J. Cosmet. Sci., 71, 351–360 (September/October 2020)

Bias in Sunscreen SPF Testing: A Review of Published Data

TRINA RICCI, ANDREW MARRA, KAREN RAUEN, and
MICHAEL CASWELL, *Consumer Product Testing Company, Inc., Fairfield, NJ (T.R., A.M., K.R., M.C.)*

Accepted for publication March 31, 2020.

Synopsis

Subversion bias, a type of selection bias, through manipulation of subject recruitment compromises data validity. This study explores the possibility of subversion bias in sunscreen sun protection factor (SPF) testing. It has been established that subjects with lower minimal erythemal dose (MED) values exhibit higher sunscreen SPF values. Consistency of this response is determined in subjects who participated in multiple sunscreen efficacy clinical trials. All trials determined the SPF of the sunscreen standard P2. Of the 652 subjects with greater than three observations ($n = 286$), 35 subjects consistently had values either well above ($n = 29$) or below ($n = 6$) the average SPF value of the dataset (15.6 ± 1.2). The difference between the average SPF by the subject exhibiting the highest average SPF for P2, 19.8 ± 0.9, and the subject exhibiting the lowest average SPF for P2, 12.3 ± 2.6, is 7.5 SPF units, or 61%. Recruitment strategies based on historical SPF values for an individual would be considered subversion bias. Foreknowledge of those subjects with consistent results either in favor or not in favor of SPF testing outcomes could be exploited and would provide a reason for variation in results among testing facilities.

INTRODUCTION

The principles of good clinical practice (GCP) include minimizing bias and maximizing precision (1,2). The ability to detect bias in a clinical trial is important to assess the validity of the results. Validity refers to the degree to which a clinical trial accurately delivers the specific concept (e.g., data) that is attempted. External validity refers to the extent to which the results of a study are generalizable or transferable.

The three types of clinical trial bias are information bias, confounding bias, and selection bias. Selection bias occurs when selection, enrollment, or continued participation of a subject in a clinical trial is somehow dependent on the likelihood of having the outcome of interest. Subversion bias, a type of selection bias, occurs when the clinical team manipulates subject recruitment. Different types of subversion bias can occur. Herein, we provide evidence for the possibility of subversion bias in sunscreen sun protection factor (SPF) testing.

The SPF of a sunscreen on a subject is inversely dependent on that subject's unprotected minimal erythemal dose (MED) (3–5). In 1993, Kawada et al. (3) reported data on 48 different subjects. In 1999, Damien et al. (4) reported data on 45 different subjects from five differ-

Address all correspondence to Trina Ricci at tricci@cptclabs.com.

ent sunscreens (i.e., P3 reference product: mean SPF 4.5; Homosalate reference product: mean SPF 15.5; commercial products 1 and 2: SPF 15+; commercial product 3: SPF 30+). Both studies determined that subjects with lower MED values exhibited higher SPF values (4). More recently, in 2019, this inverse relationship of subjects with lower unprotected MEDs exhibiting higher SPF values was confirmed by Alejandria et al. (5). In a recent article in *Journal of Cosmetic Sciences*, Alejandria et al. (5) reported data on more than 2,500 observations (652 subjects) of sunscreen standard P2. They reported a significant dependency of the SPF obtained from the sunscreen standard P2 on the unprotected MED of the subject reported (reproduced herein as Figure 1) (5). Exploitation of this relationship has the potential to influence the validity of SPF results in a clinical trial.

Obtaining the highest SPF possible is important to sunscreen economics and sunscreen marketing; thus, selection bias (e.g., inclusion of subjects with lower MEDs)

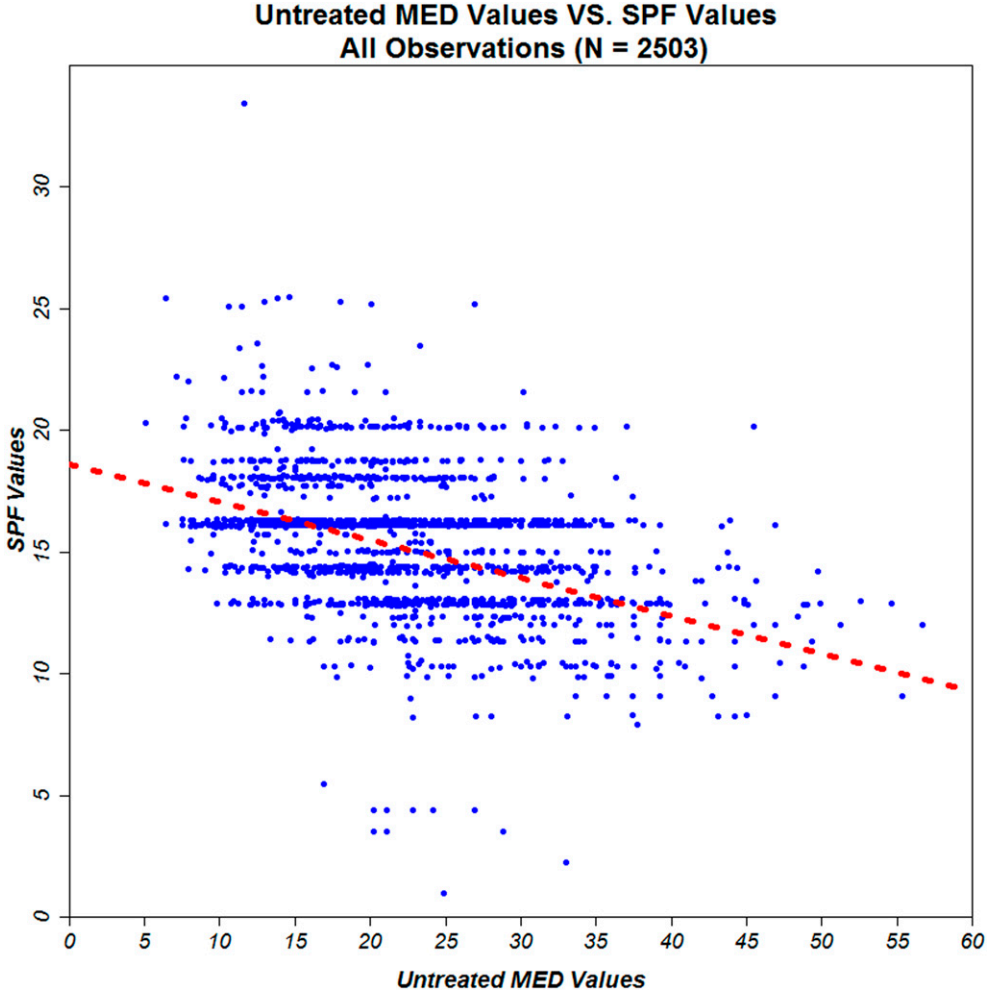

Figure 1. Relationship between a subject's unprotected MED and the SPF for standard control sunscreen P2. The dashed red line represents a regression trend line with a y intercept of 18.579 and a slope of −0.155. The regression trend line has Pearson's product moment correlation of −0.409 (5).

would influence the validity of a clinical trial to determine the SPF of a sunscreen (e.g., higher SPF values obtained). Subversion bias would occur if subjects who become known for always generating a low SPF value for the test sunscreen are excluded from future clinical trials. Similarly, subjects who become known for always generating a high SPF value might be asked to be on clinical trials. Because of the volume of observations in the data reported by Alejandria et al. (5), one might be able to discern if such individuals existed.

MATERIALS AND METHODS

STUDY DESIGN

The 2,503 observations (n = 652 subjects) depicted in Figure 1 (5) were analyzed for multiple observations on the same subject. After including only those subjects with three or more observations each, the resulting subset of data (Figure 2) consisted of 2,033 observations encompassing 286 subjects (average of seven observations per subject). The average of all observations for each of the 286 subjects' unprotected MEDs and corresponding SPF values was calculated to provide a single data point for each subject. The relationship of unprotected MED and corresponding SPF was statistically explored in this dataset.

STATISTICAL ANALYSIS

To test if the consolidation of the 2,033 observations across subjects reduced any statistical power or changed the conclusions reported on the original dataset of 2,503 observations, a linear regression and correlation analysis were performed using the same parameters presented in Alejandria et al. (5). To test for any patterns in the data, a k-means cluster analysis approach was used (6). The goal of the cluster analysis

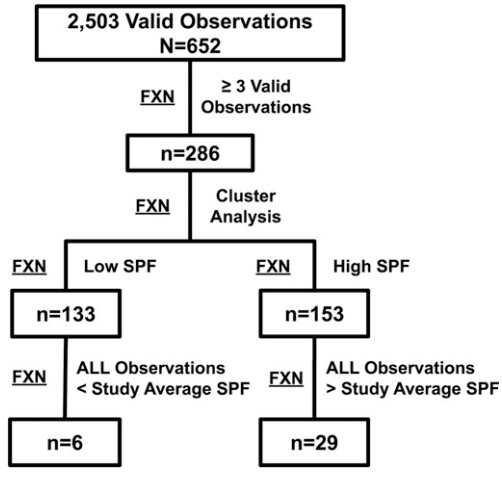

Figure 2. Study design.

was to group the 286 subjects into specific partitions based on their unprotected MED and SPF values. Only similar subjects would be present within each partition (also called a cluster), and subjects from each partition would have statistically significant different MED and SPF values versus subjects from different partitions.

To produce the clusters, the following steps were performed:

Step 1: Specify the numbers of clusters k. Randomly choose k subjects and declare them as the centers of each of the k clusters. These centers are also known as "centroids," and the value of each centroid is the average of all subjects (based on their unprotected MED and SPF values) within that specific cluster.

Step 2: Calculate the distance metric between each cluster centroid and all other data points (i.e., all other subjects) within the data set. The distance metric is the Euclidean distance between two vectors in a Euclidean space, with each vector representing a unique subject.

Step 3: Assign each subject to the cluster centroid whose distance metric is the least of all the cluster centroids. Each subject should be assigned to exactly one of the k clusters, and no subject should share multiple clusters.

Step 4: With all subjects now inside the k clusters, recalculate the new cluster centroids by calculating the average of all subjects within each cluster.

Step 5: Repeat steps 2–4, now with a newly calculated cluster centroid for each cluster. Repeat this process until the cluster centroids remain unchanged despite further iterations. When this occurs, no more subjects should be reassigned to a new cluster.

Step 1 of the k-means cluster analysis required a prespecified value for the number of clusters k. Rather than selecting an arbitrary value, an optimal value for k was determined using the silhouette method (7). This method examined a range of possible values for k. For each of these possible values, an average silhouette width was calculated. The value of k with the largest average silhouette width was selected as the optimal cluster size for the analysis.

To calculate the average silhouette width, the following steps were performed:

Step 1: Perform the k-means cluster analysis for each of the possible values of k. A range of 1–10 clusters was tested.

Step 2: Select a subject within one of the k clusters. Any subject within the cluster can be used as a starting point.

Step 3: Calculate the average distance metric between the selected subject and all other subjects within the cluster. This will be the within-cluster distance.

Step 4: Calculate the average distance metric between the selected subject and all subjects in a neighboring cluster. This will be the between-cluster distance. If there is more than one neighboring cluster, calculate the between-cluster distance for the remaining clusters.

Step 5: Compare the within-cluster distance (Step 3) with the smallest between-cluster distance (Step 4). Calculate the difference between the two values, and then divide the difference by the largest of the two values. This will produce the silhouette width, with a value ranging from –1 to 1.

Step 6: Repeat steps 2–5 by selecting another subject within the same cluster mentioned in step 2. This process will repeat until all subjects within the cluster are selected.

Step 7: Calculate the average of all silhouette widths that were calculated in step 6. This is the average silhouette width for the cluster value k.

If the clustering algorithm performed well, the within-cluster distance will be small and the between-cluster distance will be large. An average silhouette width of 1 will indicate

the most appropriate clustering, and an average silhouette width of –1 will indicate the poorest clustering performance.

After the k-means cluster analysis partitioned each subject into specific groups, descriptive statistics were calculated within each cluster. In addition, hypothesis testing using Welch's unequal variance *t*-test was performed to test for any statistically significant differences between the clusters. Statistical significance was achieved at the 95% confidence level ($p < 0.050$). Finally, subjects with extreme SPF values were evaluated within each cluster and reported.

STATISTICAL SOFTWARE

Statistical software R (version 3.6.1 for Microsoft Windows; R Foundation for Statistical Computing, Vienna, Austria) was used for all data analyses (8). In addition to the base package preinstalled with software, the packages "tidyverse," "cluster," "factoextra," and "ggplot2" were also used for the cluster analysis and for graphical plots.

RESULTS

SELECTION BIAS

Alejandria et al. (5) reported Pearson's product-moment correlation coefficient of –0.409 when evaluating the relationship between an observation's unprotected MED value and the resulting SPF value ($n = 2{,}503$) (Figure 1). In addition, the trend line from the regression analysis had an intercept of 18.579 and a slope of –0.155. By comparison, the correlation coefficient of the subject-specific data ($n = 286$ subjects) revealed a correlation coefficient of –0.478, and the trend line from the regression analysis revealed an intercept of 18.098 and a slope of –0.116 (Figure 3).

Before performing the k-means cluster analysis, the silhouette method calculated 10 average silhouette widths, one for each of the possible values of k. The average silhouette widths ranged from 0.000 to 0.395. The largest average width of 0.395 was associated with a cluster size of 2, and the second largest average width of 0.341 was associated with a cluster size of 6. Using the optimal cluster amount suggested by the silhouette method, the k-means cluster analysis revealed two groups of subjects sharing similar unprotected MED and SPF values. These two clusters—labeled as "high SPF" and "low SPF"—had sample sizes of 153 and 133, respectively (Figure 4). When comparing the two clusters, the "high SPF" cluster revealed a statistically significantly greater average SPF value (as well as a lower average unprotected MED value) than the "low SPF" cluster ($p < 0.001$). For the "high SPF" cluster, the average SPF value was 16.314 and the average unprotected MED value was 18.366. For the "low SPF" cluster, the average SPF value was 14.708 and the average unprotected MED value was 25.671.

SUBVERSION BIAS

To further characterize the potential impact of a subject's MED response on SPF results, subjects with extreme SPF values were evaluated within each cluster. In the original

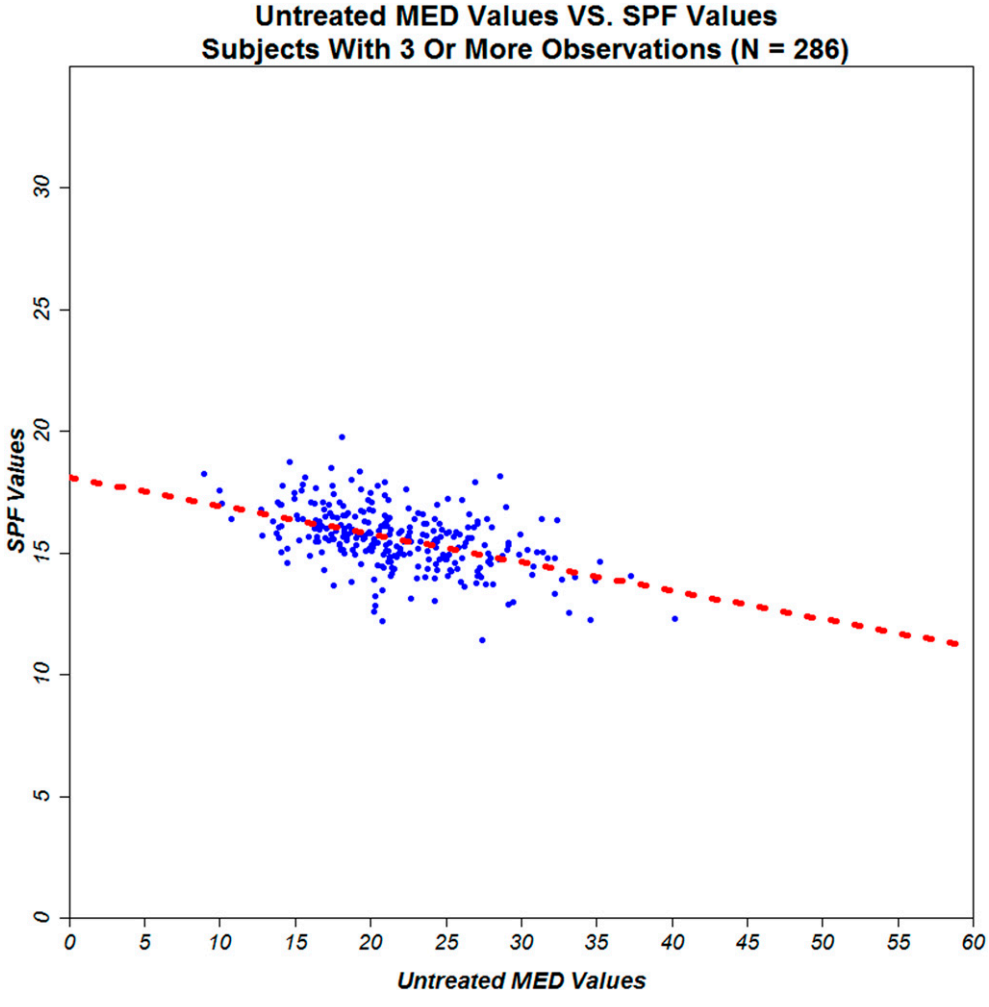

Figure 3. Relationship between a subject's unprotected MED and the SPF for standard control sunscreen P2. The dashed red line represents a regression trend line with a y intercept of 18.098 and a slope of −0.116. The regression trend line has Pearson's product moment correlation of −0.478.

dataset of 2,503 observations, Alejandria et al. (5) reported an average SPF value of 15.6 ± 2.5 for all observations. Using this threshold, subjects were declared as extreme if their observations were all consistently above the global average SPF value (or, depending on the cluster, below the global average). Of those subjects with three or more valid observations (n = 286 subjects), 29 subjects returned an SPF value that was consistently above the global average SPF of P2 (Table I) and six subjects returned an SPF value that was consistently below the global average SPF of P2 (Table II).

One subject (#53777) had 10 SPF observations that were all above the average SPF, ranging from 0.5 to 6.0 units, which gave rise to a subject average SPF value for P2 of 17.5 ± 2.0. Another subject (#81609) had only three observations, and all were above the average SPF, ranging from 3.2 to 4.7, which resulted in a subject average SPF value for P2 of 19.8 ± 0.9. By contrast, one subject (#61224) had four SPF observations that were all

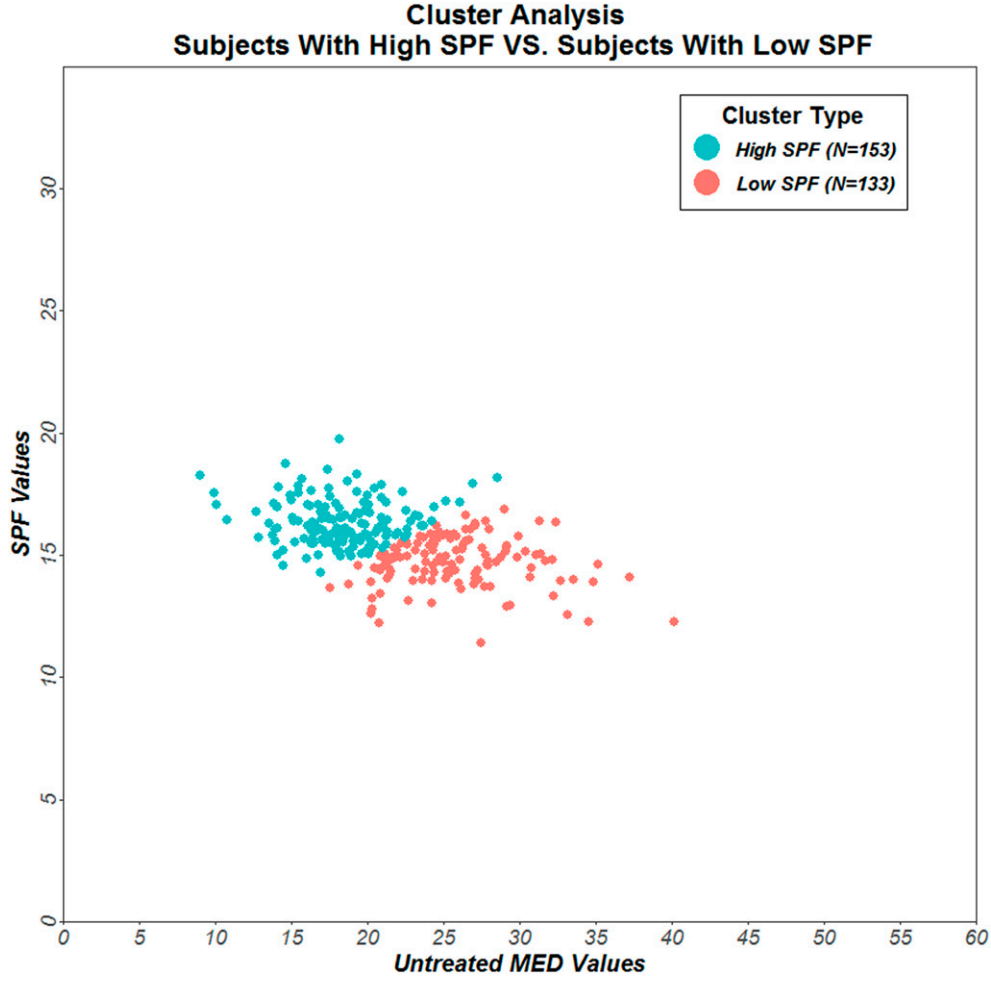

Figure 4. Subject assignments to specific clusters, based on the k-means cluster analysis. Average values for 286 subjects were partitioned into two optimal clusters. Within each cluster, subjects share similar traits regarding their unprotected MED and SPF values. A statistically significant difference between the two clusters is revealed ($p < 0.001$).

below the average SPF, ranging from 1.2 to 6.5, which resulted in a subject average SPF value for P2 of 12.3 ± 2.6.

DISCUSSION

SELECTION BIAS

The findings in this study are consistent with those reported in Alejandria et al. (5), in which the same inverse relationship of subjects with lower unprotected MEDs exhibiting higher SPF values was determined. The results of the current analysis support the findings reported in the previous three scientific articles on this topic (3–5).

Table I
Subjects (n = 29) with All SPF Values above the Average SPF of P2

Subject #	Lowest SPF value	Highest SPF value	Sample size	Counts above 15.6
1659	16.098	18.730	8	8
5032	16.086	21.571	6	6
6855	16.096	18.740	5	5
8141	16.111	18.737	5	5
10753	16.100	17.211	4	4
11577	16.297	23.584	5	5
20366	15.783	16.296	6	6
22203	15.717	16.312	3	3
28785	17.990	18.753	3	3
42199	16.099	20.106	4	4
53777	16.085	21.563	10	10
54633	16.116	18.060	4	4
55995	16.096	18.000	4	4
56638	16.101	16.180	4	4
60134	16.088	16.104	3	3
61257	16.119	20.114	8	8
61918	15.707	16.322	4	4
66482	16.092	18.751	3	3
68965	16.090	20.139	4	4
78620	16.113	20.356	3	3
78794	16.100	20.096	4	4
78860	16.100	16.295	7	7
80317	16.327	20.139	4	4
81248	16.286	18.770	6	6
81586	16.299	16.300	3	3
81609	18.772	20.300	3	3
81783	16.083	18.800	6	6
81840	16.110	20.116	3	3
81889	16.094	16.100	3	3

The cluster analysis algorithm revealed that a large sample of subjects can potentially be flagged as either a "high SPF" or "low SPF" group. This grants the ability to cull any subjects with historically low SPF values and retain only those subjects with higher SPF values. As the sample size increases with the recruitment of new subjects and with multiple iterations of the cluster analysis, the algorithm increases in precision to a point where an optimal cluster amount is only 1 (i.e., no clustering is possible). When this conclusion is reached, the dataset will consist mainly of those subjects with historically large SPF values. Regardless of the sample size, because the sample is no longer random, any statistical analysis performed will be heavily biased in favor of producing a higher SPF. This bias is an example of selection bias, where selection of subjects with lower MED values will result in higher SPF values.

Sunscreen testing facilities face not only the pressure to produce maximum SPF values which leads to this form of selection bias but also the demand to minimize testing duration. Subjects with lower MED values will require less irradiation time, which becomes more exaggerated with very high SPF sunscreens. For example, a subject with a MED of approximately 20 mJ/cm^2 will receive exposure for 46 min on a test site applied with a

Table II
Subjects ($n = 6$) with All SPF Values below the Average SPF of P2

Subject #	Lowest SPF value	Highest SPF value	Sample size	Counts below 15.6
4703	10.440	15.000	3	3
31524	12.836	14.368	4	4
61224	9.061	14.387	4	4
66837	13.064	14.384	3	3
71172	11.400	14.369	3	3
71862	11.583	14.983	3	3

sunscreen having an expected SPF of 75. A subject with a MED of 45 mJ/cm^2 will receive exposure for approximately 103 min on a test site applied with the same sunscreen. Thus, the subject with the higher MED will dramatically reduce the throughput of the testing facility, thereby reducing profitability. There is much pressure on testing facilities to commit selection bias, in conflict with GCP (1,2). Selection bias for subjects with historically low MED values for whatever reason would result in a higher SPF of a sunscreen determined by one testing facility versus another testing facility that adheres to a GCP compliant recruitment procedure.

SUBVERSION BIAS

Additional analysis presented herein reveals that a subset of subjects consistently presented with values either well above or well below the average SPF value of the dataset (15.6 ± 1.2). This subset of subjects may be instrumental in affecting the SPF value of a sunscreen in a clinical trial. For example, the difference between the average SPF by the subject exhibiting the highest average SPF for P2, 19.8 ± 0.9, and the subject exhibiting the lowest average SPF for P2, 12.3 ± 2.6, is 7.5 SPF units, or 61%. In addition, although this difference in SPF value is the maximum found within this dataset on P2, this difference in SPF value may be even greater within a dataset on a sunscreen having a higher SPF value (4).

This subset of subjects appears to be separate and independent of the relationship between MED and SPF, which causes selection bias. The subject exhibiting the highest average SPF for P2 had an average unprotected MED of 14.9 mJ/cm^2, and the subject exhibiting the second highest average SPF for P2 had an average unprotected MED of 18.1 mJ/cm^2. The subject exhibiting the lowest average SPF for P2 had an average unprotected MED of 40.2 mJ/cm^2.

A testing facility might be more likely to commit subversion bias near the end of a SPF test. At the end of an SPF test, one more or two more high SPF values might be needed to obtain the expected SPF. However, the bias would be much more dramatic if conducted throughout all 10 subjects in an SPF test. Exploiting this difference among subjects (subversion bias) would provide another reason for variation in results among testing facilities. Subversion bias for subjects with historically high SPF values, either consciously or subconsciously, would result in a higher SPF of a sunscreen for one testing facility versus another testing facility that adheres to a GCP compliant recruitment procedure.

CONCLUSION

Freeing sunscreen testing from selection bias and from subversion bias would be a worthwhile goal in enhancing the validity of SPF testing results. Currently, the FDA method (9) invalidates an SPF test if all subjects were of the same Fitzpatrick skin phototype. The correlation between Fitzpatrick skin phototype and MED is poor. A test could easily incorporate SPF test values from nine values from Fitzpatrick skin phototype 1 and one value from Fitzpatrick skin phototype 2 while committing selection bias and/or subversion bias.

External validity will be improved by requiring subjects across all MED values as suggested by Alejandria et al (5). They suggested that each valid SPF test includes at least three subjects with a MED of ≤ 15 mJ/cm^2, at least three subjects with a MED between 15 mJ/cm^2 and 40 mJ/cm^2, and at least three subjects with a MED of ≥ 40 mJ/cm^2. This would minimize the selection bias reported by Kawanda et al. (3), Damien et al. (4), and Alejandria et al. (5).

External validity will also be improved by restricting the use of individual subjects. Using a subject, no more than six times per year and no more than once every 60 days would minimize, but not eliminate, the potential for subversion bias. Until selection bias and subversion bias are eliminated, variations in SPF values from different testing facilities will continue.

REFERENCES

(1) J.A. Lewis. Statistical principles for clinical trials (ICH e9): an introductory note on an international guideline. *Statist. Med.*, **18**, 1903–1942 (1999).
(2) WHO, *Handbook for Good Clinical Research Practice (GCP)* (World Health Organization, Geneva, Switzerland, 2002), p. 30.
(3) A. Kawada, T. Noda, M. Hiruma, A. Ishibashi, and S. Arai, The relationship of sun protection factor to minimal erythema dose, Japanese skin type, and skin colour, *J. Dermatol.*, **20**, 514–516 (1993).
(4) D. L. Damian, G. M. Halliday, and R. S. Barnetson, Sun protection factor measurement of sunscreens is dependent on minimal erythema dose, *Br. J. Dermatol.*, **141**, 502–507 (1999).
(5) M. Alejandria, A. Marra, G. Roberts, and M. Caswell, Disparate SPF testing methodologies generate similar SPFs. II. Analysis of P2 standard control SPF data, *J. Cosmet. Sci.*, **70**, 181–196 (2019).
(6) J. A. Hartigan and M. A. Wong, Algorithm AS 136: a K-means clustering algorithm, *J. Roy. Stat. Soc. C Appl. Stat.*, **28**, 100–108 (1979).
(7) P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
(8) R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019), accessed January 21, 2020, www.R-project.org.
(9) Department of Health and Human Services, Food and Drug Administration, Sunscreen drug products for over-the-counter human use; proposed amendment of final monograph; proposed rule, *Fed. Regist.*, **72**, 49070–49122 (2007).