J. Soc. Cosmet. Chem., 48, 283–288 (November/December 1997)

Meta-analysis is a cost-effective tool for estimating mildness differences

P. B. NEUMANN, K. D. ERTEL, B. H. KESWICK, and G. Y. RAINS, *The Procter & Gamble Company, Sharon Woods Technical Center, 11511 Reed Hartman Highway, Cincinnati, OH 45241.*

Accepted for publication December 1, 1997.

Synopsis

Meta-analysis is a quantitative method of combining results from independent studies to form an overall conclusion based on all available data. The objective of this research was to validate the application of meta-analysis as a tool to estimate treatment differences over a series of small, randomized, blinded, pilot studies that followed the same protocol.

Three forearm controlled application technique (FCAT) screening studies were conducted to compare dryness and IBS capacitance differences induced on forearms treated with two personal cleansing bars. In each study, subjects' forearms were washed twice daily for five days on randomly assigned fixed volar sites according to a procedure that simulates normal consumer use of personal cleansing bars. Visual dryness and IBS capacitance data were collected at baseline and after five days. Both weighted and unweighted meta-analyses were performed to estimate the difference between the two treatments. An FCAT study that enrolled 105 subjects was then conducted to compare the same two cleansing bars.

Meta-analyses estimates of treatment differences derived from the screening study data closely paralleled the estimates obtained in the larger study. The visual dryness difference was estimated to be 0.167 ± 0.079 by the weighted meta-analysis and 0.168 ± 0.080 by the unweighted meta-analysis. P-values were 0.035 and 0.036, respectively. The visual dryness difference was estimated to be 0.194 ± 0.040 by the larger study. For IBS capacitance, the weighted meta-analysis estimated the difference between treatments to be −0.039 ± 0.013 (p-value = 0.004) and −0.042 ± 0.015 (p-value = 0.006) for the weighted and unweighted meta-analysis, respectively. The larger study estimated the treatment difference to be −0.023 ± 0.004 (p-value < 0.0001). This example demonstrates that estimates obtained from studies pooled by meta-analysis can adequately predict the results obtained from a single large-base-sized trial.

INTRODUCTION

The purpose of this article is to demonstrate with a practical example how closely estimates of treatment differences derived by pooling data from several small clinical trials correspond to results obtained from a larger clinical trial when the studies are all conducted according to the same protocol.

Meta-analysis is a statistical method for systematically pooling data from a series of studies to obtain a more robust estimate for a treatment effect than could be obtained from an individual study. Meta-analysis techniques to combine studies in a statistically

legitimate manner were first published by Tippett in 1931. These techniques are widely recognized today in the statistical and clinical literature, even though the actual term "meta-analysis" was introduced by Glass in 1976 in a study of psychotherapy (1–11). Chalmers estimated in 1991 that in excess of 150 meta-analyses of randomized clinical trials had been published in the English language, and that new ones were appearing at an increase of over 15% per year (12). Meta-analysis has been applied to a diverse number of problems ranging from agriculture to medicine, psychology, education, and even cloud seeding (12). It is a particularly powerful technique for pooling data from several studies conducted under a common protocol with common treatment comparisons.

In developing a new product, numerous small screening studies are normally conducted to guide final formulation. These screening studies represent a significant time and capital investment for a company, and can be an important resource for estimating treatment effects. For example, multiple small studies to investigate formulation and/or processing variables provide the researcher with information about the range of factors that can be altered without affecting product performance.

Screening studies are often considered individually when making decisions at a particular stage of a project, with a larger study conducted at the end of development to support claimed benefits prior to marketing the product. For example, suppose that a researcher has tested numerous formulation or processing variables in the development of the final product. Each of these small studies provides the researcher with information about the range of factors that can be altered without affecting product performance, even though the studies may not be large enough to show statistically significant differences versus a benchmark control. However, combining data from a number of these small studies can produce an estimate of treatment differences that is more reliable and more generally applicable than data from a single large study. The reason for this is that a single large study is usually conducted under a limited breadth of factors that could impact performance. Examples of these types of factors are weather conditions, season of the year, geographical location, local population, and product batches. By combining data from smaller studies, a more realistic estimate of the true treatment effect can be made.

Due to the wide application and popularity of meta-analysis techniques, there is a growing number of publications providing cautions and guidelines for the appropriate use of meta-analysis. These focus on issues such as establishing criteria for the inclusion of studies in meta-analysis, evaluating the quality and weighting of individual trials, and minimizing publication bias (3,8,13–15). As summarized by Olkin in *Science,* "Doing a meta-analysis is easy. Doing one well is hard" (16).

METHODS

Three forearm controlled application technique (FCAT) screening studies were conducted from April 1994 to February 1995, comparing the clinical mildness of a new mild cleansing bar with a marketed mild cleansing bar as a benchmark (17). Visual dryness and IBS capacitance data from these studies were pooled by meta-analysis techniques to estimate the difference in mildness between these two mild bars. In March 1996, an FCAT study enrolling 105 subjects was conducted to compare treatment difference estimates provided by the meta-analysis.

STUDY DESIGN

All four studies were conducted according to a forearm controlled application technique protocol (FCAT) (17). The FCAT technique has been shown to yield relative mildness rankings that correlate with consumer experience. Subjects were healthy female volunteers, aged 18–55, with Fitzpatrick skin types I–IV and forearm visual redness and dryness grades less than 3.0 on a 0–6 point scale with potential 0.5 increments. The studies were limited to female subjects because the bars compared were both intended to be marketed to women who use beauty bars. All subjects provided informed consent. Each screening study enrolled 20–25 subjects, whereas the larger study enrolled 105 subjects. The former studies were designed only to show directional rankings of product mildness. The latter was sized to show a difference as large as 0.17 in visual dryness (the estimate from meta-analysis), with at least 95% confidence and 80% power.

Each study was randomized according to a Latin Square design. In this design, subjects received all treatments, and treatments were represented approximately the same number of times on each of the eight treatment sites on the volar forearms. This is a well-recognized design that has been employed since the early 1950s (18,19). It accounts for biological differences among subjects and between application sites on the arm of a given subject, and for the order of treatment application, so that these differences do not unfairly bias treatment outcomes.

TEST PRODUCTS

The test products of interest were two commercially available mild cleansing bars. Each of the three screening studies included other personal cleansing bars and liquids that addressed objectives outside this publication. Only treatment estimates for the two mild cleansing bars of interest were common to all three screening studies and were included in the meta-analyses. The larger study included only the two mild cleansing bars of interest, with a total of four replicates of each being assigned to each subject.

EVALUATIONS

The skin condition on each treatment area was evaluated by an expert grader at baseline and three hours after the final study wash. A 0–6 scale with half-point increments was used to score dryness. Skin capacitance readings were also collected at each of these time points in duplicate, using a skin surface hydrometer manufactured by I.B.S. Co., Ltd., Japan. This measurement provides an indirect measure of moisture content in the skin.

DATA ANALYSIS

An analysis of variance model for the individual study data that accounted for subject-to-subject, side-to-side, and site-to-site variability and treatment effect was used in each study. These sources of variability have been noted historically (20). The general model for an observed response in this case can be expressed by:

$$\text{response}_{ijklm} = \mu + T_i + S_j + A_k + P_l + SA_{jk} + \epsilon_{ijklm}$$

where μ is the grand mean; T, the effect due to the ith treatment; S, the effect due to the jth treatment site; A, the effect due to the side (right or left) to which the treatment is applied; P, the effect due to the lth subject; SA, the site-by-site interaction term; and ϵ, an error term that includes experimental error and error due to uncontrolled factors.

From this model, least-squares means for treatment effects, with associated standard errors, were estimated. Skin capacitance data were log transformed prior to analyses to stabilize the variance and correct the skewness of the distribution. A logarithmic relationship between skin capacitance and moisture content had been presented previously (21).

To combine the three screening studies, both unweighted (equally weighted) and weighted meta-analyses were performed on dryness and skin capacitance data according to documented techniques to estimate overall mean differences between the products (3,4). Appropriate tests for homogeneity were performed prior to pooling the data. Weights were derived from the variability associated with treatment differences for the weighted analyses. Simpler techniques, such as Fisher's technique and the inverse normal technique (4), exist for establishing overall p-values over a series of tests. However, the techniques chosen had the advantage of estimating the actual difference between treatments as well as estimating overall p-values, whereas the simpler procedures do not provide estimates of treatment differences. The estimates from the weighted and unweighted analyses were compared with each other and with the results of the larger study for consistency.

The protocol for meta-analysis was more straightforward than those usually encountered when pooling treatment estimates from literature sources. All the data that had been generated by the FCAT protocol comparing the two bars were available and were included in the meta-analyses. Thus, publication bias and selective inclusion of studies were not issues. Publication bias is the phenomenon whereby positive results that show treatment differences are more likely to be published than those that do not. All data were generated through a contract research organization. Each of the studies was randomized, blinded, and conducted in compliance with Good Clinical Practices guidelines. Thus, the quality of the studies included in the meta-analysis is not an issue.

RESULTS AND DISCUSSION

Weighted meta-analysis combining data from the screening studies shown in Table I estimated the visual dryness difference between the two mild cleansing bars, denoted bars 1 and 2, respectively, to be 0.167 ± 0.079 (p-value = 0.035). The p-value is the probability that a difference as large as the one observed would be detected by random

Table I
Day 5 Changes From Baseline in Visual Dryness Estimates From Screening Studies

Clinical trial	n	Bar 1	Std. error	Bar 2	Std. error	Mean delta	P-value
Screening study 1	22	0.317	0.075	0.519	0.104	0.202	0.1189
Screening study 2	20	0.467	0.073	0.698	0.127	0.231	0.1183
Screening study 3	20	0.567	0.098	0.637	0.098	0.070	0.9564

chance if no true difference existed. The estimated difference from the unweighted analysis was 0.168 ± 0.080 (p-value = 0.036). The visual dryness difference between treatments estimated from the larger study was 0.194 ± 0.040 (Table III). The p-value was less than 0.0001. This showed bar 1 to be significantly less drying than bar 2.

For IBS capacitance, the weighted meta-analysis of the screening data shown in Table II estimated the difference on a logarithm scale between the two bars to be −0.039 ± 0.013 (p-value = 0.004). The estimated difference from the unweighted meta-analysis was −0.042 ± 0.015 (p-value = 0.006). The larger study estimated the treatment difference to be 0.023 ± 0.004 (p-value = 0.0001, Table III). Sites treated with bar 1 showed significantly higher capacitance than sites treated with bar 2.

Data from five subjects from the screening studies could not be used in the meta-analysis because the subjects withdrew prior to the final examination for non-product-related reasons. Six subjects were not included in the analysis of the larger study due to withdrawal prior to the final examination. Five of these subjects withdrew for non-product-related reasons. The investigator indicated that the remaining subject who withdrew may have had a product-related effect. Upon investigation it was discovered that the subject exhibited redness on one of the eight treatment sites, but had also spilled hot coffee on this site. None of the other three treatments sites receiving the same product showed redness.

CONCLUSION

Estimates of treatment differences from the meta-analyses of the screening study data closely paralleled the estimates derived from the large-based study and had the advantage of controlling for more external sources of variability. This validates the use of meta-analysis of smaller screening studies conducted under the same protocol to reliably estimate treatment differences from clinical tests to evaluate product performance attributes.

Pooling a series of smaller studies by meta-analysis is a cost-effective way of estimating the difference between treatments. When meta-analysis is used in this manner, it is important that all the data be included. Otherwise a clear explanation of why data were omitted should be provided because selective inclusion of studies can lead to a biased answer. When meta-analysis is used to combine data collected under a protocol that is reasonably consistent with normal use conditions, as with the FCAT protocol, the treatment estimates provide a reliable indication of the treatment effects that will be experienced when the products are marketed to consumers.

Table II
Day 5 Changes From Baseline in Log10 (Skin) Capacitance Estimates From Screening Studies

Clinical trial	n	Bar 1	Std. error	Bar 2	Std. error	Mean delta	P-value
Screening study 1	22	−0.033	0.009	−0.069	0.013	−0.036	0.0227
Screening study 3	20	−0.032	0.018	−0.079	0.018	−0.047	0.0635

Note: skin capacitance was not collected in screening study 2.

Table III
Changes From Baseline from the Definitive Study

Endpoint	n	Bar 1	Std. error	Bar 2	Std. error	Mean delta	P-value
Dryness	99	1.041	0.028	1.235	0.028	0.194	<.0001
Log10 (skin capacitance)	99	−0.023	0.003	−0.046	0.003	−0.023	<.0001

REFERENCES

(1) W. G. Cochran, Problems arising in the analysis of a series of similar experiments, *J. Roy. Stat. Soc. Suppl.,* **4**, 102–118 (1937).
(2) W. G. Cochran and G. M. Cox, *Experimental Designs* (John Wiley & Sons, Inc., New York, 1957), pp. 115–129, 545–568.
(3) R. B. D'Agostino and M. Weintraub, Meta-analysis: A method for synthesizing research, *Clin. Pharmacol. Therapeut.,* **58**, 606–616 (1995).
(4) L. V. Hedges and I. Olkin, *Statistical Analysis for Meta-Analysis* (Academic Press, Orlando, 1985), pp. 27–117.
(5) K. A. L'Abbe, A. S. Detsky, and K. O'Rourke, Meta-analysis in clinical research, *Ann. Intern. Med.,* **107**, 224–233 (1987).
(6) T. A. Louis, H. V. Fineberg, and F. Mosteller, Findings for public health from meta-analyses, *Ann. Rev. Publ. Health,* **6**, 1–20 (1985).
(7) H. S. Sacks, J. Berrier, D. Reitman, V. A. Ancona-Berk, and T. C. Chalmers, Meta-analysis of randomized control trials, *N. Engl. J. Med.,* **316**, 450–455 (1987).
(8) S. B. Thacker, Meta-analysis: A quantitative approach to research integration, *J. Am. Med. Assoc.,* **259**, 1685–1689 (1988).
(9) K. W. Wachter, Disturbed by meta-analysis?, *Science,* **241**, 1407–1408 (1988).
(10) N. M. Laird and F. Mosteller, Some statistical methods for combining experimental results, *Intl. J. Technol. Assess. Health Care,* **6**, 5–30 (1990).
(11) G. V. Glass, Primary, secondary, and meta-analysis of a large collection of research, *Educ. Res.,* **5**, 3–8 (1976).
(12) I. Olkin, Keynote address: Meta-analysis: reconciling the results of independent studies, *Stat. Med.,* **14**, 457–472 (1995).
(13) T. C. Chalmers, Problems induced by meta-analyses, *Stat. Med.,* **10**, 971–980 (1991).
(14) J. C. Bailar III and F. Mosteller, Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations, *Ann. Intern. Med.,* **108**, 266–273 (1988).
(15) R. J. Light, Accumulating evidence from independent studies: What we can win and what we can lose, *Stat. Med.,* **6**, 221–228 (1987).
(16) C. Mann, Meta-analysis in the breech, *Science,* **249**, 476–480 (1990).
(17) K. D. Ertel, B. H. Keswick, and P. B. Bryant, A forearm controlled application technique for estimating the relative mildness of personal cleansing products, *J. Soc. Cosmet. Chem.,* **46**, 67–76 (1995).
(18) B. J. Winer, *Statistical Principles in Experimental Design* (McGraw-Hill, New York, 1971), pp. 685–711, 397–398.
(19) G. D. Steel and J. H. Torrie, *Principals and Procedures of Statistics* (McGraw-Hill, New York, 1960), pp. 146–159.
(20) V. Rogiers, M. P. Derde, G. Verleye, and D. Roseeuw, Standardized conditions needed for skin surface hydration measurements, *Cosmet. Toiletr.,* **105**, 73–82 (1990).
(21) S. Dikstein, M. Katz, A. Zlotogorski, Y. Broun, D. Wilson, and H. Maibach, Comparison of different instruments for measuring stratum corneum moisture content, *Intl. J. Cosmet. Sci.,* **8**, 289–292 (1986).