm. Sci.* **1997**, *86* (6), 690–700. DOI: 10.1021/js960473x.
25. Adams, E.; De Maesschalck, R.; De Spiegeleer, B.; Vander Heyden, Y.; Smeyers-Verbeke, J.; Massart, D. L. Evaluation of dissolution profiles using principal component analysis. *Int. J. Pharm.* **2001**, *212* (1), 41–53. DOI: 10.1016/S0378-5173(00)00581-0.
26. Adams, E.; Coomans, D.; Smeyers-Verbeke, J.; Massart, D. L. Non-linear mixed effects models for the evaluation of dissolution profiles. *Int. J. Pharm.* **2002**, *240* (1–2), 37–53. DOI: 10.1016/S0378-5173(02)00127-8.
27. Yuksel, N.; Kanik, A. E.; Baykara, T. Comparison of in vitro dissolution profiles by ANOVA-based, model-dependent and -independent methods. *Int. J. Pham.* **2000**, *209* (1–2), 57–67. DOI: 10.1016/S0378-5173(00)00554-8.
28. Bartoszynski, R.; Powers, J. D.; Herderick, E. E.; Pultz, J. A. Statistical Comparison of Dissolution Curves. *Pharmacol. Res*. **2001**, *43* (4), 369–387. DOI: 10.1006/phrs.2001.0796.
29. Eaton, M. L.; Muirhead, R. J.; Steeno, G. S. Aspects of the Dissolution Profile Testing Problem. *Biopharm. Rep.* **2003**, *11* (2), 2–7.
30. *Guideline for Bioequivalence Studies of Generic Products;* English translation of Attachment 1 of Division-Notification 0229 No. 10, Appendix 1; Japanese National Institute of Health Sciences, Pharmaceutical and Food Safety Bureau: Tokyo, Japan, 2012.
31. WHO Expert committee on Specifications for pharmaceutical preparations. *Dissolution Profile Comparison*; WHO Technical Report Series, No. 937, Fortieth Report; World Health Organization: Geneva, 2006; p 382.
32. Berger, R. L.; Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.* **1996**, *11* (4), 283–319. DOI:10.1214/ss/1032280304.
33. Sathe, P. M.; Tsong, Y.; Shah, V. P. In-Vitro Dissolution Profile Comparison: Statistics and Analysis, Model Dependent Approach. *Pharm. Res.* **1996**, *13* (12), 1799–1803. DOI: 10.1023/A:1016020822093.
34. Goodman, S. N. Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy. *Ann. Intern. Med.* **1999**, *130* (12), 995–1004. DOI: 10.7326/0003-4819-130-12-199906150-00008.
35. Goodman, S. N. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Ann. Intern. Med.* ***1999***, *130* (12), 1005–1013. DOI: 10.7326/0003-4819-130-12-199906150-00019.
36. The R Project for Statistical Computing Home Page. R: A Language and Environment for Statistical Computing. http://www.R-project.org/ (accessed Jan 17, 2016).
37. Lunn, D. J.; Thomas, A.; Best, N.; Spiegelhalter, D. WinBUGS— A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **2000**, *10* (4), 325–337. DOI: 10.1023/A:1008929526011.
38. Stan Home Page. Stan: A C++ Library for Probability and Sampling, Version 2.5.0, 2014. http://mc-stan.org/ (accessed Jan 17, 2016).
39. Plummer, M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (DSC 2003), Vienna, Austria, March 20–22, 2003. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Plummer.pdf (accessed Jan 17, 2016).
40. Novick, S.; Shen, Y.; Yang, H; Peterson, J.; LeBlond, D.; Altan, S. Dissolution Curve Comparisons Through the F_2 parameter, a Bayesian Extension of the f_2 Statistic. *J. Biopharm. Stat.* **2015**, *25* (2), 351–371. DOI: 10.1080/10543406.2014.971175.