

Paper 4655-2020

Transdermal Medicine – Challenges and Solutions with SAS®

Marina Komaroff, Noven Pharmaceuticals, Inc.

ABSTRACT

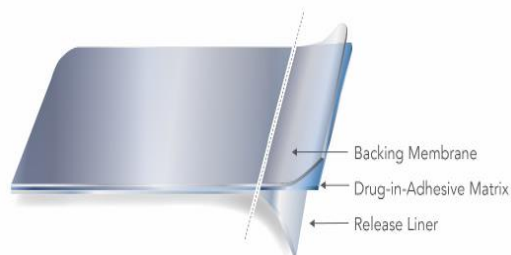
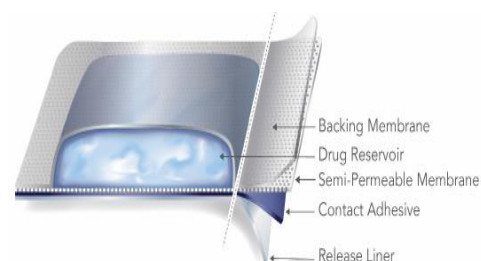
In 21 century Transdermal Drug Delivery System (TDDS or TDS) has become the fastest growing division in pharmaceutical industry because of innovative technologies and various advantages over the oral medicine. Fast growth came along with multiple challenges. The challenging dermal safety studies must be conducted to determine transdermal system performance, like adhesion, discomfort, irritation, and sensitization. Non-inferiority tests are also required to justify acceptable adhesion and irritation of a new product in comparison to an active reference. The rationale for data analyses and clinically meaningful choices for non-inferiority margins is still an area of research and development.

This paper concentrates on the analyses and reporting of transdermal data utilizing SAS. The author will: 1) review the existing measurement scales for transdermal safety; 2) highlight the difference in scales and requirements between regulatory agencies; 3) concentrate on the planning and conducting non-inferiority studies, and 4) demonstrate how p-values facilitate conclusions of superiority, non-inferiority, and bioequivalence of treatments. Sample size and power calculations will be demonstrated. The author is convinced that TDDS has not been fully appreciated as an effective alternative to oral and injection drug delivery methods, and hopes this paper helps to convince others that efficient transdermal studies can be conducted with confidence.

INTRODUCTION

The history of transdermal drug delivery goes back to the oldest existing medical record in human history. The use of ointments, balms, and patches with plant, animal or mineral extracts was popular in ancient time for cosmetic as well as therapeutic purposes. ^[1] In the middle of 20th century the first transdermal patch Scopolamine came to the United States market to treat the motion sickness. In 1986 the nicotine patch for smoking cessation became a **“blockbuster” in transdermal medicine**. The use of TDS became a common practice at the end of 20th century.

“Transdermal delivery systems are designed to deliver an active ingredient (drug substance) across the skin and into systemic circulation, while topical delivery systems are designed to deliver the active ingredient to local tissue.” ^[2] Transdermal or topical delivery system can be matrix type and liquid or gel reservoir type delivery systems (**Figure 1** and **Figure 2**; Source: FDA guidance for Transdermal and Topical Drug Delivery for Product Development and Quality Consideration ^[2]).

**Figure 1 Matrix Type****Figure 2 Reservoir Type**

TDS **“passively”** delivers the active substance(s) through the intact skin in a constant systemic absorption rate, where rate is dependent on the absorption through the skin or dissolving the active substance in a (semi solid) reservoir. It is a challenge to find the drug candidate for transdermal administration capable to penetrate the skin and sufficiently potent to be active. Those candidates have to be moderately lipotropic drugs (log P from 1 to 5) with low molecular weight (MW < 500 Da), and a low melting point (MP < 250C). The beginning of 21st century spiked with a progress in noninvasive or minimally invasive technology like iontophoresis, **microporation, sonophoresis and microdermabrasion, which are characterized as “active”** enhancement strategies due to permeation enhancement. The criteria for drug candidates is **not so demanding compared to “passive” TDS; however,** finding of suitable drug candidates for an appropriate cost-effective transdermal technology remains a challenge.

The pivotal pharmacokinetics studies are to determine drug product strength, dose proportionality between strengths if necessary, and to identify the residual active substance content. Pharmaceutical development should establish the links between the pharmacokinetic drug product properties and clinical efficacy including in vivo skin adhesion. In addition to traditional safety performance assessed by adverse events, vital signs, laboratory, ECG assessments, generic and novel TDS must also demonstrate acceptable level of dermal safety: adhesion, irritation, discomfort, pain, and adhesive residue. Each of these parameters has a specific scale, recommended time of assessments, and statistical methodology for analyses. The goal in this paper is to present the scales of measurements for each aspect of dermal safety, explain the specifics of design and analysis of skin adhesion, and irritation studies, and concentrates on the statistical analyses required for submission: non-inferiority tests for adhesion, and irritation studies. The author will stress the importance of margins of equivalence and non-inferiority and how they relate to sample size and power of the study. This should help the readers to perform their non-inferiority or equivalence trials with confidence.

To bring a generic transdermal product to market, a strict regulatory acceptance bioequivalence criteria must be fulfilled for test product (T) to demonstrate similar biopharmaceutical properties to a previously approved pharmaceutical reference product (R).^[3, 4] In addition, the majority of regulatory agencies require clinical endpoint studies to demonstrate therapeutic equivalence for generic topical products. Understanding the complexity and high cost of clinical endpoints studies, in the last decade the US Food and Drug Administration (FDA) expressed the acceptance, for some products, *in vitro* release test (IVRT) or *in vitro* permeation test (IVPT) as the alternative bioequivalence assessment methodology. This approach relaxed economic burden in clinical endpoint studies and increased investments in transdermal generic products by many companies. Another goal in this paper is to explain and demonstrate the statistical analyses required for Abbreviated New Drug Application (ANDA) submission: bioequivalence test for pharmacokinetics (PK) parameters.

DERMAL SAFETY

A TDS patch includes a backing sheet, impermeable to the active substance and to water; and a protective liner over the releasing surface of the patch that should be removed before applying the patch to the skin (Figure 1 and Figure 2). The drug product performance is dependent of the choice of excipients including the adhesives, backing layer and release liners and rate control membrane, their concentration and characteristics. There should not be adverse effects from the excipients, or exacerbation of the adverse effects from the active substance. The pressure-sensitive adhesives should assure the adhesion of the patch to the skin. The generic TDS products may utilize modern technologies that may not have been available at the time when the R was developed. For ANDA, it is necessary to balance the adhesive characteristics of the T with R through adhesiveness, cohesiveness and stability properties, and to ensure a consistent and uniform T adhesion of its entire surface area to the skin for the entire duration of wear.

In summary, to be safe and effective, the active substance(s) of T should be delivered through the skin at an adequate rate that is maintained for an appropriate time during patch wear, and should not generate or exacerbate the adverse effects of the active substance, irritate or sensitize the skin.

ADHESION

It is important for TDS to have a good adhesion because the amount of drug delivered into and through the skin partially depends on the dose surface area. If TDS loses its adherence during wear, the rate and extent of drug absorption, and the amount of drug delivered to the patient is uncertain. The recommended adhesion study design is a single-dose, randomized (including site of application), two-treatment, two-period crossover study where all subjects are dosed with the same strength of T and R TDS.

The evaluation of TDS adhesion can be performed with staggered or simultaneous application. Simultaneous application is preferable because it minimizes the intra-subject variability and **account for effect of subject's behavior on patch adhesion; however, simultaneous application** may not be acceptable if daily dose with applied patches is exceeded.

The adhesion studies can be standalone or added as another primary objective of the (1) PK studies using crossover application of T and R (usually used at highest strength), or (2) with the first application of irritation studies that designed with simultaneous application of T and R (usually at lower strength). For the R, a single dose, full-size application for the dose duration should be used. For the T, a single application of the largest patch size is usually used, but single or multiple application of smallest patch size is also acceptable. European Medicines Agency (EMA) recommends to test smallest and largest patch sizes.^[5] FDA assumes that adhesion performance is independent of patch size; however, if ratio of active drug area to peripheral adhesive varies with patch size, then smallest and largest patches should be tested.

Adhesion scoring scales differ for EMA and FDA (Table 1). The assessment should be done by trained evaluator (preferable the same through the study), or using dot-matrix assessment which is more objective approach.

The FDA recommended *primary endpoint* for evaluating adhesion as the overall equally-spaced mean adhesion score (MAS) derived for a T from individual adhesion scores at each assessment time point averaged across all the equally spaced time points (except the baseline). The weights corresponding to the interval lengths should be used in case on unequally-spaces time points of assessments.

The *secondary endpoints* (descriptive statistics only) have to assess the potential treatment group difference in clinically meaningful extreme values or events using the 5-point adhesion scale ^[6]:

- Proportion of subjects with an adhesion score >2 at any time point
- Proportion of subjects with T mean adhesion score greater than R by 1 or more points, compared to the proportion of subjects with an R mean adhesion score greater than the corresponding T mean adhesion score by 1 or more
- Time to adhesion score > 2 compared between T and R using a Kaplan Meier cumulative incidence can be plotted if sufficient number of events.

To be approved, the T product must be non-inferior (NI) to R with regards to MAS and with a non-inferiority (NIMa) **margin of 0.15 ($\delta = 0.15$); and also show no meaningful difference** with regard to degree of detachment. ^[6] A sufficient number of subjects should be enrolled to power the study at a level of 0.80 or higher (type 1 error=0.05). A larger sample size than what might be ordinarily calculated is recommended in order to ensure the validity of any large-sample Gaussian assumptions. ^[6]

EMA ^[5]	FDA ^[6]
Adhesion Scale	
<p>7-score system for adhesion of transdermal patches should be scaled in 5 % increments: 0=more than 95 % of the patch area adheres 1= more than 90 % of the patch area adheres 2=more than 85 % of the patch area adheres 3=more than 80 % of the patch area adheres 4=more than 75 % of the patch area adheres 5=more than 70 % of the patch area adheres 6=less than 70 % adheres or patch detachment is regarded as significant patch adhesion failure</p>	<p>5-score system 0 = ≥ 90% adhered (essentially no lift off the skin) 1 = ≥ 75% to < 90% adhered (some edges only lifting off the skin) 2 = ≥ 50% to < 75% adhered (less than half of the patch lifting off the skin) 3 = > 0% to < 50% adhered but not detached (more than half of the patch lifting off the skin without falling off) 4 = 0% adhered - patch detached (patch completely off the skin)</p> <p>Additionally, the following information should be collected:</p> <ul style="list-style-type: none"> • At each time point when adhesion is assessed on the above described 5-point scale, the actual percent adherence estimate (e.g., if the observer scores =2, the estimates can be 60% percent adhered, a score of two and a 60% should be recorded for that time point). • Photographic evidence showing the extent of TDS adherence to the skin at each time point should be provided.
Data	
<p>Consider adhesion failure if score less than 70% of the patch area adheres or score = 6. In general:</p> <ul style="list-style-type: none"> • a mean adhesion score of greater than 90% should be expected (or score < 2) and • No instances of detachment should be seen; Poor adhesion events should be investigated and possible causes and risk factors determined 	<p>Consider adhesion failure defined by a meaningful degree of detachment when adhesion score ≥3</p> <p>In general, consider:</p> <ul style="list-style-type: none"> • the number of subjects that experience detachment (score=4) or meaningful degree of detachment (scores= 3 or 4), and • how early in the application period those unacceptable scores are observed

Table 1. Adhesion

IRRITATION

The application site is examined for signs of skin irritation before patch application and at time points after patch removal. All subjects will be evaluated by a trained skin evaluator using the Berger and Bowman scale that consists of two parts: "Dermal Response" (Table 2), and "Other Effects" (Table 3). ^[7] If reactions fall between the unit grades than more severe of the two grades should be assigned.

Score	Skin Appearance
0	No evidence of irritation
1	Minimal erythema, barely perceptible
2	Definite erythema, readily visible; or minimal edema or minimal papular response
3	Erythema and papules
4	Definite edema
5	Erythema, edema, and papules
6	Vesicular eruption
7	Strong reaction spreading beyond test (i.e., application) site

Table 2. Dermal Response Rate

Score (Numeric Equivalent)	Observation
N (0)	No other effects observed
A (0)	Slightly glazed appearance
B (1)	Marked glazed appearance
C (2)	Glazing with peeling and cracking
F (3)	Glazing with fissures
G (3)	Film of dried serous exudates covering all or part of the patch site
H (3)	Small petechial erosions and/or scabs

Table 3. Other Effects

A combined irritation score is calculated by adding the "Dermal Response" score and the **numeric equivalent for "Other Effects" letter score**. The primary endpoint for analysis is mean cumulative irritation score (MIS) that calculated as the sum of all combined scores observed at each observation divided by the total number of observations. The T product should be statistically non-inferior to the R based on evaluating the **difference in the products' overall mean MIS, with a Non-Inferiority Margin of Irritation (NIMi) of 0.20 ($\delta = 0.20$)**. [7,8]

OTHER DERMAL SAFETY ASSESSMENTS

It is expected that a trained individual evaluates the patch application site for adhesion, irritation, discomfort, pain, and adhesive residue. The evaluator observes the TDS application site and interviews the subject. Post-removal, the evaluator assesses the application site and the surrounding area for adhesive residue and irritation. In order to ensure reliability between successive ratings, only one evaluator has to follow each subject through the completion of the study. If this was not possible, every effort should be made to limit the number of evaluators to 2 responsible for evaluation of an individual subject.

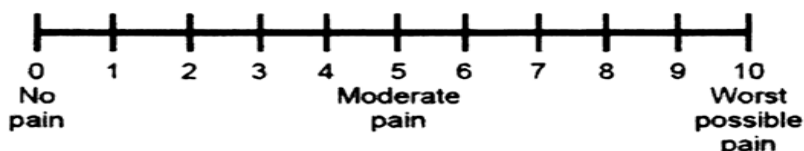
Discomfort: Discomfort assessments are performed at the scheduled time points by asking the subject, "Are you experiencing any discomfort related to the patch?" If no, the overall level of discomfort is rated as 0. If yes, the evaluator should ask the subject to rate the discomfort as mild, moderate, or severe and record as presented in Table 4. If the discomfort was mild, moderate, or severe, the evaluator asks **the subject further, i.e., "Describe your**

discomfort.” The subject’s verbatim description of discomfort should be recorded, the most common are itching, burning, etc.

Score	Definition
0	No discomfort
1	Mild discomfort
2	Moderate but tolerable discomfort
3	Severe, intolerable discomfort
4	Patch not present

Table 4. Discomfort

Pain: Subjects are asked to rate whether they had any pain at the patch site using a 0-10 point Numeric Pain Rating Scale



Adhesive Residue: Immediately following (within 1 minute) removal of the patch, the amount of adhesive remaining at the application site should be examined and graded according to the scale shown in Table 5.

Score	Definition	
0	None	0%
1	Light	<25% of patch site
2	Medium	≥25% to ≤75%
3	Heavy	>75% of patch site

Table 5. Adhesive Residue

Adverse events at application sites: Any spontaneous complaints of adverse reaction including dermal reactions at unscheduled times should be recorded. They can include, but not limited to initial dermal response, exacerbation of an existing response, discomfort, or pain. If adverse event happened at application site, it should be marked as adverse event at application site, and summarized and listed separately.

In the event that a patch was removed early, discomfort and adhesion should be assessed prior to removal as an unscheduled assessment. Irritation assessments after patch removal should be performed as scheduled relative to the time of actual removal. For any patches that detached prior to the scheduled time of removal, irritation assessments are to be performed as scheduled relative to the time of detachment.

TESTING REQUIREMENTS

1. The requirements for Adhesion and Irritation studies are Non-Inferiority tests with regard to mean adhesion scores and with a NI margin of 0.15 (NIMa = 0.15) for adhesion studies, and mean irritation scores with NI margin of 0.20 (NIMi = 0.20) for irritation studies. The goal is to justify that T product is non-inferior than R in terms of adhesion and irritation properties.

2. In 21 CFR part 320, the requirements are set forth for submitting bioavailability (BA) and bioequivalence (BE) data in the investigational new drug applications (INDs), new drug applications (NDAs), abbreviated new drug applications (ANDAs), and supplements, the definitions of BA and BE, and the types of *in vivo* studies that are appropriate to measure BA and establish BE.

BE compares PK parameters of a T versus R drug product, where results of comparisons rely on (1) a criterion, (2) a confidence interval for the criterion, and (3) a predetermined BE limit. To establish BE, the calculated confidence interval for the ratio of the products' geometric means of the measures should fall within a BE limit defined as 80-125%. [3,4] BE comparisons is also used in certain pharmaceutical product line extensions, such as additional strengths, new dosage forms (e.g., changes from immediate release to extended release), and new routes of administration. [3,4]

In permeation studies, the permeation rate (flux) should be calculated for each diffusion cell and the mean flux reported together with the corresponding standard deviation (SD), coefficient of variation. For the comparison of products, relevant permeation parameters (flux) should be statistically compared using the same approach: the 90% confidence interval for the ratio of the two products should be contained within the 80-125% limit, unless justified.

For statistical analyses of bioequivalence studies, the PROC MIXED in SAS is used having treatment sequence, subject, period, and treatment as independent, and the response measure (e.g., log(AUC), log(Cmax)) as dependent parameter. [3]

There are many papers that particularly review statistical methodology for superiority, non-inferiority, and equivalence testing in SAS/STAT® software. However, multiple issues arise in the literature with application and interpretation of such methods.

That's why the author's choice was to concentrate on statistical aspect of the testing, and classical interpretation by using traditional p-values.

SUPERIORITY

In a superiority trail, the T product should demonstrate the superiority versus R or placebo.

The hypotheses for mean difference between treatment groups can be determined in two ways:

A. $H_0: \mu_T - \mu_R = 0$ (there is no difference between treatments)

$H_a: \mu_T - \mu_R \neq 0$ (there is a difference between treatments).

The null hypothesis is rejected based on the two group 2-sided t-test ($\alpha=0.05$) when the p-value < 0.05 .

B. $H_0: \mu_T - \mu_R = 0$

$H_a: \mu_T - \mu_R < 0$ (the beneficial response from T when it is smaller than R).

The null hypothesis is rejected based on the two group 1-sided t-test ($\alpha=0.05$) when the p-value < 0.05 . SAS code: Left-Tailed t-test for $\Delta = \mu_T - \mu_R$.

NON-INFERIORITY

A non-inferiority trail aims to demonstrate that T is not worse than R by more than a pre-specified an acceptable small amount called *non-inferiority margin* (NIM). It implies that the statistical test should be one-sided, and treatment T could be even better (superior) than R [1-5].

The hypotheses for mean difference between treatment groups:

$$H_0: \mu_T - \mu_R = NIM \Rightarrow H_0: \mu_T - \mu_R - NIM = 0$$

$$H_a: \mu_T - \mu_R < NIM \Rightarrow H_a: \mu_T - \mu_R - NIM < 0$$

(the beneficial response from T if it is smaller, or bigger than R by NIM points or less).

The statistical test is similar to the last example of superiority where the null hypothesis is rejected based on the two group 1-sided t-test ($\alpha=0.05$) having the p-value < 0.05 . SAS code: Left-Tailed t-test for $\Delta = \mu_T - \mu_R - NIM$.

EQUIVALENCE

Clinical equivalence trials are active control trials with the objective to demonstrate that new treatment is equivalent to active control with an acceptable level called *margin of equivalence* (ME). ^[1-5]

$$H_0: |\mu_T - \mu_R| = ME \Rightarrow H_{01}: \mu_T - \mu_R + ME \leq 0 \text{ and } H_{02}: \mu_T - \mu_R - ME \geq 0$$

$$H_a: |\mu_T - \mu_R| < ME \Rightarrow H_{a1}: \mu_T - \mu_R + ME > 0 \text{ and } H_{a2}: \mu_T - \mu_R - ME < 0$$

Only when both H_{01} and H_{02} are rejected, then equivalence is concluded. It means that the difference in effect falls within the pre-specified margins of equivalence. SAS code: Two One-Sided T-test (TOST), **each with $\alpha=0.05$** .

APPLICATIONS

The imaginary new treatment T was developed by company X for pain management. The properties of adhesion and irritation need to be established in comparison with the referenced listed drug R.

ADHESION: NON-INFERIORITY TEST

To simplify this example, Adhesion study was a single-dose, randomized, two treatments, parallel group design; and Adhesion data for T and R TDS were simulated and summarized in Output 1.

Treatment	Adhesion					
Frequency Percent	0 >=90%	1 >=75 - <90%	2 >=50 - <75%	3 > 0 - <50%	4 0% adhered	Total
T	22 44%	19 38%	7 14%	0 0%	2 4%	50
R	26 52%	15 30%	4 8%	5 10%	0 0%	50

Output 1. Output from a PROC FREQ procedure

The hypotheses for mean difference between treatment groups and $NIMa=0.15$ are following (where μ is Mean Adhesion Score):

$$H_0: \mu_T - \mu_R = 0.15 \Rightarrow H_0: \mu_T - \mu_R - 0.15 = 0$$

$$H_a: \mu_T - \mu_R \leq 0.15 \Rightarrow H_a: \mu_T - \mu_R - 0.15 \leq 0$$

(the beneficial response from T if it is smaller, or bigger than R by 0.15 points or less).

The null hypothesis is rejected based on the two group 1-sided t-test ($\alpha=0.05$) when the p-value < 0.05 . SAS code: Left-Tailed t-test for $H_0=0.15$:


```

PROC TTEST DATA=MYDS CI=NONE SIDES=L HO=0.15;
CLASS TREATMENT;
VAR ADHESION;
RUN;

```

Mean Adhesion scores, standard deviation and 95% Confidence Limit (CL) are presented in Output 2. Mean difference between T and R is equal to 0.06.

Treatment	Mean Adhesion Score	95% CL Mean		Standard Deviation
Test	0.8200	0.5465	1.0935	0.9624
Reference	0.7600	0.4813	1.0387	0.9806
Diff (T-R)	0.0600	-Infty	0.3827	0.9715
Method	Variances	DF	t Value	Pr < t
Pooled	Equal	98	-0.46	0.3221

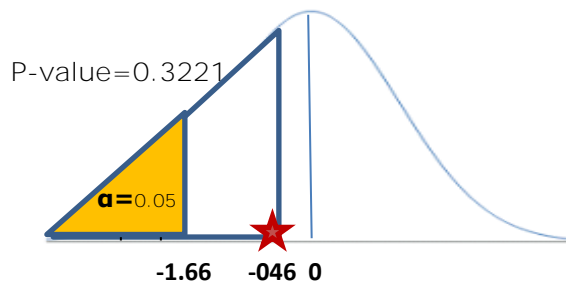
Output 2. Output from a PROC TTEST

Calculate T-value = -0.46 using a formula:

$$t = \frac{\bar{x}_T - \bar{x}_R - 0.15}{\sqrt{\frac{\sigma_T}{n_T} + \frac{\sigma_R}{n_R}}}$$

Calculate critical t as T-critical=TINV (0.05, 98) = -1.66. Because T-value > T-critical, the p-value=0.3221 > 0.05. Thus, null hypothesis cannot be rejected, and we cannot conclude that T is non-inferior than R with NIMa=0.15 (Figure 3).

Figure 3. Non-inferiority test for Adhesion



IRRITATION: NON-INFERIORITY TEST

In this example, Irritation study was a randomized, two treatments study design where T and R were applied simultaneously to each subject. Irritation data for T and R were simulated and summarized in Output 3.

Treatment	Combined Irritation Scores					
Frequency /Percent	0	1	2	3	4	Total
Test	23 46.0%	15 30.0%	4 8.0%	8 16.0%	0 0%	50

Treatment	Combined Irritation Scores					
Frequency /Percent	0	1	2	3	4	Total
Reference	29 58.0%	10 20.0%	5 10.0%	2 4.0%	4 4.0%	50

Output 3. Output from a PROC FREQ

The hypotheses for difference between mean irritation scores of treatment groups and NIMI=0.20 are following (where μ is Mean Combined Irritation Score):

$$H_0: \mu_T - \mu_R = 0.20 \Rightarrow H_0: \mu_T - \mu_R - 0.20 = 0$$

$$H_a: \mu_T - \mu_R < 0.20 \Rightarrow H_a: \mu_T - \mu_R - 0.20 < 0$$

(the beneficial response from T if it is smaller, or bigger than R by less than 0.20).

The null hypothesis is rejected based on the paired 1-sided t-test ($\alpha=0.05$) when the p-value < 0.05. SAS code: Left-Tailed t-test for $H_0=0.20$:

```
PROC TTEST DATA=MYDS CI=NONE SLIDES=L HO=0.20;
  PAIRED TRT1*TRT2;
RUN;
```

Mean Combined Irritation scores, standard deviation and 95% Confidence Limit (CL) are presented in Output 4. Mean difference between T and R is equal to -0.20 because it is smaller for T versus R.

Treatment	Mean Irritation Score	95% CL Mean		Standard Deviation
Diff (T-R)	-0.2000	-Infity	0.1387	1.4286
		DF	t Value	Pr < t
		49	-1.98	0.0267

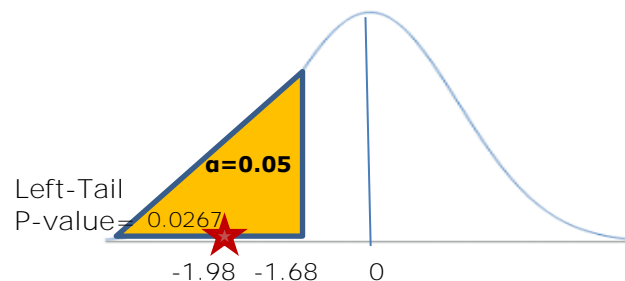
Output 4. Output from a PROC TTEST

Calculate T-value for the same formula = -1.98

$$t = \frac{\bar{x}_T - \bar{x}_R - 0.20}{\sqrt{\frac{\sigma_{diff}}{n}}}$$

and T-critical=TINV (0.05, 98) = -1.68. Because T-value < T-critical, the p-value=0.0267 < 0.05. Thus, null hypothesis is rejected, and we conclude that T is non-inferior than R with NIMI=0.20 (Figure 4).

Figure 4. Non-inferiority test for Irritation



EQUIVALENCE

Clinical equivalence trials are trials with the objective to demonstrate that new treatment T is equivalent to T with an acceptable level called *margin of equivalence* (ME). [2-5] Statistical analysis for pharmacokinetic measures, such as area under the curve (AUC) and peak concentration (Cmax), should be based on the *two one-sided tests procedure* to determine whether the average values for the pharmacokinetic measures determined after administration of the T and R products were comparable. To establish BE, the calculated 90% confidence interval should fall within a BE limit: 80-125% for the ratio of the product geometric means.

The null and alternative hypotheses for bioequivalence are based on the ratios of means:

Lower limit: $H_0: \mu_T/\mu_R = 0.80$; and Upper limit $\Rightarrow H_0: \mu_T/\mu_R = 1.25$

$H_a: \mu_T/\mu_R \geq 0.80$ $\Rightarrow H_a: \mu_T/\mu_R \leq 1.25$

which (assuming $\mu_R > 0$); where μ_T is the mean response for the T product and μ_R is the mean response for the R for any PK parameter: Cmax, AUC, etc..

After logarithmic transformation of data and keeping in mind that $ME = \text{Log}(1.25) = \text{Log}(0.80) = 0.22314$, the hypotheses of equivalence can be rewritten as following:

$H_0: |\log(\mu_T/\mu_R)| = ME \Rightarrow H_{01}: \log\mu_T - \log\mu_R + ME \leq 0$ and $H_{02}: \log\mu_T - \log\mu_R - ME \geq 0$

$H_a: |\log(\mu_T/\mu_R)| < ME \Rightarrow H_{a1}: \log\mu_T - \log\mu_R + ME > 0$ and $H_{a2}: \log\mu_T - \log\mu_R - ME < 0$.

Only when both H_{01} and H_{02} are rejected, then equivalence is concluded. It means that the difference in effect falls within the pre-specified margins of equivalence (-0.22314, 0.22314).

SAS code: Two One-Sided T-test (TOST) for cross-over design, **each with $\alpha=0.05$** :

```
PROC TTEST DATA=MYDS CI=NONE TOST = (-0.22314, 0.22314);
  PARIED LCMAX1*LCMAX2;
RUN;
```

Please note, data was transposed for each subject to get two variables with C_{max} log-transformed values for T as LCMAX1 and for R as LCMAX2. The results are presented in Output 5, Output 6, and Output 7.

T-R Mean	95% CL Mean		Std Dev
0.2928	0.1548	0.4309	0.3193

Output 5. Output from a paired PROC TTEST: Difference LCMAX1-LCMAX2

T-R Mean	Lower Bound		90% CL Mean			Upper Bound	Assessment
0.2928	-0.223	<	0.1785	0.4071	>	0.2231	Not equivalent

Output 6. Output from a paired PROC TTEST: TOST Level 0.05 Equivalence Analysis

Calculate T-value = 7.75 and 1.05 (using a formula from adhesion section) for $ME = -0.2231$ and 0.2231, respectively; and calculate critical t as $T\text{-critical} = \text{TINV}(0.05, 22) = -1.72$, and $\text{TINV}(0.95, 22) = 1.72$. For Right-tail test, because $T\text{-value} = 7.75 > T\text{-critical} = 1.72$, the $p\text{-value} < 0.001 < 0.05$.

Test	Null	DF	t Value	P-Value
Upper	-0.223	22	7.75	<.0001
Lower	0.2231	22	1.05	0.8467

Output 7. Output from a paired PROC TTEST: TOST Level 0.05 Equivalence Analysis

Thus, null hypothesis is rejected for one-side Right-tail test. For Left-tail test, because T-value = 1.05 > T-critical = -1.72, the p-value = 0.8457 > 0.05, and null hypothesis is not rejected. Thus, because one out of two null hypotheses cannot be rejected, we cannot conclude that T is equivalent to R with ME=0.22134.

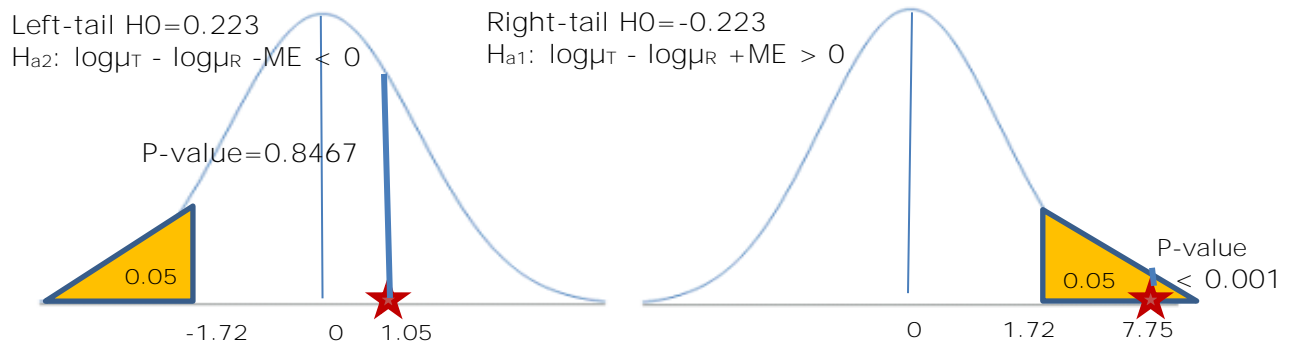


Figure 5. Equivalence: Left-tail and Right-tail tests

FDA recommends to perform an analysis of variance (ANOVA) on PK parameters using PROC MIXED in SAS. The PK parameters should be natural logarithm transformed and used as a dependent variable in the model. The model includes a fixed effect for treatment, sequence and period, and a random effect for subject within sequence. From this model the least-squares means (LSMeans) for each treatment will be calculated. The ratios of the LSMeans for the comparison of the T to the R will be obtained in the second step by exponentiation. The bioequivalence will be established if the CI falls inside the limits: 80-125%. The following program statements were used on the log-transformed data:

```
PROC MIXED DATA=MYDS;
  CLASSES SEQ SUBJID APERIOD TRT;
  MODEL LCmax = SEQ APERIOD TRT/ DDFM=KR;
  RANDOM SUBJID(SEQ);
  ESTIMATE 'T vs. R' TRT 1 -1/E CL ALPHA=0.1;
  LSMEANS TRT/DIFF CL ALPHA=0.1;
  ODS OUTPUT ESTIMATE=ESTIMATE COVPARAMS=COV;
RUN;
```

From the Output 8, the mean difference and 90% CI interval is exactly the same as in Output 6 using t-test. After performing exponentiation for ratio Exp (0.2930) = 134.04% and Lower and Upper bounds of CI 90% (119.66%, 150.15%), the conclusion is the equivalence cannot be established because the right bound falls beyond 125%.

Effect	Estimate Diff	Alpha	Lower	Upper
T/R	0.2930	0.1	0.1795	0.4065

Output 8. Output from a paired PROC MIXED

The use of T-test is more straight forward and does not require exponentiation as extra step. On the other hand, PROC MIXED gives you another good estimate of intra-subject variability, calculated as 100%*SQRT(EXP(MSE)) -1), where MSE is mean square error. In our example, it was Intra-CV%= 22% that may be helpful for power analyses and sample size calculation.

MARGINS, SAMPLE SIZE AND POWER

A pharmaceutical industry follows FDA guidance for margins of non-inferiority for adhesion and irritation studies, and to establish bioequivalence. However, the choice of margins relies on statistical and clinical judgement. [9,10,11] NI studies should demonstrate that T product has effect that is not substantially inferior to the R. Thus, NI margins should be as tight as possible, and the rationale of choice should be presented in the protocol. From statistical point of view, the choice of margin influence sample size and power of the study. [9,10,11] For example, if we relax margins of equivalence in the last example as ME=0.35 instead of 0.2231, then sample size to reach of BE will be 282 (power =80%), and 410 (power= 90%). For adhesion example, it would require 1442 subjects per arm to demonstrate non-inferiority for adhesion data with NIMa=0.15, but only 204 if NIMa=0.30. Using SAS code, sample size is calculated for different margins for non-inferiority and equivalence testing:

```
PROC POWER; * CALCULATE SAMPE SIZE FOR NON-INFERIORITY***;
  TWOSAMPLEMEANS TEST=DIFF
  NULLDIFF = 0.15 TO 0.30 BY 0.05
  MEANDIFF = 0.06
  SIDES = L
  ALPHA = 0.05
  STDDEV= 0.9715
  NPERGROUP=.
  POWER = 0.80;
RUN;
PROC POWER; * CALCULATE SAMPE SIZE FOR EQUIVALENCE***;
  TWOSAMPLEMEANS TEST=EQUIV_DIFF
  LOWER = -0.15
  UPPER = 0.15
  MEANDIFF = 0.06
  ALPHA = 0.1
  STDDEV= 0.9715
  NPERGROUP=.
  POWER = 0.80;
RUN;
```

The results from both are summarized in Table 6. There are a few important observations: (1) wider margins decrease sample size; (2) more power requires bigger sample size; and (3) more sample size is required for non-inferiority versus equivalence testing.

Power	Margin	Non-Inferiority Sample size per arm	Equivalence Sample size per arm
0.80	0.15	1442	1052
0.80	0.20	597	440
0.80	0.25	324	246
0.80	0.30	204	160
0.90	0.15	1997	1532
0.90	0.20	826	634
0.90	0.25	449	347
0.90	0.30	282	221

Table 6. Margins, Sample Size and Power for Non-Inferiority and Equivalence

CONCLUSION

This paper highlights the history and current challenges in transdermal industry. It describes the assessments of TDS performance. The author particularly concentrates on statistical analysis for adhesion, irritation and bioequivalence, and clarified the hypotheses and the meaning of p-values for non-inferiority and equivalence testing. The application and interpretation was presented for comparing means for two treatment groups on simulated data. Nevertheless, the same logic should be used for other types of data.

REFERENCES

1. Pastore MN, Kalia YN, Horstmann M, et. Al. Transdermal patches history, development and pharmacology, *Br J Pharmacol*. 2015; 172 (9): 2179-2209
2. FDA: Transdermal and Topical Delivery System – Product Development and Quality Consideration; November 2019
3. FDA: Statistical Approaches to Establishing Bioequivalence; January 2001
4. Miranda M, Sousa JJ, Veiga F, Cardoso C, Vitorino C: Bioequivalence of topical generic products, Part Paving the way to a tailored regulatory system. *Eur J Pharm Scie*, 2018 Sep 15: 122: 264-272
5. EMA: Guideline on quality of transdermal patches - 23 October 2014
6. FDA: Assessing Adhesion with Transdermal Delivery Systems and Topical Patches for ANDAs, Draft Guidance for Industry, June 2016
7. Berger, RS, Bowman, JP. A reappraisal of the 21-day cumulative irritation test in man. *J. Toxicol. Ot. & Ocular Toxicol*. 1982; 1(2); 109-115.
8. FDA: Assessing the Irritation and Sensitization Potential of Transdermal and Topical Delivery Systems for ANDAs; October 2018
9. FDA Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness; November, 2016
10. Marina Komaroff, 2019; Equivalence, Superiority and Non-inferiority with Classical Statistical Tests: Implementation and Interpretation; PharmaSUG 2019 - Paper SA-91; <https://www.pharmasug.org/proceedings/2019/ST/PharmaSUG-2019-ST-091.pdf>
11. Marina Komaroff, Sailaja Bhaskar (2014); Defining Non-Inferiority Margins for Adhesion Studies; PharmaSUG2014, San Diego, CA; Available at <http://www.pharmasug.org/proceedings/2014/SP/PharmaSUG-2014-SP03.pdf>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Marina Komaroff, Dr.P.H.
Director-Biostatistics, Product Development
Noven Pharmaceuticals, Inc.
mkomaroff@noven.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.