J. Soc. Cosmet. Scientists Korea
Vol. 39, No. 2, June 2013, 105-115

기능성 화장품의 인체시험 설계 및 통계적용 방법에 대한 고찰

서 영 경† · 고 재 숙 · 이 원 철*

㈜더마프로 피부과학연구소, *가톨릭대학교 의과대학 예방의학교실
(2012년 10월 30일 접수, 2013년 1월 17일 수정, 2013년 2월 26일 채택)

Investigation of the Study Plan and Statistical Method of Functional Cosmetics on Human Skin

Young Kyoung Seo†, Jae Sook Koh, and Won Chul Lee *

Dermapro Skin Research Center, Dermapro Co., LTD.,
4F Jiho B/D, 919-1, Bangbae-Dong, Seocho-Gu, Seoul 137-843, Korea
*Department of Preventive Medicine, College of Medicine, Catholic University of Korea
(Received October 10, 2012; Revised January 17, 2013; Accepted February 26, 2013)

요 약: 국내의 주름개선 혹은 미백 효능을 평가하기 위한 인체시험 방법은 식약청 가이드라인에 근거하여 시행되어왔으며, 인체시험에서 육안평가 및 기기평가 결과에 대해 시험군과 대조군 간의 효과를 비교하기 위해서 unpaired *t*−test를 주로 이용하였고, 시술 전후의 효과를 비교하기 위해서 paired *t*−test를 이용하였다. 설문평가 결과에 대해서는 빈도분석을 이용한 기술통계법이 이용되고 있다. 미국 및 유럽의 임상 평가기관에서도 이와 유사한 시험법 및 통계분석 방법을 이용하고 있다. 그러나 동일 개체에 대하여 처치를 반복 적용하여 얻은 자료는 서로 관련성이 높아 이를 감안한 분석법을 적용해야 한다. 따라서 본 연구에서는 화장품 분야에서는 처음으로 기능성 화장품 중 주름 개선 및 미백 효능 시험의 육안평가 및 기기평가 자료에 대해 repeated measures ANCOVA (RM ANCOVA)와 repeated measures ANOVA (RM ANOVA)를 적용하여 통계 방법의 타당성 여부를 검증함으로서 현재의 인체시험 방법에 적합한 새로운 통계분석 방법을 제시하였다.

Abstract: In Korea, the human skin tests to evaluate the anti-wrinkles and whitening effect have been accomplished in accordance with the KFDA guideline. Regarding the data of the visual assessment and machinery evaluation of the results for the human skin test, unpaired *t*-test have been used in order to compare between the test and the control groups and paired *t*-test for the comparison of effects for before and after. Descriptive statistics such as frequency analyses was used for the questionnaire evaluation data. In many cases of the European and American clinical test centers, the methodology and the statistical analysis were similar to ours. But, the documentation obtained by repeated application from identical individual has high relation. For this reason, it is desirable to apply RM ANCOVA and RM ANOVA to a visual assessment and machinery evaluation. We suggested that RM ANCOVA and RM ANOVA is the new approach to statistical analysis of human test data of functional cosmetics.

Keywords: *clinical statistics, functional cosmetics, repeated measures ANOVA, repeated measures ANCOVA*

† 주 저자 (e-mail: dermapro@dermapro.co.kr)

1. 서 론

제정된 화장품법에 따르면 “화장품”이라 함은 “인체를 청결, 미화하여 매력을 더하고, 용모를 밝게 변화시키거나 피부 모발의 건강을 유지 또는 증진하기 위하여 인체에 사용되는 물품으로서 인체에 대한 작용이 경미한 것을 말한다. 다만, 이러한 사용목적 이외에 사람 또는 동물의 질병의 진단, 치료, 경감, 처치 또는 예방의 목적과 사람 또는 동물의 구조기능에 약리학적 영향을 주기 위한 목적을 겸하여 사용되는 물품은 제외로 한다.”로 정의하고 있다. 이에 관하여 화장품법 시행규칙 제2조에서 기능성 화장품의 범위를 지정하고 있는데 그 범위는 다음과 같다. 1) 피부의 멜라닌 색소가 침착하는 것을 방지하여 기미, 주근깨 등의 생성을 억제함으로써 피부의 미백에 도움을 주는 기능을 가진 화장품, 2) 피부에 침착된 멜라닌색소의 색을 엷게 하여 피부의 미백에 도움을 주는 기능을 가진 화장품, 3) 피부에 탄력을 주어 피부의 주름을 완화 또는 개선하는 기능을 가진 화장품, 4) 강한 햇볕을 방지하여 피부를 곱게 태워주는 기능을 가진 화장품, 5) 자외선을 차단 또는 산란시켜 자외선으로부터 피부를 보호하는 기능을 가진 화장품으로 기능성 화장품의 범위를 규정하고 있다[1,2].

현재 국내에서 화장품법이 시행된 이래 장업사 연구소, 의과대학 피부과 및 전문임상 평가기관 등을 중심으로 인체시험 가이드라인에 준하여 기능성화장품 효능 검증 연구가 진행되고 있다. 효능 검증을 위해 얻어진 데이터의 통계 분석방법은 임상 시험의 해석과 논문 분석에 필수적이며[3,4] 연구의 디자인과 직결되며 논문의 완성도와 근거의 비중을 결정하는 중요한 요소이다[5]. 현재 인체적용시험에서 육안평가 및 기기평가 결과에 대해 시험군과 대조군 간의 효과를 비교하기 위해서 unpaired *t*-test를 주로 이용하였고, 시술 전후의 효과를 비교하기 위해서 paired *t*-test를 이용하였다. 설문평가 결과에 대해서는 빈도분석을 이용한 기술통계법이 이용되고 있다.

그러나 화장품의 사용처럼 동일 개체에 대하여 처치를 반복 적용하여 얻은 자료는 서로 관련성이 높아 이를 감안한 분석법을 적용해야 한다. 따라서 본 연구에서는 화장품 분야에서는 처음으로 기능성 화장품 중 주름 개선 및 미백 효능 시험의 육안평가 및 기기평가 자료에 대해 반복측정 분석(Repeated measurement ANOVA)을 적용하여 통계 방법의 타당성 여부를 검증함으로서 현재의 인체적용시험 방법에 적합한 통계분석 방법을 제시하고자 하였다.

2. 대상 및 방법

2.1. 기능성 화장품의 인체시험법 및 통계분석법 조사

기능성 화장품 중 주름개선 및 미백 효능 평가와 관련한 국내외 화장품 회사, 원료 회사, 임상기관의 인체 시험 보고서를 수집하여 시험 설계법 및 통계 방법에 관해 조사하였다. 또한, 화장품 및 피부과학 분야를 검색하여 화장품의 주름개선 및 미백 효능 평가와 관련한 자료를 수집하였다.

2.2. 제3상 임상시험의 설계법 조사

기능성 화장품의 인체시험의 설계 및 통계분석의 타당성을 입증하기 위해서는 질병치료의 효과를 평가하고자 했던 의학 분야에서의 임상시험을 참조할 필요가 있다. 의학분야의 임상시험은 임상 약리학 및 독성을 평가하고자 한 제1상 시험, 치료 효과를 위한 첫 단계로서의 제2상 시험, 적극적이고 광범위한 치료효과 평가단계로서의 제3상 시험, 시판 후 조사 방법인 제4상 시험으로 분류될 수 있다. 이 중 광의의 ‘임상시험’을 대변하는 제3상 시험의 설계에 관해 조사하여 기능성 화장품 관련 인체시험의 설계에 참고하고자 하였다.

2.3. 기존 기능성 화장품 평가의 통계적 타당성 고찰

2.3.1. 반복측정 자료에 대한 Repeated Mesaures ANCOVA의 적용에 대한 검토

현재 기능성 화장품의 인체 시험에서 각 시점에서의 두 군간 차이를 비교하고자 할 때 independent *t*-test를 주로 적용하고 있다. 구체적인 방법으로는 관찰시점에서의 관측치에서 기준시점에서의 관측치를 뺀 차이값(difference value)을 변수로 하여 시험군과 대조군의 두 군에서 차이가 있는지를 검정한다. 또는 기준시점에 대한 고려 없이 관찰시점에서의 관측치를 변수로 하여 시험군과 대조군의 두 군에서 차이가 있는지를 검정한다. 이에 본 연구에서는 이의 타당성을 논하

Table 1. Report of the Anti-wrinkle Effect [6]

Report summary	
Test center	IEC, France
Report date	2002. 3. 21
Subjects	Twenty-eight female subjects
Test period	12 weeks (W0, W6, W12)
Instruments	Silflo, Skin Image Analyzer (QUANTRIDES^{TM}software)
Test site	Crow's foot
Visual assesment	9-scale scoring
Product	Test product / Placebo
Statistical analysis	Skin relief parameters Levene test and Kolmogorov-Smirnov test ($p < 0.05$) : Test for homogeneity of variances and the normality of distributions → Student *t*-test in case of homogeneity of variances and of normality of distributions → Non parametric Wilcoxon test (two-tail, $p < 0.05$) in adverse case Clinical evaluation and self-assessment Paired Wilcoxon test (two-tail, $p < 0.05$)

Table 2. Clinical Efficacy Evaluation of s Skin Care Product, Containing Indole-3-acetic Acid, Designed to Reduce the Appearance of Surface Fine Lines and Wrinkles [7]

Report summary	
Test center	RCTS Inc, USA
Report date	1999. 1. 14.
Subjects	Thiry-two subjects
Test period	6 weeks (W1, W2, W3, W4, W6)
Instruments	Image Analyser : Rz (roughness) : Number of wrinkles : Breadth of lines detected (normal direction, parallel direction) : Area of shadows (Normaldirection, paralleldirection)
Test site	Crow's foot
Visual assesment	None
Statistical analysis	paired *t*-test, $p \leq 0.05$

고 이를 보정한 Repeated Mesaures ANCOVA를 제시하고 실제 주름 측정 자료에 대해 적용하였다.

2.3.2 반복측정 자료에 대한 Repeated Mesaures ANOVA에 대한 검토

현재 기능성화장품의 인체시험에서 한 군에서의 사용 전후의 차이를 비교하고자 할 때 paired *t*-test를 주로 적용하고 있다. 즉 동일한 대상에 대하여 시간이 진행에 따라 반복적으로 관찰한 측정치에 대하여 기준시점에 대한 변화를 검정하고자 하는 경우 기준시점 값과 관찰 시점 1, 2, 3 시점의 각각 값에 대하여 paired *t*-test를 수행하는 방법을 적용하고 있다. 본 연

Table 3. Evaluation, *in vivo* on Human Subjects, of the Anti-wrinkle, Moisturizing and Firming Effects of Aproduct [8]

Report summary	
Test center	Dermscan, France
Report date	2000. 6. 21.
Subjects	Thiry subjects
Test period	42 days (D0, D42)
Instruments	Skin Image Analyzer (QUANTIRIDES^{TM}-MONADERMsoftware)
Test site	Crow's foot
Statistical analysis	paired *t*-test, $p \leq 0.05$
	Excel 7.0, ver 95

Table 4. Study Report-marketing, Skin Care Home-in-use [9]

Report summary	
Test center	proDERM, Germany
Subjects	Thirteen sbujects
Test period	12 weeks (D1, D29, D57, D85)
Instruments	Skin color: Chromameter Skin wrinkle: PRIMOS (Skin roughness parameters: Ra, Rz)
Test site	Skin color: back of hands Skin wrinkle: periorbital regions
Statistical analysis	Skin color - ITA values (difference to baseline / difference to the basic formulation, $p < 0.05$) Skin wrinkles - Ra, Rz (difference to baseline / difference to the basic formulation, $p < 0.05$, $p < 0.1$)

Table 5. Study of Anti-wrinkle Effect of Argireline [10]

Report summary	
Test center	Advancell, Barcellona, Spain
Report date	2001. 10.
Subjects	Ten subjects
Test period	30 days (D0, D15, D30)
Instruments	PRIMOS (Skin roughness parameters: Ra, Rz)
Test site	Crow's foot
Statistical analysis	paired *t*-test, $p < 0.05$

구에서는 이의 타당성을 논하고 이를 보정한 repeated measures ANOVA의 적용을 제시하고 실제의 측정치에 적용하고자 하였다.

3. 결 과

3.1. 기능성 화장품의 인체시험법 및 통계분석법 조사

국내의 주름개선 및 미백 효능을 평가하기 위한 인체시험 방법은 식약청에서 제시한 가이드라인에 근거하여 시행되었다. 인체시험의 육안평가 및 기기평가 자료에 대해서는 시험군과 대조군의 효과를 비교하기 위해 unpaired *t*-test를 주로 이용하였고, 전후의 효과를 비교하기 위해서는 paired *t*-test를 이용하였다. 설문평가 자료에 대해서는 빈도 분석을 이용한 기술통계 기술이 이용되었다.

미국 및 유럽의 임상평가기관의 보고서를 분석한 결과 인종의 특성상 미백의 경우보다 대다수가 주름

Table 6. Skin Roughness Parameters of at 0 Weeks and 12 Weeks

Group	Test group		Control group	
No	W0	W12	W0	W12
1	0.50	0.45	0.49	0.48
2	0.45	0.39	0.42	0.40
·	·	·	·	·
·	·	·	·	·
22	0.41	0.38	0.42	0.49
23	0.47	0.28	0.41	0.40
Mean	0.46	0.41	0.47	0.44

Table 7. Change of Skin Roughness Parameters at 0 Weeks and 12 Weeks

Group	Test group	Control group
No	W12-w0	W12-w0
1	-0.05	-0.01
2	-0.06	-0.01
·	·	·
·	·	·
22	-0.02	0.07
23	-0.19	-0.02
mean	-0.05	-0.022

에 관한 사례로서 국내에서 시행하고 있는 방법과 큰 차이가 없었으며 통계적용 방법도 유사하였다. 다만, 피험자 수 및 정규분포성을 만족하지 않는 경우 비모수적 검증 방법을 적용하는 사례가 있었다. 미백관련 인체시험의 경우도 주름의 경우와 마찬가지로 시험의 원리 및 프로토콜 및 통계방법이 국내의 경우와 유사하였다(Table 1-5).

3.2. 기존 기능성 화장품 평가의 통계적 타당성 고찰 및 새로운 통계분석법 적용

3.2.1. 통계기법인 ANCOVA의 적용에 대한 검토

기능성 화장품의 인체시험 자료 중 기기평가값, 육안평가 값의 분석들을 살펴보면, 대부분 두 관찰시점의 값에 대한 차이(⊿)값을 이용하여 분석하는 경우가 대부분이다. 그러나 원 자료를 사용치 않고, 차이인 ⊿값만을 사용하여 두 군의 차이를 검정(unpaired *t*-test)하는 것은 chance의 문제를 일으킬 수 있다. 즉 자료가 20과 10인 경우의 차이도 10이고, 190과 180인 경우의 차이도 10인데, 차이인 ⊿값만을 사용 시 두 가지의 경우가 모두 동일하게 분석된다. 그러나 실제 원 자료를 분석하는 repeated measures ANCOVA를 사용하면, 그와 같은 문제를 해결할 수 있다.

Tables 6, 7은 각각 시험군과 대조군의 기기측정값과 두 시점 간의 차이값을 보여주는 예이다. 이 두군의 평균값을 unpaired *t*-test를 적용하면, Table 8과 같이

Table 8. The Result of Statistical Analysis by Unpaired *t*-test

The TTEST Procedure

Statistics

			Lower CL		Upper CL		Lower CL		Upper CL
Variable	gr	N	Mean	Mean	Mean	Std Dev	Std Dev	Std Dev	Std Err
W12_0	대조군	23	-0.046	-0.022	0.002	0.0433	0.056	0.0793	0.0117
W12_0	시험군	23	-0.075	-0.05	-0.026	0.0443	0.0573	0.081	0.0119
W12_0	Diff (1-2)		-0.005	0.0283	0.0619	0.0469	0.0566	0.0715	0.0167

t-Tests

Variable	Method	Variances	DF	t Value	Pr > | t |
W12_0	Pooled	Equal	44	**1.69** ①	**0.0977** ②
W12_0	Satterthwaite	Unequal	44	1.69	0.0977

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
W12_0	Folded F	22	22	1.05	0.9181

Table 9. The Result of Statistical Analysis by Repeated Measures ANCOVA

The GLM procedure

Dependent variable: W12 W12

Source	DF	Sum of Squares	Mean square	F Value	Pr > F
Model	2	0.04175376	0.02087688	9.14	0.0005
Error	43	0.09818078	0.00228327		
Corrected total	45	0.13993454			

R-Square	Coeff Var	Root MSE	W12 Mean
0.298381	11.17083	0.047784	0.427754

Source	DF	Type I SS	Mean square	F Value	Pr > F
GROUP	1	0.01116546	0.01116546	4.89	0.0324
W0	1	0.03058830	0.03058830	13.40	0.0007
Source	DF	Type III SS	Mean square	F Value	Pr > F
GROUP	1	0.01022808	0.01022808	**4.48 ①**	**0.0401 ②**
W0	1	0.03058830	0.03058830	13.40	0.0007

Parameter	Estimate	Standard Error	t Value	Pr > | t |
Intercept	0.2004209824 B	0.05870534	3.41	0.0014
GROUP 1	0.0298326475 B	0.01409528	2.12	0.0401
GROUP 2	0.0000000000 B	.	.	.
W0	0.4577365983	0.12505958	3.66	0.0007

Note : The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM procedure

Least squares means

GROUP	W12 LSMEAN	Standard Error	H0:LSMEAN = 0 Pr > | t |	H0:LSMean1 = LSMean2 Pr > | t |
1	0.44266995	0.00996522	< .0001	0.0401
2	0.41283730	0.00996522	< .0001	

t = 1.69(①), *p*-value는 0.0977(②)로써, 두 군은 통계적으로 차이가 없다고 판정할 수 있다.

이 자료를 repeated measures ANCOVA 분석한 결과 Table 9에 나타난 바와 같이 F = 4.48(①), *p*-value = 0.0401(②)로써, 두 군은 유의한 차이가 있다고 해석이 가능하다.

따라서, 동일한 자료에 대하여 두 시점 간 차이(w12-w0)를 가지고 분석한 unpaired *t*-test를 적용한 결과에 대하여는 두 군의 차이가 없는 것으로 나타났으나, 원 자료 그대로를 이용하여 ANCOVA로 분석한 결과에서는 두 군의 차이가 있는 것으로 나타나 통계적 검정력이 크게 향상되었음을 보여주는 예이다.

3.2.2. 반복측정 자료에 대한 Repeated Measures ANOVA에 대한 검토

현재 연구에서 사용하고 있는 자료는, 시험이 모두 동일한 사람들을 대상으로 시점을 달리하여 반응값을 얻은 자료로서 이를 반복측정 자료라 하며 R.A. Fisher

Table 10. Skin Roughness Paramenters

No	Control group				Test group			
	W0	W4	W8	W12	W0	W4	W8	W12
1	0.49	0.39	0.46	0.48	0.50	0.37	0.42	0.45
2	0.42	0.41	0.34	0.40	0.45	0.39	0.39	0.39
3	0.44	0.46	0.37	0.49	0.42	0.41	0.31	0.40
4	0.48	0.45	0.51	0.42	0.43	0.42	0.34	0.41
5	0.49	0.48	0.43	0.49	0.56	0.46	0.46	0.45
6	0.49	0.45	0.41	0.40	0.46	0.47	0.36	0.40
7	0.54	0.50	0.47	0.50	0.42	0.37	0.38	0.37
8	0.49	0.37	0.42	0.51	0.50	0.39	0.49	0.58
9	0.44	0.39	0.38	0.35	0.42	0.37	0.39	0.38
10	0.37	0.48	0.32	0.40	0.40	0.49	0.35	0.41
11	0.45	0.49	0.48	0.52	0.46	0.35	0.43	0.45
12	0.42	0.45	0.51	0.44	0.63	0.48	0.49	0.48
13	0.48	0.42	0.39	0.40	0.43	0.38	0.37	0.32
14	0.48	0.50	0.50	0.48	0.42	0.39	0.36	0.43
15	0.56	0.34	0.35	0.41	0.49	0.41	0.42	0.39
16	0.43	0.48	0.55	0.44	0.45	0.53	0.46	0.43
17	0.43	0.43	0.44	0.42	0.40	0.41	0.41	0.36
18	0.48	0.49	0.44	0.44	0.52	0.42	0.39	0.45
19	0.47	0.45	0.50	0.43	0.42	0.39	0.36	0.39
20	0.44	0.41	0.42	0.41	0.41	0.45	0.39	0.37
21	0.59	0.55	0.56	0.47	0.61	0.48	0.56	0.52
22	0.42	0.41	0.41	0.49	0.41	0.39	0.41	0.38
23	0.41	0.38	0.35	0.40	0.47	0.42	0.33	0.28

가 제안한 분산분석법(ANOVA, analysis of variance)을 적용하여 시점 간, 처리 간의 차이를 검정하였다. 주름 평가의 경우 측정시점별로 비교할 때 0주와 4주, 0주와 8주, 0주와 12주를 각각 검정하였다. 각 검정에 대한 유의수준으로 α = 0.05를 택하면, 각 검정에서 평균의 차이가 없다는 가설을 기각하지 않을 확률은 0.95이다. 확률의 곱에 따라, 각 검정들이 서로 독립적일 때, 3개의 검정(0주와 4주, 0주와 8주, 0주와 12주)을 통하여 모든 평균들이 서로 차이가 있다는 가설을 기각하지 않을 확률은 0.95^3 = 0.8573이다. 그러면 여기서 차이가 없다는 가설 중 적어도 하나를 기각하는 확률은 1 - 0.8573 = 0.1427이다. 우리가 이 3가지 경우에 있어서 귀무가설이 옳다는 것을 알고 있다면, 귀무가설을 기각하는 것은 제 I 종 오류를 저지르게 됨을 의미한다[11]. 그러므로 결국 14.27%나 되는 제 I 종 오류를 저지르게 되는 셈이다. 게다가 서로 독립이 아닌 *t*-검정이므로 더욱 복잡하게 된다. 따라서, 전체 검정의 유의수준이 α (= 0.05)가 되도록 조정해 주는 검정법, 즉 다중비교법을 적용해야 한다. 반복 측정자료의 많은 다중비교법 중 본 연구에서는 Bonferroni 방법을 제시하였다[12]. 이는 Bonferroni 부등식에서 비롯된 것으로, 서로 독립이 아닌 검정의 실시 횟수가 m번일 때, α' = α/m을 각 검정의 유의수준으로 정하여 검정하는 방법이다. 대부분의 통계 패키지에서는 각 검정에 대한 *p*값이 제공되는데 이를 그대로 사용치 않고, p' = pXm을 구하여 사용한다. 모든 집단에서의 효과가 다 같지는 않다는 결론을 얻은 경우 집단 간의 차이도 같은 방법을 적용하여 검정한다[13,14].

Table 11. Statistical Analysis of Skin Roughness Parameters (Mean ± SD)

Control group				Test group			
W0	W4	W8	W12	W0	W4	W8	W12
0.47± 0.05	0.44± 0.05	0.44± 0.07	0.44± 0.04	0.46± 0.06	0.42± 0.05	0.40± 0.06	0.41± 0.06

(1) RM ANOVA 적용 사례–주름기기 평가자료

Table 10은 주름평가 시 대조군과 시험군에 대하여 0주에서 12주까지 4회 반복측정한 값과 이 값의 평균과 표준편차를 이용한 시점별 평균 변화량을 보여주는 예이다(Table 11).

SPSS 10.0을 사용하여, 실제 자료에 RM ANOVA를 적용한 결과는 다음과 같다(Table 12).

Table 12의 ① 'WEEK'는 0, 4, 8, 12주의 평균값들이 모두 동일한지 아니면 어느 한 주에라도 차이가 있는지를 평가하기 위함이다. 이 결과 ④에서 유의확률 0.001을 나타내어 모두 동일하지 않음을 표시하고 있다. 즉, w0 = w4 = w8 = w12는 기각이 되고, 어느 한 주에서라도(또는 모든 주에서) 평균값이 다르다는 것을 의미한다. 각 평가 시점별 변화량은 단계 분석결과인 Table 12에 나타난다.

Table 12. Statistical Analysis of Skin Roughness Parameters by RM ANOVA

Effect		Value	F	Hypothesis df	Error df	Sig.
① Week	Pillai's Trace	0.542	7.880[a]	3.000	20.000	0.001 ④
	Wilks' Lambda	0.458	7.880[a]	3.000	20.000	0.001
	Hotelling's Trace	1.182	7.880[a]	3.000	20.000	0.001
	Roy's Largest root	1.182	7.880[a]	3.000	20.000	0.001
② Group	Pillai's Trace	0.228	6.494[a]	1.000	22.000	0.018 ⑤
	Wilks' Lambda	0.772	6.494[a]	1.000	22.000	0.018
	Hotelling's Trace	0.295	6.494[a]	1.000	22.000	0.018
	Roy's Largest root	0.295	6.494[a]	1.000	22.000	0.018
③ Week* Group	Pillai's Trace	0.199	1.660[a]	3.000	20.000	0.207 ⑥
	Wilks' Lambda	0.801	1.660[a]	3.000	20.000	0.207
	Hotelling's Trace	0.249	1.660[a]	3.000	20.000	0.207
	Roy's Largest root	0.249	1.660[a]	3.000	20.000	0.207

a: Exact statistic
b: Design : Intercept
W0-Group within subjects Design: Week + Group + Week*Group

② 'GROUP'은 대조군과 시험군이 동일성을 검정하고 있다. 이 결과는 ⑤의 0.018로 0.05보다 작으므로, 두 군은 유의한 차이가 있으므로 주름개선이 되었음을 보여주는 예이다. 즉, 대조군과 비교 시 시험군이 통계적으로 유의하게 주름개선이 되었음을 알 수 있다.

③ 'WEEK* GROUP'은 두 군과 평가 시점 사이에 교호작용이 있는지를 검정하고자 함이다. 여기서 교호작용이라 함은 각 평가 시점에 따라 두 군의 차이가 동일한지를 검정하는 것으로, 그 결과 ⑥의 0.207은 p = 0.05보다 커 교호작용은 인정되지 아니하며, 각 시점마다 두 군의 차이는 동일함을 나타내고 있다.

기기 측정이 시점별로 0, 4, 8, 12주 반복되었으므로, 기준시점인 0주와 다른 시점들 간에 차이가 있는지에 대한 검정 결과가 WEEK의 수준별 유의확률(p-value)이다.

여기에 Bonferroni method를 적용하여 3을 모든 유의확률(p-value)에 곱하면 다음과 같은 결과를 볼 수 있다.

0주와 4주(수준2 대 수준1) = 0.009(①) × 3 = 0.027 ← ④

0주와 8주(수준3 대 수준1) = 0.000(②) × 3 = 0.000 ← ⑤

0주와 12주(수준4 대 수준1) = 0.001(③) × 3 = 0.003 ← ⑥

분석 결과 기준시점과 비교 시 4, 8, 12주 후 유의한 주름개선이 되었음을 보여주고 있다. GROUP별 차이는 WEEK와는 달리, 수준이 시험군과 대조군 둘 뿐이므로 Table 12-⑤와 같이 p = 0.018로 나타났다.

(2) Unpaired t-test와 RM ANOVA의 결과 비교

Table 14는 0주와 4주차의 기기측정값 차이를 시험군과 대조군으로 unpaired t-test한 결과이다. 0주와 4주차의 유의확률은 0.295(①), 0주와 8주차의 유의확률은 0.085(②), 0주와 12주차의 유의확률은 0.098(③)이다. 세 군 모두 α = 0.05보다 크기때문에, 유의한 차이가 없으므로 시험군의 주름개선효과가 없음을 나타낸다.

그러나, RM ANOVA 분석 시 시험군이 대조군에 비해 유의하게 나타났다(Table 12). 따라서 적절한 통계분석방법의 선택은 결과를 뒤집을 수 있으며 통계

Table 13. Changes of Skin Roughness Parameters on Control Group Before and After Treatment Measure: Measure_1

Source	Week	Group	Type III Sum of Squares	df	Mean square	F	Sig.
Week	Level 2 vs. level 1		2.467E-02	1	2.467E-02	8.323	0.009 ①
	Level 3 vs. level 1		4.628 E-02	1	4.628 E-02	24.460	0.000 ②
	Level 4 vs. level 1		3.031 E-02	1	3.031 E-02	13.949	0.001 ③
Error (Week)	Level 2 vs. level 1		6.522 E-02	22	2.965E-03		
	Level 3 vs. level 1		4.162 E-02	22	1.892E-03		
	Level 4 vs. level 1		4.781 E-02	22	2.173E-03		
Group		Level 2 vs. level 1	1.161 E-02	1	1.161E-02	6.494	0.018
Error (Group)		Level 2 vs. level 1	3.932 E-02	22	1.787E-03		
Week* Group	Level 2 vs. level 1	Level 2 vs. level 1	9.469 E-02	1	9.469E-03	1.896	0.182
	Level 3 vs. level 1	Level 2 vs. level 1	2.030 E-02	1	2.030E-02	3.672	0.068
	Level 4 vs. level 1	Level 2 vs. level 1	1.837 E-02	1	1.837E-02	4.442	0.047
Error (Week* Group)	Level 2 vs. level 1	Level 2 vs. level 1	0.110	22	4.995E-03		
	Level 3 vs. level 1	Level 2 vs. level 1	0.122	22	5.529E-03		
	Level 4 vs. level 1	Level 2 vs. level 1	9.097 E-02	22	4.135E-03		

Table 14. Statistical Analysis of Skin Roughness Parameters by Unpaired *t*-test

Source	F	Sig.	t	df	Sig. (2-tailed)	Mean difference	
W4_W0	0.268	0.607	1.060	44	0.295	2.029E-02	←①
W8_W0	1.796	0.187	1.761	44	0.085	2.971E-02	←②
W12_W0	0.057	0.813	1.692	44	0.098	2.826E-02	←③

Table 15. Statistical Analysis of Skin Roughness Parameters by Paired *t*-test

Source	Paired difference					t	df	Sig.
	Mean	Std. Deviation	Std. Error mean	95% Confidence interval of the difference				
				Lower	Upper			
0W(control)-4W(control)	2.261E-02	6.413E-02	1.337E-02	-5.12E-03	5.034E-02	1.691	8.323	0.105 ①
0W(control)-8W(control)	3.000D-02	6.724E-02	1.402E-02	9.233E-04	5.908E-04	2.140	24.460	0.044 ②
0W(control)-12W(control)	2.217D-02	6.600E-02	1.168E-02	-2.04E-03	4.639E-02	1.899	13.949	0.071 ③

적인 검정력을 높일 수 있다.

(3) Paired *t*-test와 RM ANOVA의 결과 비교

Table 13은 대조군의 각 시점 변화값을 검정한 것이다. 그 유의확률을 보면, 0.105(①), 0.044(②), 0.071(③)로 0주와 비교 시 4, 12주 간의 차이는 유의하지 않게 나타났다. 그러나 RM ANOVA에서는 0주와 비교 시 4, 8, 12주차에 통계적으로 유의하게 나타났다(Table

13 ①, ②, ③). Paired *t*-test분석에서는 차이가 없게 나온 결과가 동일한 자료를 RM ANOVA로 분석하였을 경우 유의한 차이 있게 나타남을 이 예에서도 마찬가지로 보여주고 있다(Tables 14, 15).

4. 결 론

임상시험의 결과가 통계적 검정력을 가지려면 연구목적에 맞게 연구대상자를 모집하여 이들을 연구종료 시점까지 컨트롤 해야 하고[15] 임상시험 종료 후 적절한 분석법을 실시하여야 한다. 적절하지 않은 결과분석법 적용시 효과가 없는데도 불구하고 효과가 있다고 판정되는 제I종 오류와 효과가 있는데도 불구하고 효과가 없다고 판정되는 제II종 오류를 범할 수 있다[16,17]. 예를 들어 결과 분석 시 모수적 방법을 시행하기엔 연구대상자가 너무 적거나 정규분포성을 만족하지 않아 모수적 분석을 할 수 없는 경우[5] 비모수적 검증 방법을 적용하는 경우가 있다. 특히, 인체시험 결과에서 두 군의 차이를 인정할 만한 차이를 보일 지라도 연구대상자 수가 너무 작을 경우 통계 검정 시 제II종 오차(Type II error)를 초래할 수 있으므로 연구 설계 단계부터 적절한 최적의 연구대상자 수 추정 또한 매우 중요한 것으로 사료된다[18].

따라서 본 연구는 인체시험 결과 자료에 대한 국내외의 기능성 화장품 평가와 관련한 다양한 인체시험 자료를 수집하여 인체시험 설계법 및 통계분석법을 조사하였다. 현재 국내의 주름의 인체시험 자료 검증에 이용되는 통계분석 방법의 타당성을 검토한 결과 반복측정 자료로서 서로 독립이 아닌 기존에 검정방법(unpaired *t*-test, paired *t*-test)으로 여러 번 반복할 경우 제I종 오류를 범할 확률이 높아진다는 사실을 알 수 있었다.

기능성 화장품에서 얻어지는 데이터는 한사람에게 처리를 하고 반복적으로 측정한 결과를 얻게 된다. 반복측정 자료는 한 사람에게서 반복측정한 데이터로 주로 의학, 약학 분야의 연구에서 흔히 다루어진다. 이렇게 반복 측정된 데이터들은 상호 의존적이며 다른 설계 기법과는 구별되어야 한다. 그리하여 본 연구에서는 Repeated measures ANCOVA, Repeated measures ANOVA는 반복측정 자료의 분석에서 평균의 변화 양상이나 평균들 간의 차이에 대한 다중 비교가 가능한 분석법을 제시하였다. 이들 통계법의 적용으로 본 연구에서는 실제로 통계적으로 유의한 차이가 없다고 판정된 동일한 결과를 Repeated measures ANCOVA, Repeated measures ANOVA를 적용하여 분석한 결과 반대의 결과를 보여주는 예를 확인하였다. 결론적으로 반복측정 자료의 분석에서는 처치 평균 변화 양상이나 처치 평균들 간의 차이에 관한 다중비교가 주요 관심이 되므로 시험 설계 기법에 따른 적절한 통계 처리의 필요성은 필수적이다.

본 연구를 통하여 기능성 화장품의 인체시험 설계 시 반복측정 자료의 적합한 통계방법을 제시함으로서 향후 기능성 화장품에 대한 임상시험이 보다 체계적으로 수행될 뿐 아니라 통계적 검정력을 높이게 되는 역할을 하게 될 것으로 기대된다.

감사의 글

본 연구는 2004년도 식약청 용역연구개발사업비에 의해 수행되었다(04072기화안 191).

Reference

1. 국내임상시험의 현황과 발전방안, 34th Forum on Health Industry Promotion, Korea Health Industry Development Institute (2003).
2. 화장품안전성관리사업 연구보고서II. 식품의약품안전청, **1**(2) (2001).
3. K. Katz, G. Crawford, D. Lu, J. Kantor, and J. Margolis, Statistical reviewing policies in dermatology journals: results of a questionnaire survey of editors, *J. Am, Acad, Dermatol.,* **51**(2), 234 (2004).
4. D. Altman, Related Articles, The scandal of poor medical research. *BMJ.* **29**, 283 (1994).
5. H. Ahn, S. Kim, Y. Kye, Statistical Methods Used in Articles of the Korean Journal of Dermatology. *Korean Journal of Dermatology*, **44**(3), 281 (2006).
6. IEC, Anti-wrinkle effect. France (2001).
7. RCTS Inc., Clinical efficacy evaluation of skin care product, containing indole-3-aceticacid, designed to reduce the appearance of surface fine lines and wrinkles. USA (1999).

8. DERMSCAN, Evaluation, *IN VIVO* on Human Subjects, of the Anti-Wrinkle, Moisturizing and Firming Effects of a Product. FRANCE (2000).
9. preDerm, Study Report-Marketing, Skin care home-in-use. Germany (2004).
10. LIPOTEC S.A., Study of anti-wrinkle effect of Argireline. Spain (2001).
11. J. Cohen. Statistical Power Analysis. *Current Directions in Psychological Science,* **1**(3), 98 (1992).
12. S. Nakagawa. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology,* **15**(6), 1044 (2004).
13. 박용규, 송혜향. 반복측정과 교차계획, 자료의 분석법. 자유 아카데미사, 서울 (1998).
14. M. Lawrence, C. Friedman and D. Furberg, Fundamentals of Clinical Trials, Third Edition (1998).
15. R. Carter, S. Sonne, and K. Brady. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC Med. Res. Methodol.,* **30**, 11 (2005).
16. M. Shermer. The Skeptic Encyclopedia of Pseudoscience 2 volume set. ABC-CLIO. p455 (2002).
17. D. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures. 59, CRC Press, New York, Washington D.C. (2004).
18. R. Carter, Application of stochastic processes to participant recruitment in clinical trials. *Control Clin. Trials.,* **25**(5), 429 (2004).