**This is a pre- or post-print of an article published in
Gómez-Mantilla, J.-D., Casabó, V.G., Schaefer, U.F.,
Lehr, C.-M.**

**Permutation Test (PT) and Tolerated Difference Test
(TDT): Two new, robust and powerful nonparametric tests
for statistical comparison of dissolution profiles
(2013) International Journal of Pharmaceutics, 441 (1-2),
pp. 458-467.**

PERMUTATION TEST (PT) AND TOLERATED DIFFERENCE TEST (TDT). TWO NEW, ROBUST AND POWERFUL NONPARAMETRIC TESTS FOR STATISTICAL COMPARISON OF DISSOLUTION PROFILES.

J.D.Gomez-Mantilla^{1,2}, V.G. Casabo³, U. Schaefer¹, C.M. Lehr.^{1,4}

¹Biopharmaceutics and Pharmaceutical Technology, Saarland University, Campus A4.1, D-66123 Saarbruecken, Germany.

² Department of Pharmacy, National University of Colombia. Bogota, Colombia.

³ Department of Technological Pharmacy, University of Valencia, Burjassot, Spain.

⁴ Helmholtz-Institute for Pharmaceutical Research (HIPS), Helmholtz Center for Infection Research (HZI), Saarbruecken 66123 Germany.

Corresponding author at: Helmholtz Institute for Pharmaceutical Research Saarland, Saarland University, P.O. Box 15 11 50, D-66041 Saarbruecken, Germany. Tel.: +49 681 302 3039; fax: +49 681 302 4677. E-mail address: lehr@mx.uni-saarland.de (C.-M. Lehr).

Abstract

The most popular way of comparing oral solid forms of drug formulations from different batches or manufacturers is through dissolution profile comparison. Usually, a similarity factor known as (f_2) is employed; However, the level of confidence associated with this method is uncertain and its statistical power is low. In addition, f_2 lacks the flexibility needed to perform in special scenarios. In this study two new statistical tests based on nonparametrical permutation test theory are described, the Permutation Test (PT), which is very restrictive to confer similarity, and the Tolerated Difference Test (TDT), which has flexible restrictedness to confer similarity, are described and compared to f_2 . The statistical power and robustness of the tests were analyzed by simulation using the Higuchi, Korsmayer, Peppas and Weibull dissolution models. Several batches of oral solid forms

were simulated while varying the velocity of dissolution (from 30 mins to 300 mins to dissolve 85% of the total content) and the variability within each batch (CV 2% to 30%). For levels of variability below 10% the new tests exhibited better statistical power than f2 and equal or better robustness than f2. TDT can also be modified to distinguish different levels of similarity and can be employed to obtain customized comparisons for specific drugs. In conclusion, two new methods, more versatile and with a stronger statistical basis than f2, are described and proposed as viable alternatives to that method. Additionally, an optimized time sampling strategy and an experimental design-driven strategy for performing dissolution profile comparisons are described.

1. Introduction

Comparing time profiles for dissolution data or for any other type of data is a complex statistical challenge. The highly correlated nature of this type of data, which exists in spite of its mostly unknown mechanisms, the many types of curves observed in dissolution profiles, the high variability combined with the finite nature of the variable ($\leq 100\%$), and the fact that two curves may cross producing both positive and negative differences, make it difficult to determine whether two curves should be regarded as similar or different, and therefore represent a major barrier to an adequate solution to this problem (, and). When a variable is measured over time and compared under two or more conditions, a simple and commonly used technique is to compare the value of the variable at one or two particular time points and to test hypotheses about differences in the variable between the different conditions at these precise time points . Although this approach is adequate in a broad variety of situations, it fails, when the major interest lies in the kinetic of the process, as when drug dissolution profiles are compared.

Drug dissolution assays of oral solid dosage forms are designed to predict the performance of these formulations in the gastro-intestinal tract (GIT) and ultimately provide information

about the bioavailability of oral formulations. The information obtained at each time point can be crucial because the absorption of drugs varies across the GIT due to the different membrane properties of the mucosal cells, the local microclimate and, the presence or absence of transporters, enzymes and other substances (, , and). Because of this, comparisons using data obtained at only one or two time points are insufficient. Comparisons of areas under the curve are also inadequate because two curves can have very similar areas under the curve but present important differences at single time points, especially if the two profiles cross . To date, there is no satisfactory statistical tool, either for dissolution profiles or for other types of data that completely solves this particular problem.

In 1996, Moore and Flanner described the use of an expression that they called f_2 (Equation 1) to compare dissolution curves, a method especially recommended for use in stability studies and optimization during product development and scale-up. Since f_2 has been proposed, several publications have explored the advantages and disadvantages of f_2 and some modifications, such as the constructions of confidence intervals have been proposed . Presently , f_2 is employed and recommended by regulatory authorities for scale-up and post-approval changes; in addition it can be used to waive clinical bioequivalence studies (at least under certain conditions) for immediate release and modified release solid formulations , and . However, the level of confidence of the method is uncertain and several publications have shown it to have low statistical power , .

$$f_2 = 50 \times \log \left\{ \left[1 + \frac{1}{n} \sum_{t=1}^n w_t (R_t - T_t)^2 \right]^{-0.5} \times 100 \right\} \quad (\text{Equation 1})$$

In equation 1, R_t is the mean of the dissolved drug from the Reference batch at time t , T_t is the mean of the drug dissolved from the Test batch at time t , n is the number of time points and w_t is a weight factor that can be used to enhance the influence of particular time points. If the calculation yields $f_2 \geq 50$, similarity of R and T is declared.

FDA guidelines recommend testing 12 tablets of each batch. Theoretically, if the difference in drug dissolution between R and T is exactly 10% at every time point, the value of f_2 is 50; if the differences are >10% (at every time point) f_2 becomes smaller than 50, and if the differences are smaller (<10%) f_2 becomes larger than 50. However, values of f_2 above 50 can be obtained with differences greater than 10% at some particular time points if the differences at the other points are small enough to compensate for the larger differences; thus, the basis for choosing a value of 50 as the rejection criterion is questionable. To alleviate this problem, Moore and Flanner included the weight factor w_t in the expression; however, there is no clarity on how to employ the weight factor and it can highly favor (intentionally or unintentionally) either similarity or non-similarity. The FDA guidelines also mention the weight factor but allow the researcher to decide whether to use it or not. Another disadvantage of using f_2 is that the arithmetic mean is very susceptible to extreme values and this may result in large differences between individual tablets being ignored.

Apart from f_2 , several methodologies for comparing dissolution profiles have been described like adaptations of single value comparisons of level B parameters, (area under curve, mean dissolution time, time to reach 85% of dissolution etc.), that have not been well accepted because it is often not possible to properly include the information for every time point in the comparison, or multivariate analysis, which require assumptions that are difficult to fulfill. Moreover, it is questionable whether comparison of dissolution profiles should be considered as a multivariate problem, because the same variable is measured repeatedly over time. Model-dependent methods have also been used, but these rely highly on fitting to a specific dissolution model, and in some cases such a model is not available. Moreover, model-dependent methods are still bound to multivariate distances with the aforementioned problems. Factors as f_2 which are easy to implement have been widely employed, but normally lack scientific justification or statistical support.

Most available statistical tests are designed to detect differences, rather than to prove similarities, and a lack of difference does not necessarily imply similarity. However, demonstration of a lack of difference with quantified and adequate type-I and type-II errors would provide a more solid statistically method for detecting similarities than a method based on subjective limits.

The power of a statistical test is the probability that the test will reject the null hypothesis (in this case similarity) when the null hypothesis is false (i.e. the probability of not committing a type-II error, or making a false negative decision). With the help of dissolution models (equations 2-5), scenarios when the null hypothesis is false can be generated (differences in the value of one or more parameters), and the power of the tests can be evaluated. The more powerful a test is, the smaller difference it can detect in the value of model parameters.

A more powerful test (able to detect small differences between two profiles) would be very valuable for comparing the dissolution profiles of formulations containing drugs with very narrow therapeutic windows and/or drugs classed as II, III and IV in the Biopharmaceutical Classification for which *in-vivo* bioequivalence can require a more strict, almost identical *in-vitro* similarity . It can also be postulated that for a transporter substrate (active transport or efflux) of a transporter present in enterocytes, the effective concentration in the intestinal lumen may play a decisive role in determining the bioavailability of the compound. Very powerful statistical tests are needed, indeed, to detect small differences in dissolution profiles to assure similarity of two products from different manufacturers or from the same manufacturer after a major or minor change in production technology. In a large number of cases, the bioavailability of two different drug products with the same active molecule will be very similar if their dissolution profiles, evaluated under the relevant conditions , are very similar.

On the other hand, for some compounds, large differences in dissolution profiles are necessary to produce significant differences in bioavailability, and a test less strict than f_2 is

also needed . In general, a flexible test that offers variable power according to specific needs, but still retains adequate levels of robustness and statistical uncertainty, is highly desirable.

Aware of the expectations that dissolution and drug release will play an even wider role in regulating quality generic drug products in the future , two major characteristic are needed in statistical tests for dissolution profile comparison: High statistical power and flexibility, as these two properties are not likely to be fulfilled by the same test, two separate tests, each with one of the mentioned properties may be an adequate solution.

In this study two new statistical tests based on nonparametric permutation test theory are presented, and their ability to satisfy the above mentioned requirements (more restrictiveness and more flexibility) is assessed. The first, called Permutation Test (PT), is capable of detecting small differences in dissolution profiles and is very exigent to confer similarity. The second one, called Tolerated Difference Test (TDT), the level of exigency to confer similarity can be modified to detect large or small differences according to the requirements of any particular case. Both tests were explored in terms of statistical robustness and power and were compared to f_2 and bootstrap 95% Confidence Intervals of f_2 (C.I.) .

2. Materials and methods

2.1 Data Simulations.

Dissolution data were simulated following four different mathematical dissolution models. i.e., the Higuchi model (Equation 2) , the Korsmayer model (Equation 3) , the Peppas model (Equation 4) and the Weibull model (Equation 5).

$$\frac{Q_t}{Q_\infty} = bt^{\frac{1}{2}} \quad (\text{Equation 2})$$

$$\frac{Q_t}{Q_\infty} = kt^n \quad (\text{Equation 3})$$

$$\frac{Q_t}{Q_\infty} = k_d t^m + k_r t^{2m} \quad (\text{Equation 4})$$

$$\frac{Q_t}{Q_\infty} = \left(1 - e^{-(at)^B}\right) \quad (\text{Equation 5})$$

In equations 2-5, Q_t is the accumulated mass dissolved at time t and Q_∞ is the mass dissolved at infinite time, b and k are kinetic constants and n is the diffusion exponent which depends on the release mechanism. Values of $n = 0.5$ indicate Fickian release, values of $0.5 < n < 1.0$ indicate an anomalous (non-Fickian or coupled diffusion/relaxation) drug release, and values of $n = 1.0$ indicate a case II (relaxation-erosion controlled) drug release.

In these equations, k_d is the diffusional constant, k_r is the relaxation constant and m is the diffusional exponent, the latter depends on the geometrical shape of the releasing device through its aspect ratio. a is a scale parameter and B is a shape parameter that can also indicate the release mechanism .

The parameters of each model were varied to obtain a wide range of dissolution profiles (85% of labeled drug dissolved in 30 to 300 min). For every condition, a set of values for 12 tablets was generated. Time points were established according to the following scheme:

$t_{85} \leq 40$ minutes	sampling every 5 minutes
$40 < t_{85} < 60$ minutes	sampling every 10 minutes
$60 \leq t_{85} < 90$ minutes	sampling every 15 minutes
$90 \leq t_{85} < 150$ minutes	sampling every 20 minutes
$t_{85} \geq 150$ minutes	sampling every 30 minutes

where t_{85} = time at which 85% of labeled drug is dissolved. According to the current guidelines only one time point with average dissolved drug higher to 85% is considered.

For each individual tablet, intrinsic (parameters of the model) and residual variability (experimental error), variability was included (Supplementary Data S1-S2).

2.2 Profile Comparison Tests.

In a typical dissolution profile comparison (Reference vs. Test) two Matrices (Reference and Test) of data points; $R (m \times n)$ and $T(m \times n)$ are evaluated, being m the number of tablets (normally 12) and n the number of time points sampled. The data for every tablet are expressed as a vector of length n .

2.2.1 Confidence Interval (CI):

Bootstrap 95% Confidence Intervals (CI) of f_2 were calculated similar to ones described in the literature . Initial simulations were performed to establish the number of repetitions to be used (S3). 5000 repetitions produced acceptable estimations and allowable computation time.

2.2.2 Permutation Test (PT):

Figure 1 illustrates the procedure for the Permutation Test (PT). In this procedure, the mathematical distance D_0 , between the two profiles, which is similar to f_2 is first calculated. The same principle used for f_2 of square difference between means at every time point is used but without any constant terms. This value is stored as the Original Distance or D_0 . Data from every tablet can be represented as a vector of length n , a set of $2 \times m$ vectors now representing the data of the Reference and Test batches. The first subset of m vectors represents the Reference batch and the last subset of m vectors represents the Test batch. After D_0 is calculated, the vectors are randomly sampled without replacement. In this way, each vector is relocated randomly in the Reference or Test subsets, creating two new ($m \times n$) matrices R_i and T_i . The same distance D between R_i and T_i is calculated and the value is stored as D_i . This cycle is repeated 5000 times and an empirical distribution of the D_i values (5000 values of D_i) is built. According to a predetermined type I error (typically, $\alpha = 0.05$), a rejection value that is greater than the 1-alpha percent of all D_i values in this empirical distribution is calculated. If the profiles are similar, D_0 is expected to be below this rejection value; D_0 above the rejection value indicates lack of similarity between the two profiles.

2.2.3 Tolerated Difference Test (TDT).

This test is based on a tolerated difference (δ) in dissolution between two tablets at each time point. Following the concept that there is some difference in percentage of dissolved drug that can be tolerated, this test attempts to statistically prove whether the differences between the Reference and Test samples exceed the predetermined tolerated difference or not.

Having at any time point the dissolved drug for m tablets from the Reference and m from the Test

$R_{t1}, R_{t2}, R_{t3}, \dots, R_{tm}$; R_{ti} = dissolved drug of i-th tablet from the Reference at time t

$T_{t1}, T_{t2}, T_{t3}, \dots, T_{tm}$; T_{ti} = dissolved drug of i-th tablet from the Test at time t

differences between all R_{ti} and T_{ti} are evaluated, and the number of events for which this difference is greater than the established tolerated difference (δ) is counted.

For a single time point, under the null hypothesis $\mu_{Tt} - \mu_{Rt} \leq \delta$ the random variable D_d

(the number of events in which difference is greater than δ) has a discrete distribution easy to calculate. For several time points the same procedure is followed but the statistic D_d is expressed as:

$$D_d = \frac{1}{n} \sum_{i=1}^n D_i \quad (\text{Equation 6})$$

where D_i = the sum of differences greater than δ at the i-th time point. In this work, values of δ = 5 (TDT-1) and δ = 10 (TDT-2) were analyzed. The distributions of D_d for $m = 12$ and $n = (1,2,3, \dots, 12)$ are available in the Supplementary Data (S4).

For $m = 12$ or less the exact discrete distribution of D_d can be calculated without difficulty for values of $n \leq 5$. For any increment in the value of n the computational time required to calculate the distribution of D_d is m^2 times longer, and not easily shortened by using parallel computing. In this work we calculated for $m = 12$ the exact distribution of D_d for values of $n = (1, 2, 3, 4 \text{ and } 5)$. Distributions for higher values of n were built by simulation with 100.000.000 repetitions. This value was sufficient to produce a distribution that differed less than 10e-5% from the exact distribution for the case of $n = 5$.

It must be considered that Equation 6 is completely operative only if all D_i are independent and identically distributed (iid) which may be not the case in a typical dissolution profile, because only one determination is allowed for each tablet and for each time point, this implies that for a typical case of 12 observations at 5 different time points, 60 individual tablets must be evaluated under independent conditions (iid-conditions). Although this may seem fatiguing, it could be well worth the effort in order to reduce or completely avoid *in-vivo* studies. Therefore, all explorations were made under normal conditions (typical dissolution profile with questionable independence) as well as under independent conditions (iid-conditions). The options of using only 6 and 3 observations per time point (options requiring 30 or 15 tablets, respectively, for 5 time points), were also evaluated. To simulate iid-conditions, a new tablet (with the same parameters and intrinsic and residual variability) was generated to estimate the dissolved drug at each time point; $n \times m$ tablets are needed and each tablet was evaluated just one time. Results were obtained under both conditions for all the comparison tests to evaluate the influence of independence in the comparisons. Additionally, sample size of $n = 6$ and $n = 3$ tablets were generated under iid-conditions to evaluate the influence of sample size in the comparisons.

2.3 Robustness and Power Explorations

Under conditions of similarity (equal parameters values in the model employed) pairs of Reference-Test batches were generated at different levels of variation. Every pair of batches was compared using the four procedures described (f2, CI, TDT and PT). At every level of variation 5000 pairs of batches were generated and the percentage of rejections (no similarity) was evaluated for each method. Ideally, under conditions that satisfy the null hypothesis (in this case, similarity), a robust statistical test does not increase the level of rejections at increasing levels of variation. In the best case, the level of rejections should be constant and very similar to the set type I error of the test (normally 5%) in order to quantify uncertainty. Variation in the models were generated including intrinsic and residual variability,

the 95% percentile of all the measured CV's at all-time points was recorded as CV_{95} as a measure of global variability.

Under conditions of non-similarity (different parameters values in the model employed) pairs of Reference-Test batches were generated at different levels of variation. As in the Robustness analysis, each pair of batches was compared using the four procedures described (f_2 , CI, TDT and PT). Differences in parameters were designed to produce values of t_{85} ranging from 60 to 300 minutes. For Korsmayer, Peppas and Weibull models in which more than one parameter describes the kinetic of the process, differences in single parameters (keeping the others constants) and bidirectional differences (varying two parameters simultaneously) were explored. At every condition, 5000 pairs of batches were generated and the percentage of rejections (%detections of no similarity) was evaluated for each method. More powerful tests are expected to detect smaller differences in the parameters used.

In preliminary experiments, stable (no difference with increment in repetitions) values of percentage of rejections for both robustness and power were found at 2000 repetitions, internal validation with sets of 2000 from the 5000 repetitions were also made and there was no difference in the results. All the analyses, simulations and statistical tests were performed using the R software environment for statistical computing and graphics from the R Foundation for Statistical Computing.

3. Results and Discussion

3.1 Robustness

All presented tests showed good robustness for standard ($CV_{95} \leq 0.1$) conditions. Figure 2 shows that the percentage of rejections remains under 0.05 for values of $CV_{95} \leq 0.12$ and sample size $n = 12$ tablets for all the tests under the four models. The robustness of the tests was always in the same order (from less robust to more robust) i.e., $TDT < CI < f_2 < PT$. For

conditions of high variability ($CV_{95} \geq 0.2$) just f2 and PT showed acceptable levels of rejections. PT was the only test in which the level of rejections remained under 5% for $CV_{95} \geq 0.3$ in all the models and conditions studied. Sample size and iid-conditions did not affect significantly the robustness of the tests, in all cases the rejection levels remained within the desired limits (≤ 0.05) for values of $CV_{95} \leq 0.1$ (Supplementary data S5-S8, ST1).

3.2 Power

Figure 3A illustrates how, under the Higuchi model, the level of rejections increases in all the tests when the difference between b_{Test} and $b_{\text{Reference}}$ becomes greater, in other words, when the simulated Reference and Test batches are more different. As shown here for the Higuchi model with sample size $n = 12$ and no-iid conditions, PT was the most powerful test, followed by TDT, CI and f2. These results were very similar regardless of the T_{85} of the Reference used in the simulations (Supplementary Data S9). Analogous results were obtained under the other models for differences in single parameters (Figure 3B-D and S10). The magnitude of the differences detected for single parameters under all the models are summarized in Table 1.

The capacity of the tests to detect simultaneous differences in more than one parameter (Power) is shown in figure 4, In this Power contour plots, two parameters are varied simultaneously (X and Y axis) and the combination of differences in these parameters required by each test to reach a power ≥ 0.8 is represented by a point on the contour plot. More powerful tests are able to detect smaller combination of differences with a power ≥ 0.8 (points closer to the origin on the diagram). Again, in these cases, PT was the most powerful test, detecting the smallest combination of differences (Points closer to the origin of the contour plots) of the parameters studied, followed by TDT, CI and finally f2 under all the models employed (Figure 4A-C, and S11-S13), highlighting that in the Peppas model, TDT and CI have very similar statistical power.

Consistent with the results for robustness, the power of the CI and TDT tests under high variability conditions was rather poor, f2 performance was slightly better but still not acceptable, and PT showed the best performance in this scenario although its power was significantly decreased compared to low variability conditions (Supplementary Data S9).

The demand for a very powerful and robust statistical tool, able to detect small differences in dissolution profiles can be satisfied with the introduced PT. PT was able to detect with statistical power ≥ 0.8 the smallest differences in each model parameters, normally more than two times smaller than the differences detected with the same power with f2 (Table 1). For example, in Korsmayer model, PT was able to detect differences of 4% in the kinetic constant while f2 is able to detect just differences greater than 20%, this can represent a 10% detectable difference in T85 with PT against a 40% detectable difference in T85 with f2.

As we have shown, PT can be used to compare profiles even with high levels of variation, moreover, PT allows the user to choose the level of statistical uncertainty. Furthermore, this test is not especially sensitive to the sample size employed in the comparisons, provided that the sample size is greater than $n = 3$ (due to the permutation nature of PT, the sample size of $n = 3$ highly compromised the power of the test and should not be employed S24-25). A sample size of $n = 6$ could be used without significantly altering its good performance compared to a sample size of $n = 12$. PT appears ideal for situations in which high similarity should be proven, e.g., in cases of drugs with a narrow therapeutic window, or with low permeability and/or solubility or susceptible of intestinal transport or metabolism, and currently there is no test as powerful and robust that can do so with similar statistical consistency.

As previously mentioned, In other situations, however, detection of significant but small differences in dissolution profiles may not be the objective and a more tolerant and flexible test is needed, This flexibility to vary the tolerated δ in order to detect larger or smaller differences in dissolution profiles is precisely one of the designed properties of the TDT test.

In addition TDT takes into account information on every tablet at every single point and does not rely on measures of central tendency as do f2, CI and PT, therefore, the analysis it provides may be more comprehensive than those of the other tests.

In Figure 4D TDT with $\delta = 5$ and $\delta = 10$ are explored and compared to f2. It can be appreciated that increasing the value of δ decreases the power of the test, in this particular case TDT with $\delta = 5$ was more powerful than f2, while TDT with $\delta = 10$ was less powerful than f2. Although iid-conditions had a minor effect on the power of the three tests, this did not alter the relative power of the tests (TDT-1 > f2 > TDT-2) in any of the studied models (S27-S29). The robustness of TDT for $\delta = 5$ was good and even better for higher values of δ (S30).

As previously stated, the underlying principle of the TDT demands that the data from every time point of every tablet be independent and identically distributed (iid-conditions). Effects of iid-conditions were analyzed and compared with no-iid conditions to determine how necessary iid-conditions are to a proper performance of the test. The effect of iid was shown to be of no practical importance because, in all the cases and models studied, the differences in power between iid and no-iid conditions were typically less or equal to 5% (Figure 3D, table 2 and S14-S29). In principle, the TDT test will perform similarly under no-iid conditions or iid-conditions and the former may be preferred for convenience (a smaller number of tablets is needed).

The effect of sample size on the power of the tests is also summarized in table 2 for each test and each dissolution model (all the graphical data is shown in S13-S25), in general, smaller sample sizes reduced the power of PT and TDT and increase the power of CI and f2. According to standard statistical theory, the power of a test increase with sample size and

should not be increased by reduction in sample size as happened with f_2 and CI in these simulations, it shows the limitations of the f_2 similarity factor..

Figures 3B and 3C (also S10 and S11) show some apparent discontinuities in the power curves of f_2 , CI and TDT tests. For example, in Figure 3B for $n_{\text{Reference}} = 0.5$, the power of f_2 , CI and TDT first increases continuously at increasing values of n_{Test} (Korsmeyer model), but at $n_{\text{Test}} = 0.5275$ (difference of 5.25%) the power of the three tests is reduced. This unexpected phenomenon was identified as an artifact due to sampling times. The time sampling scheme was designed to be as realistic as possible (intervals of 5, 10, 15, 20, or 30 min. see materials and methods section). According to this rules, solving equation 4 for $n_{\text{Test}} = 0.52625$, $t_{85} = 100.8161$, meaning that samples must be collected at 6 time points (20,40,60,80,100 and 120 min), in contrast, when $n_{\text{Test}} = 0.5275$, t_{85} is 99.72 and just 5 time points need to be sampled (20,40,60,80 and 100 min). This reduction in the number of sampling time points can produce a 50% decrease in power in the f_2 , CI and TDT tests. To counteract this effect, an optimized sampling scheme was developed. In optimized sampling, the number of time points is fixed, and t_{85} (the smaller between Reference and Test) is divided into equidistant time points with t_{85} as the last time point, for example, fixing 6 time points for a $t_{85} = 95$ min, the time points are: 15.833, 31.667, 47.5, 63.333, 79.167 and 95 min. Optimized sampling with 6 and 5 points, respectively, was employed in the analysis (Figure 5 and S10). In either case of such optimized sampling the discontinuity in power was no longer present. No significant difference was found between results obtained using 5 or 6 time points, confirming that the apparent discontinuity is due to the sampling strategy and not to the number of time points sampled. These finding suggest that optimized sampling should be employed as a first option for dissolution profile comparisons.

Combining the information presented here with basic principles of experimental design, we propose an experimental design driven strategy for performing optimal dissolution profile comparisons; this strategy is illustrated in Figure 6.

- If the goal is to detect small differences (table 2) in dissolution profiles, PT must be employed with a sample size greater or equal to $n = 6$ and a standard time sampling can be employed.
- If a less strict comparison is needed or if there is no certainty about the degree of similarity that can be accepted, the following procedure should be followed:
 - Preliminary experiments must be conducted to determine the t_{85} of the Reference and Test formulations and to fit the dissolution data to one or several models, (including models not presented in this work).
 - The minimum detectable difference (the difference to be considered as not similarity), must be determined either, by finding the adequate difference at t_{85} or t_{50} (time at which 50% of the labeled drug is dissolved) or a combination of both, or ideally, through either an *in-vitro/in-vivo* Correlation (IVIVc) models indicating the differences in dissolution that may lead to differences in bioavailability, or by fitting the dissolution data for preliminary experiments to available dissolution models, and estimating the difference in parameters acceptable as similar. This step is the most complicated and the most susceptible to produce under- or over-estimation of the detectable difference due to personal interpretation.
 - After the acceptable difference is determined and the dissolution model selected, simulations with TDT must be done to determine the value of δ and sample size at which the determined minimal difference in parameters

or t_{85} or possible combinations is detected with an acceptable statistical power (power of 0.8 or higher is recommended) always using an optimized time sampling strategy.

- Preferably, effects of iid-conditions on variability, robustness and power of TDT should be addressed in simulation or laboratory experiments.
- Finally, the dissolution assays must be performed under the conditions (iid/no-iid-conditions, δ and sample size) found and the statistical comparison must be done using TDT.

In this way the flexibility of TDT is used to customize a comparison test (setting a specific δ value) able to detect the differences in dissolution that can produce a difference in the bioavailability/bioequivalence of the formulations. The procedure described may seem arduous compared to the current f2 standard, however, it follows the typical procedure employed in any experimental design aimed at detecting significant differences with a quantified statistical uncertainty and known type-I and type-II errors. The procedure involves the following steps: i) preliminary data analysis, ii) determination of minimum detectable acceptable difference, iii) determination of sample size according to a desired level of power and robustness and iv) experimentation and statistical computation.

4. Conclusions and Outlook

Two new statistical tests, the Permutation Test (PT) and the Tolerated Difference Test (TDT), are presented for dissolution profile comparison in which type-I and type-II errors can be quantified, and have a stronger statistical basis than the current alternatives (e.g., the f2 similarity factor). The two new tests showed acceptable robustness at standard conditions of

variation ($CV_{95} \leq 0.1$). PT was the most robust and powerful test in all the conditions studied (even in conditions of high variability $CV_{95} \leq 0.2$ and reduced sample sizes). This test is strongly recommended for identifying small differences in dissolution (S30). For $\delta = 5$, TDT showed good robustness and very good power in all the conditions studied.

The impact of iid-conditions in TDT was not particularly large, therefore the more usual no-iid conditions could be employed (experimental confirmation of this is still pending). The possibility to modify the value of δ confers great versatility on TDT and allows it to be customized for any specific formulation. To make the best use of the two new tests, a strategy to design and perform a dissolution profile comparison is presented under typical premises of statistical experimental design. Finally, it was shown that optimized time sampling should be employed when possible to avoid artificial discontinuities in the statistical power of the tests, except for PT which is not susceptible to this effect.

Due to the simulated nature of the data presented here, experimental verification of the lack of effect for iid-conditions and examples of how to customize TDT with specific formulations are recommendable. Also, evaluation of additional available expressions for modeling drug dissolution, including models for controlled released mechanisms might be an obvious subject for future studies.

Acknowledgments: JDGM thanks Deutscher Akademischer Austauschdienst (DAAD) for financial support. We are grateful to Prof. Dr. Luisa Fernanda Ponce (Universidad Nacional de Colombia, Bogota, Colombia) , for her invaluable help in the conception of the project, and Dominik Selzer (Saarland University, Saarbruecken, Germany) for his help in computing and programming.

References

Amidon, G.L., Lennernas, H., Shah, V.P., Crison, J.R., 1995. A theoretical basis for a biopharmaceutical drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharmaceutical research* 12, 413-420.

Anand, O., Yu, L.X., Conner, D.P., Davit, B.M., 2011. Dissolution testing for generic drugs: an FDA perspective. *Aaps J* 13, 328-335.

Berry, M.R., Likar, M.D., 2007. Statistical assessment of dissolution and drug release profile similarity using a model-dependent approach. *J Pharm Biomed Anal* 45, 194-200.

Bolton, S., 1997. *Pharmaceutical Statistics*. Marcel Dekker.

Costa, P., Sousa Lobo, J.M., 2001. Modeling and comparison of dissolution profiles. *Eur J Pharm Sci* 13, 123-133.

Dahan, A., Miller, J.M., Hilfinger, J.M., Yamashita, S., Yu, L.X., Lennernas, H., Amidon, G.L., 2010. High-Permeability Criterion for BCS Classification: Segmental/pH Dependent Permeability Considerations. *Molecular pharmaceuticals*.

Dahan, A., West, B.T., Amidon, G.L., 2009. Segmental-dependent membrane permeability along the intestine following oral drug administration: Evaluation of a triple single-pass intestinal perfusion (TSPIP) approach in the rat. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences* 36, 320-329.

Dokoumetzidis, A., Papadopoulou, V., Macheras, P., 2006. Analysis of dissolution data using modified versions of Noyes-Whitney equation and the Weibull function. *Pharmaceutical research* 23, 256-261.

Dressman, J.B., Vertzoni, M., Goumas, K., Reppas, C., 2007. Estimating drug solubility in the gastrointestinal tract. *Adv Drug Deliv Rev* 59, 591-602.

Duan, J.Z., Riviere, K., Marroum, P., 2011. In vivo bioequivalence and in vitro similarity factor (f₂) for dissolution profile comparisons of extended release formulations: how and when do they match? *Pharm Res* 28, 1144-1156.

Fagerholm, U., Lindahl, A., Lennernas, H., 1997. Regional intestinal permeability in rats of compounds with different physicochemical properties and transport mechanisms. *J Pharm Pharmacol* 49, 687-690.

FDA, 1997a. *Guidance for Industry Extended Release Oral Dosage Forms: Development, Evaluation, and Application of In Vitro/In Vivo Correlations*, Rockville, MD 20857.

FDA, 1997b. *Guidance for industry, Dissolution testing of immediate release solid oral dosage forms*, Rockville, MD.

FDA, 1997c. *Guidance for Industry, 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, CMC 8*. Center for Drug Evaluation and Research (CDER), Rockville, MD 20857.

FDA, 2003. *Guidance for Industry, Bioavailability and Bioequivalence Studies for Orally Administered Drug Products — General Considerations*, Rockville, MD 20857.

Higuchi, T., 1963. Mechanism of Sustained-Action Medication. Theoretical Analysis of Rate of Release of Solid Drugs Dispersed in Solid Matrices. *J Pharm Sci* 52, 1145-1149.

Jiang, W., Kim, S., Zhang, X., Lionberger, R.A., Davit, B.M., Conner, D.P., Yu, L.X., 2011. The role of predictive biopharmaceutical modeling and simulation in drug development and regulatory evaluation. *Int J Pharm* 418, 151-160.

Kocic, I., Homsek, I., Dacevic, M., Parojcic, J., Miljkovic, B., 2011. An Investigation into the Influence of Experimental Conditions on <i>In Vitro</i> Drug Release from Immediate-Release Tablets of Levothyroxine Sodium and Its Relation to Oral Bioavailability. *AAPS PharmSciTech* 12, 938-948.

Korsmeyer, R.W., Gurny, R., Doelker, E., Buri, P., Peppas, N.A., 1983. Mechanisms of potassium chloride release from compressed, hydrophilic, polymeric matrices: effect of entrapped air. *J Pharm Sci* 72, 1189-1191.

Langenbucher, F., 1972. Linearization of dissolution rate curves by the Weibull distribution. *J Pharm Pharmacol* 24, 979-981.

Moore, J.W., Flanner, H.H., 1996. Mathematical Comparison of Dissolution Profiles. *Pharmaceutical Technology* 20, 64-74.

O'Hara, T., Dunne, A., Kinahan, A., Cunningham, S., Stark, P., Devane, J., 1997. Review of methodologies for the comparison of dissolution profile data. *Adv Exp Med Biol* 423, 167-171.

Pang, K.S., 2003. Modeling of intestinal drug absorption: roles of transporters and metabolic enzymes (for the Gillette Review Series). *Drug Metab Dispos* 31, 1507-1519.

Papadopoulou, V., Kosmidis, K., Vlachou, M., Macheras, P., 2006. On the use of the Weibull function for the discernment of drug release mechanisms. *Int J Pharm* 309, 44-50.

Peppas, N.A., Sahlin, J.J., 1989. A simple equation for the description of solute release. III. Coupling of diffusion and relaxation. *Int J Pharm* 57, 169-172.

Saranadasa, H., Krishnamoorthy, K., 2005. A multivariate test for similarity of two dissolution profiles. *J Biopharm Stat* 15, 265-278.

Sathe, P.M., Tsong, Y., Shah, V.P., 1996. In-vitro dissolution profile comparison: statistics and analysis, model dependent approach. *Pharmaceutical research* 13, 1799-1803.

Shah, V.P., Tsong, Y., Sathe, P., Liu, J.P., 1998. In vitro dissolution profile comparison--statistics and analysis of the similarity factor, f_2 . *Pharmaceutical research* 15, 889-896.

Siepmann, J., Siepmann, F., 2008. Mathematical modeling of drug delivery. *Int J Pharm* 364, 328-343.

Siepmann, J., Siepmann, F., 2011a. Mathematical modeling of drug release from lipid dosage forms. *Int J Pharm* 418, 42-53.

Siepmann, J., Siepmann, F., 2011b. Modeling of diffusion controlled drug delivery. *J Control Release*.

Vertzoni, M., Symillides, M., Iliadis, A., Nicolaidis, E., Reppas, C., 2003. Comparison of simulated cumulative drug versus time data sets with indices. *Eur J Pharm Biopharm* 56, 421-428.

Zhang, Y., Huo, M., Zhou, J., Zou, A., Li, W., Yao, C., Xie, S., 2010. DDSolver: an add-in program for modeling and comparison of drug dissolution profiles. *Aaps J* 12, 263-271.