Paper AS10

Prentice-Wilcoxon Test to Study Time to Healing for Split Face, Double Blind, Dermatology Clinical Trials

Shrutimita Pokhariyal, Symbio, Chandigarh, India
Craig Wesselman, Symbio, Colorado, USA

ABSTRACT

Dermatology clinical trials often have randomized, blinded studies that have a split face comparison for two treatments, usually investigational medicinal product (IMP) and a vehicle or active comparator. In indications such as rosacea, acne vulgaris, healing after cosmetic treatment etc., a participant is enrolled in the study and is randomly allocated the study treatment to each side of the face. The objectives of such clinical trials are to assess the safety and efficacy of the IMP and one of the clinical endpoints can be time to healing. While time to event is a common endpoint in drug development, tests such as Log-Rank or Gehan test prove to be inadequate when within subject correlation is involved. This paper will talk about the Prentice-Wilcoxon Test used to test for differences between the time to first healing between two treatments.

INTRODUCTION

Topical medications are commonly used as the first line of treatment for various skin conditions. They are also used for post-procedure care after aesthetic treatments such as ablative fractional resurfacing. When it comes to dermal fillers used for enhancing facial symmetry, they must undergo testing to obtain regulatory approval. This testing is carried out through controlled, randomized, evaluator-blind clinical trials under an Investigational Device Exemption (IDE) application. In aesthetics and dermatology trials, a split-face design is often employed to compare the safety and effectiveness of different products.

For statisticians, to plan an effective analysis for a study, it is important to understand the design of experiments and the implications of statistical tests to be involved for the optimal outcome. In a systematic review focusing on design and methodological issues of split-body randomized controlled trials, 54 trials (published between 2017 and 2021) were included, and data was extracted to study with the aim of providing a methodological overview of split-body randomized trials in dermatology. It was highlighted that 14 out of 54 studies used a test for independent observations, thus ignoring the correlation during analysis. A total of 33 articles described a sample size calculation out of which 6 studies accounted for a within-patient between-lesions correlation, without reporting the value of the assuming correlation. Out of 54, 33 studies performed a paired test, and one study fitted a mixed effects model to consider between-lesion correlation. No study reported the within-patient between-lesions correlation coefficient estimate. Thus, the key message of the systematic review was that the planning and reporting of within-person trials in dermatology are not optimal. [**Reference**: 2]

Let us briefly study the key design features, advantages and challenges of split body design trails.

DESIGN AND KEY FEATURES

Split-body clinical trials involve administering different interventions or treatments to separate, anatomically distinct areas of the same participant. This design enables researchers to compare the effects of treatments within the same individual. The intent of using this design may be to remove the effect of the factors confounded with the treatment effect or to decrease the sample size.

Advantages:

1. Reduced Variability: Using the same individual for both the treatment and control conditions minimizes variability due to individual differences.
2. Ethical Considerations: It allows researchers to study the effects of treatments within the same person, potentially reducing the need for larger sample sizes and minimizing exposure to placebo or control conditions.

Applications:

Split-body trials are particularly useful in various areas of clinical research:

1. Dermatological Studies: Different areas of the body may receive new treatments while others receive standard treatments or placebos.
2. Orthopedic Research: Distinct joints or limbs might be treated differently to compare the effects.
3. Pain Management: Various areas of the body might be subjected to different pain management techniques.
4. Cosmetic and Reconstructive Surgery: Surgeons may use split-face studies to evaluate the outcomes of different surgical techniques. By comparing pre- and post-operative images, they can assess changes in appearance and functionality.

Challenges:

However, split-body trials come with challenges, including:

1. Site-Specific Effects: The effects of the intervention may be influenced by the specific anatomical or physiological characteristics of the treatment site.
2. Blinding and Bias: It can be difficult for blind participants and researchers, as they might be aware of which side or area is receiving which treatment. This can introduce bias into the study.

Examples:

Examples of split-body trials include:

1. Topical Treatments: Testing a new cream or ointment where one half of a patient's body or a specific area of their skin receives experimental treatment while the other half receives a placebo or standard treatment.
2. Injections: Administering an experimental injection into one side of the face while the other side receives a standard or placebo injection.

In summary, split-body trials are a valuable design in clinical research, especially for comparing the effects of different treatments within the same individual. In survival analysis, comparing the time for an event to occur within the same patient in a split-body design clinical trial presents challenges for conventional survival analysis methodology. However, statistical tests such as the Prentice-Wilcoxon Test, considering within-subject correlation, provide a more accurate estimate than conventional methods such as Log-Rank or Gehan. The paper focuses on the advantages of a split-face design and the associated analysis of the time to first healing clinical endpoint.

Difference between Prentice-Wilcoxon, Gehan and Log Rank Test

The Prentice-Wilcoxon test and the Gehan test are both non-parametric statistical tests used for comparing time-to-event data, but they differ in their approaches to handling censored data and the types of hypotheses they emphasize. The Prentice-Wilcoxon test and the log-rank test are also used for comparing time-to-event distributions, but they differ in how they handle different parts of the survival curve and the emphasis they place on early versus late events.

In summary, these tests compare time-to-event curves with censored data, but they differ in how they weight the events, especially in the context of early vs. late differences and the hypotheses they emphasize. The Prentice-Wilcoxon test is more sensitive to early time-to-event differences.

Test	Purpose	Weighting	Censoring	Sensitivity	Applicability
Prentice-Wilcoxon	The Prentice-Wilcoxon test is a modification of the Wilcoxon rank-sum test, adapted for survival analysis, especially when censored data is present.	It gives more weight to observations with smaller event times, making it more sensitive to differences in early survival times.	It can handle censored data, but it places more emphasis on early events compared to later ones, which may make it less effective if late events are of more interest.	More sensitive to early differences between survival curves.	More suitable for paired time to event data when early differences in survival are the primary interest.
Gehan	The Gehan test (also known as the Gehan-Wilcoxon test or Gehan-Breslow test) is used for comparing survival curves and is often considered a generalization of the Wilcoxon rank-sum test for censored data.	It assigns weights to observations based on the number of subjects at risk at each event time, giving more weight to earlier events where more individuals are still at risk.	This test also accounts for censored data but is less sensitive to late-time differences compared to tests like the log-rank test, as it emphasizes early differences.	Generally, also emphasizes early events but balances weights differently compared to Prentice-Wilcoxon.	Appropriate for situations where one wants to account for all differences but with an emphasis on earlier events due to the at-risk weighting.
Log Rank Test	The log-rank test gives equal weight to all events across the survival curve, regardless of when they occur. This makes it sensitive to detecting differences across	Weights all events equally, making it more sensitive to detecting overall differences throughout the entire observation period. The log-rank test is better at	Like the Prentice-Wilcoxon test, the log-rank test can also handle censored data, but it provides an equal emphasis throughout the follow-up period.	Equally sensitive to differences over the whole follow-up time, making it suitable for comparing	Used when differences may arise at any point during follow-up, making it a more general-purpose test for comparing

	the entire follow-up period.	identifying consistent differences in survival across the entire duration of the study. It is often used in clinical trials and survival analysis when differences between groups are expected throughout the observation period.		overall survival curves.	survival distributions.

Data Requirements for Prentice-Wilcoxon Test:

- Survival Time (Ti): The time from the start of the study to the occurrence of the event or censoring.
- Censoring Indicator (δi): An indicator variable showing whether an event has occurred (δi=1) or the observation is censored (δi=0).
- Groups: The test compares two or more groups, each with their own time to event time and censoring information.

The application of the Prentice-Wilcoxon test is facilitated through the utilization of the **Lashley, et al. (2000)** provided method. This method necessitates the availability of valid time-to-event data for both variables from patients; any individuals lacking such data are excluded from the analysis.

Summary

- Rank Survival Times: Rank all survival times across both groups.
- Weight the Ranks: Apply weights based on the number of individuals still at risk, emphasizing early events.
- Calculate Weighted Rank Sums: Sum the weighted ranks separately for each group.
- Compute the Test Statistic: Calculate the test statistic U and standardize it to obtain Z.
- Interpret the Result: Use the standard normal distribution to determine if there is evidence of a difference between the treatment groups.

The Prentice-Wilcoxon test is powerful in detecting differences in survival between two treatment groups when those differences occur primarily in the early stages of the study. While the outcome of the test is a p-value, it does not offer insight into the direction or magnitude of the difference. Consequently, the employment of Kaplan-Meier median estimates and plots is indispensable for discerning the direction of any significant difference. The results can be reproduced in R software as well. The code for the analysis provided is in the appendix.

SPLIT-FACE DESIGN EXAMPLE

As an example, consider randomized, blinded, controlled, comparator split-face study of an IMP and a Vehicle to promote skin healing following ablative fractional resurfacing treatment over 21-days for 45 participants. On Day 1, the baseline visit, subjects were randomly assigned to IMP twice daily to one side of the face (right or left) and vehicle cream twice daily to the opposite side (left or right). The subjects were then instructed to apply the study products, each on one side of the face, as assigned, for a period of 14 days. Study visits took place on Days 2, 4, 7, 10, 14, 21 and 28.

The following endpoint was used for analysis:

- Investigator Assessment: Time to Healing - A comparison of investigator assessment of IMP and comparator side
- Time to healing was defined as the number of days from fractional resurfacing procedure to the first time that healing is achieved. If healing was not achieved, it was right censored by the number of days to the last visit assessment without healing. (Healing was defined as other than Poor on Overall skin healing and none of the other 6 items scores exceeding Mild in Investigator Assessment Questionnaire. Refer to appendix.)

(Example data in appendix)

```
proc sort data=one; by time;
run;
** compute s_sub_i and score for each point;
data one; set one nobs=nobs;
retain s_sub_i 1;
** get time to event for last obs;
lasttime=lag(time);
** compute s_sub_i;
if ((lasttime ^= time) and (censor ^= 1)) then
s_sub_i=s_sub_i*((nobs-_n_+1)/(nobs-_n_+2));
** compute score;
if (censor=1) then score=1-s_sub_i;
else score=1-(2*s_sub_i);
drop lasttime;
run;
proc sort data=one; by subject treatment;
run;
data two;
merge one(where=(trt1="A") rename=(score=score1 treatment=trt1) drop= time s_sub_i censor)
one(where=(trt2="B") rename=(score=score2 treatment=trt2) drop= time s_sub_i censor) end=eof;
by subject;
retain sumd 0 sumd2 0;
delta=score1-score2;
delta2=delta**2;
run;
proc sql noprint;
select sum(delta) into:sumd from two;
select sum(delta2) into:sumd2 from two;
quit;
%put &sumd &sumd2;
data _null_;
z=&sumd/(sqrt(&sumd2));
probz=2*(1-probnorm(abs(z)));
call symput('z',put(z,4.2));
call symput('probz',put(probz,6.4));
run;
%put &z &probz;

 ods listing close;
 ods OUTPUT CensoredSummary=cnsr  Quartiles=md ;

 proc lifetest data=one;
   strata treatment ;
   time time*censor(1);
 run; quit;
 ods output close;
 ods listing;
```

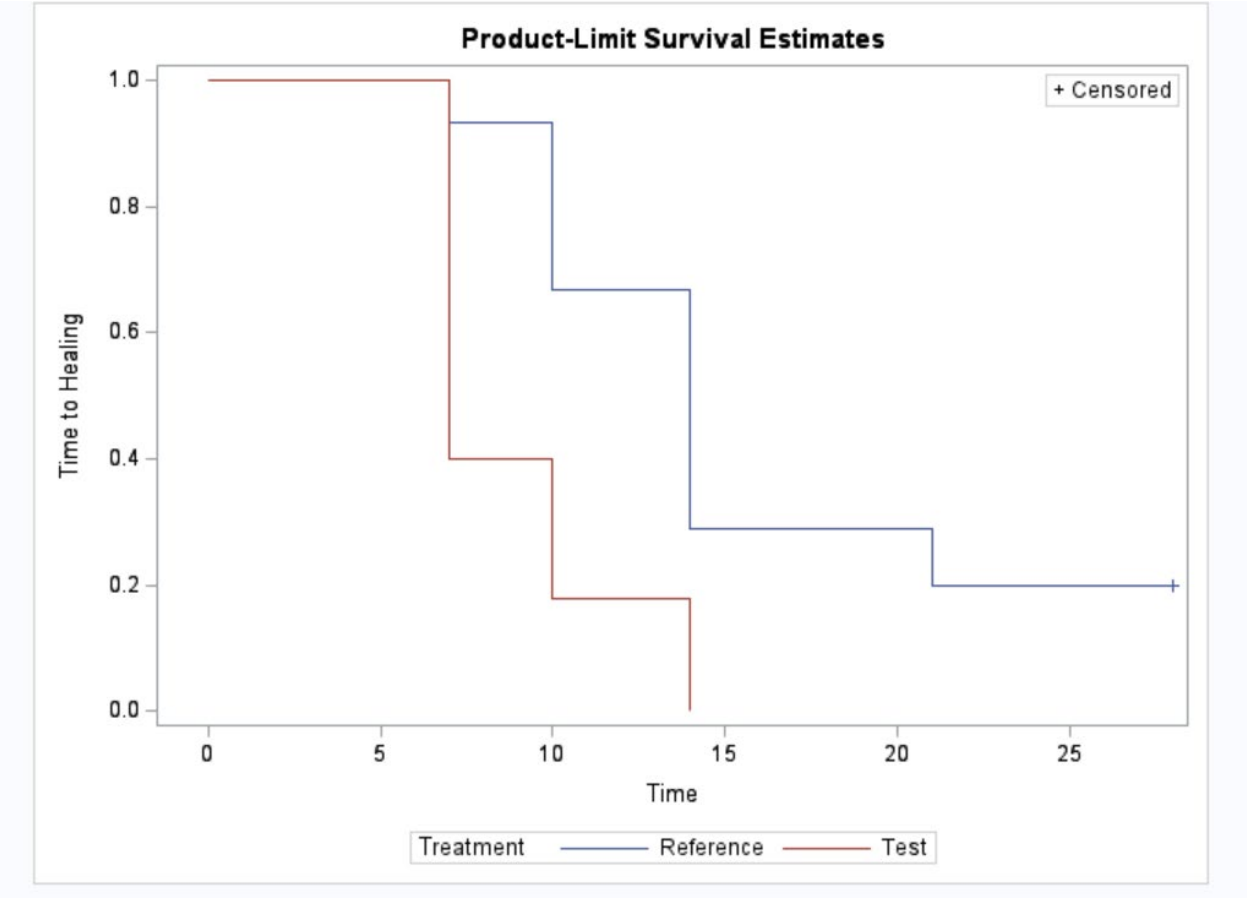
Statistical Analysis of Time to Healing: Investigator Assessment During Comparator Period (Day 1 - Day 21)

Statistics	Test	Reference	p-value[2]
N	45	45	0.0003
n (%) of Censored[1]	0	9 (20%)	
25th Quartile of Time to First Healing	7	21	
Median Time to First Healing	7	14	
75th Quartile of Time to First Healing	10	10	

Time to healing is calculated as number of days from fractional resurfacing procedure to the first time that healing is achieved. Subjects who did not achieve healing at the last assessment of the study period are right censored at the time of the last assessment.

[1] Statistics of censoring and quartiles of time to first healing derived using the Kaplan–Meier product-limit estimate.

[2] P-value for treatment comparison is derived from Prentice-Wilcoxon test for paired data with respect to time to healing.



CONCLUSION

From the above example using Prentice-Wilcoxon test and Kaplan-Meier plot, it is evident that treatment A is significantly better than treatment B in time it takes for overall healing in a split-face design study. This experimental design decreases the required number of subjects by utilizing each individual as their own control. However, the analysis of the resultant time to first event data presents greater complexity compared to parallel studies.

REFERENCES

1. A Paired Prentice-Wilcoxon Test for Censored Paired Data: Peter C. O'Brien and Thomas R. Fleming

2. Design and methodological issues of within-person (split-body) randomized controlled trials evaluating a topical treatment: A systematic review. Dermatology, 2023, 10.1159/000530149. hal-04106133

3. A Nonparametric Approach for the Statistical Analysis of Time to First Event in Censored Paired Data: Roderick Lashley, Dennis King, Pam Marks

4. Rank Tests for Matched Survival Data: Sin-Ho Jung 1999

5. The Statistical Analysis of Failure Time Data JOHN D. KALBF'LEISCH ROSS L. PRENTICE

6. Testing Equality of Survival Functions Based on Both Paired and Unpaired Censored Data- Michael J. Dallas and P.V Rao

7. https://www.quanticate.com/blog/bid/106718/prentice-wilcoxon-test-for-paired-time-to-event-data

CONTACT INFORMATION

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

Author Name: Shrutimita Pokhariyal
Company: Symbio
Address: Chandigarh, India
Work Phone: +91 7528806366
Email: shruti.pokhariyal@symbioresearch.com

Co-Author Name: Craig Wesselman
Company: Symbio
Address: Colorado, USA
Work Phone: +1 (970) 371-5767
Email: craig.wesselman@symbioresearch.com

APPENDIX:

Example Data

```
data one;
input Subject        Treatment$        Time Censor;
datalines;
1          A          7          0
1          B          10         0
2          A          7          0
2          B          14         0
3          A          7          0
3          B          10         0
4          A          7          0
4          B          14         0
5          A          14         0
5          B          28         1
6          A          10         0
6          B          14         0
7          A          10         0
7          B          28         1
8          A          14         0
8          B          28         1
9          A          7          0
9          B          10         0
10         A          7          0
10         B          14         0
11         A          10         0
11         B          14         0
12         A          14         0
12         B          21         0
13         A          7          0
13         B          10         0
14         A          14         0
14         B          21         0
15         A          10         0
15         B          14         0
16         A          10         0
16         B          28         1
17         A          7          0
17         B          10         0
18         A          14         0
18         B          28         1
19         A          10         0
19         B          14         0
20         A          7          0
20         B          28         1
21         A          7          0
21         B          14         0
22         A          7          0
22         B          10         0
23         A          7          0
23         B          14         0
24         A          14         0
24         B          21         0
25         A          7          0
25         B          10         0
26         A          7          0
26         B          10         0
27         A          10         0
27         B          14         0
28         A          7          0
28         B          10         0
29         A          7          0
29         B          10         0
30         A          10         0
30         B          28         1
31         A          7          0
31         B          7          0
32         A          7          0
32         B          10         0
33         A          14         0
```

```
33          B           28          1
34          A           7           0
34          B           14          0
35          A           7           0
35          B           14          0
36          A           7           0
36          B           14          0
37          A           7           0
37          B           21          0
38          A           7           0
38          B           14          0
39          A           14          0
39          B           28          1
40          A           10          0
40          B           14          0
41          A           7           0
41          B           7           0
42          A           7           0
42          B           14          0
43          A           7           0
43          B           10          0
44          A           10          0
44          B           14          0
45          A           7           0
45          B           7           0
run;
```

R CODE

```
install.packages("survival")

# Load necessary packages
library(dplyr)
library(survival)


# Create the data frame
one <- data.frame(
  Subject = c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10,
              11, 11, 12, 12, 13, 13, 14, 14, 15, 15, 16, 16, 17, 17, 18, 18,
              19, 19, 20, 20, 21, 21, 22, 22, 23, 23, 24, 24, 25, 25, 26, 26,
              27, 27, 28, 28, 29, 29, 30, 30, 31, 31, 32, 32, 33, 33, 34, 34,
              35, 35, 36, 36, 37, 37, 38, 38, 39, 39, 40, 40, 41, 41, 42, 42,
              43, 43, 44, 44, 45, 45),

  Treatment = c("A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A",
                "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B",
                "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A",
                "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B",
                "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A",
                "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B",
                "A", "B", "A", "B", "A", "B", "A", "B", "A", "B", "A", "B"),

  Time = c(7, 10, 7, 14, 7, 10, 7, 14, 14, 28, 10, 14, 10, 28,
           14, 28, 7, 10, 7, 14, 10, 14, 14, 21, 7, 10, 14, 21,
           10, 14, 10, 28, 7, 10, 14, 28, 10, 14, 7, 28, 7, 14,
           7, 10, 7, 14, 14, 21, 7, 10, 7, 10, 10, 14, 7, 10, 7,
           10, 10, 28, 7, 7, 7, 10, 14, 28, 7, 14, 7, 14, 7, 14,
           7, 21, 7, 14, 14, 28, 10, 14, 7, 7, 7, 14, 7, 10, 10,
           14, 7, 7),

  Censor = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1,
             0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
             0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
             0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
             0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
```

```
          0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
)


# Sort the data by Time
one <- one[order(one$Time), ]

result <- one %>%
  mutate(
    s_sub_i = 1,
    nobs = n(),  # Total number of observations
    lasttime = lag(Time)
  )
for (i in 2:nrow(result)) {
  if (!is.na(result$lasttime[i]) && result$lasttime[i] != result$Time[i] && result$Censor[i - 1] == 0) {
    result$s_sub_i[i] <- result$s_sub_i[i - 1] * ((result$nobs - i + 1) / (result$nobs - i + 2))
  } else {
    result$s_sub_i[i] <- result$s_sub_i[i - 1]  # Retain previous value
  }
}

# Compute score based on Censor
result <- result %>%
  mutate(
    score = ifelse(Censor == 1, 1 - s_sub_i,
                   ifelse(Censor == 0, 1 - (2 * s_sub_i), NA))

  )%>%

select (-lasttime)

# Step 2: Prepare for merging
one_A <- result %>% filter(Treatment == "A") %>%
  select(Subject, score1 = score)

one_B <- result %>% filter(Treatment == "B") %>%
  select(Subject, score2 = score)

# Merge the datasets
two <- one_A %>%
  left_join(one_B, by = "Subject") %>%
  mutate(
    delta = score1 - score2,
    delta2 = delta^2
  )

# Step 3: Calculate sums
sumd <- sum(two$delta, na.rm = TRUE)
sumd2 <- sum(two$delta2, na.rm = TRUE)

# Print the results for sumd and sumd2
cat("Sum of delta:", sumd, "\n")
cat("Sum of delta squared:", sumd2, "\n")

# Step 4: Calculate z and probz
z <- sumd / sqrt(sumd2)
probz <- 2 * (1 - pnorm(abs(z)))
```

Investigator Assessment: Time to Healing Questionnaire

Parameter	Investigator Assessment/ Time to Healing	Rating Scale				
	Sign/Symptom	**0**	**1**	**2**	**3**	**4**
i.	Erythema (redness)	none/absent	mild	moderate	severe	very severe
ii.	Edema (swelling)	none/absent	mild	moderate	severe	very severe
iii.	Dyspigmentation (darkening or lightening of the skin)	none/absent	mild	moderate	severe	very severe
iv.	Desquamation (peeling)	none/absent	mild	moderate	severe	very severe
v.	Crusting or Scabbing	none/absent	mild	moderate	severe	very severe
vi.	Acne or acne-like eruptions in area of treatment	none/absent	mild	moderate	severe	very severe
vii.	Overall skin healing	excellent	very good	good	fair	poor

- Lower the score better the result of the product in the skin healing.
- Healing was defined as other than Poor on Overall skin healing and none of the other 6 items scores exceeding Mild
- For a specific pair of subscales, a score is calculated for each patient, taking into consideration both subscales and the censored nature of the data.
- The visit at which the score first indicates healing is defined as the time (in days) to first-time healing.
- If treatment does not provide improvement before the end of the treatment period or discontinuation day, the number of elapsed days is recorded, and the value is considered censored.