J. Soc. Cosmetic Chemists, **18**, 323–331 (May 27, 1967)

Pitfalls and Problems in Predictive Testing

MATTHEW J. BRUNNER, M.D.*

Presented September 20–21, 1966, Seminar, New York City

Synopsis—Predictive tests of cosmetics for sensitization potential and irritancy, as now constituted, serve as guides rather than absolute criteria. Alterations in the conditions employed in these test procedures can increase or decrease the number of sensitized subjects, but the crucial question of the relationship between sensitizations in the laboratory test and in actual usage remains unanswered. Properly supervised consumer use tests are still required to supplement the laboratory studies. Further study of the basic mechanisms of sensitization is required before tests can be significantly improved.

Since the most common reactions to cosmetics involve untoward effects on the skin, a number of test procedures for new products have been suggested to aid in predicting the probable incidence of these skin reactions, as produced either by direct irritancy or by sensitization of the contact dermatitis type. The multiplicity of these procedures suggests that none is perfectly satisfactory. A good predictive test should, in advance of consumer use, be able to determine the irritating or sensitizing powers of a new formulation when these are at such a low level that they may not be revealed in a small-scale trial of ordinary usage. Even very low reaction rates may present a problem when multiplied by the millions of usages of a nationally sold product. To determine the actual rate of reaction in these low ranges, the test procedure must in some way exaggerate exposure conditions so that reactions become frequent enough for one to compare new formulations with controls in the small test groups which are practical to use. Otherwise, the "test" would consist

* 910 Via De La Paz, Pacific Palisades, Calif. 90272.

only in the supervised normal use of a product. To make sure with 95% certainty that the reaction rate in the general population is no more than 1:1000, it would be necessary to have some 3000 subjects use the product without finding a reaction. Groups of such size are impractical for preliminary tests. On the other hand, when a test involves exaggeration of exposure conditions, one cannot directly apply the results to normal consumer usage, since irritancy and sensitization rates depend to a great extent on conditions of contact.

In evaluating the irritancy of new cosmetics, predictive tests are usually considered to be useful only for screening out the more violently reactive agents. Animal test procedures, such as the Draize (1) test for irritancy which involves a simple occlusive application, can be helpful for range finding. Such tests serve for preliminary evaluation of entirely new agents, the toxicity and irritancy of which are unknown. Even in those tests employing human subjects the single application of a substance to the skin by the usual twenty-four to forty-eight hour patch test technic is an unreliable means for predicting irritancy. Both false positive and false negative reactions are possible. The greater the difference between use conditions and patch test conditions the more potentially misleading are the results. Some substances need multiple applications before irritancy results. With other agents, such as nail polish remover, a single covered patch application will give misleading positive results. It is necessary to design tests specifically for each class of formulations, changing the factors of occlusion and repetition or adding some damaging stimulus. By such means the test can be made stringent enough to produce relatively low level but measurable irritant reactions with a control formulation which in actual use has produced a tolerable level of irritancy. Then, if a new formulation produces no more irritation than the control by paired comparison, it can be assumed that the two probably will also give comparable results in actual usage. This is a safe assumption only if the usage conditions are similar. The test used must be reliable, that is, give reproducible results on repetition, and the results must, of course, be tested for statistical significance. It is necessary to choose the test panel carefully, since individuals vary considerably in their general level of reactivity to irritants, and some individuals rarely show irritancy even with the more reactive substances. The composition of panels should be such that equivalent numbers of strong and weak reactors are present in different panels. There is also an important seasonal effect on irritancy, with wintertime increases in reactivity. Absolute values are therefore not comparable for tests performed at dif-

ferent times, even on the same panel. A test subject who experiences reactions should not be used again for sixty to ninety days, since the reaction site is more sensitive to irritants and may give a false positive. It must also be remembered that the usual irritancy tests will generally not uncover such undesirable effects as drying, stinging, scaling, acneform reactions, etc., which may be of the greatest significance in consumer use.

A test procedure which incorporates features adequate for irritancy evaluation has been reported by Finkelstein, Laden, and Miechowski (2). It involves the repetitive application of the test substance and of a suitable control formulation on cotton flannel pads, occluded by polyethylene film. For materials which are actually used on a daily basis, application is made for a seventeen-hour period on four consecutive days. Substances used once or twice weekly are applied for five hours a day on five consecutive days. Irritancy reactions are scored daily, and weight is given both to the severity of reaction and the number of days required to produce the reaction. Finkelstein and his co-workers found this a fairly sensitive and reliable indicator of the irritancy of agents which fall in the low range on the Draize test. Results of the test indicate whether (as regards irritancy) it is worthwhile going ahead with usage studies and trial sales with the new formulation. With this and all similar procedures, selection of proper control formulations is important.

Predictive tests for sensitization are even more complicated than irritancy tests. It is relatively simple to screen out potent sensitizing agents, since even casual contact with these may induce reactions in laboratory or factory personnel. However, the problem is to predict accurately the sensitization rate when the formulation is only a weak sensitizer and is used by the consumer in a way which differs from the usual test procedure exposure. The predictive tests for sensitization in general use a formalized method of application, repeated one or more times in the various methods. The aim is, of course, to determine how often sensitization can be purposely induced by repeated application of the test substance. The frequency of reactions is regarded as an indicator of how the substance will act under use conditions; a judgment can then be made as to whether the level is tolerable. However, questions have been raised concerning the sensitivity of the test methods and the relationships of test results to usage results.

The first of the predictive tests was the Schwartz (3) prophetic patch test. Schwartz suggested that new cosmetics should be tested by the closed patch test method on at least 200 subjects, using as a control an old formula with a known record of safety. The patches are left on for

forty-eight hours. Repeat patch tests are made two to three weeks later. Those subjects showing reactions on the second contact which were not present on the first are deemed sensitized. According to Schwartz, if the new formula shows more reactions than the control it is unsafe. Schwartz admits that the test may "not give an accurate idea of what may happen under conditions of actual use." Therefore, a four-week paired comparison use test of the cosmetic on the same 200 subjects is recommended before trial sale. In this part of the test the occurrence of more than one case of dermatitis indicates that the formulation is unsafe. Trial sale, the final step by Schwartz' definition is the sale and use of 5–10,000 units in one community. The Schwartz prophetic patch test has certain inherent defects. False positives may occur, since borderline primary irritants can sometimes produce reactions which can be confused with sensitizations when only a patch test reading is made. More important are the false negatives, which are due to the fact that the single application of a small amount of the cosmetic is often inadequate to produce sensitization, except in the case of strong allergens. The use test which Schwartz recommended to follow the initial patch application is probably a recognition of this inadequacy. Most of the predictive burden is shifted to the use test, but this part of the procedure is probably numerically inadequate to reveal low reaction rates. Also, some products are used only once weekly, or every four to six weeks, and cannot be evaluated for sensitization in a four-week use test.

The Brunner-Smiljanic test (4), reported in 1952, attempted to surmount some of the problems of the prophetic patch test by increasing the frequency of application and using a larger area of contact with the test substance. Following an initial standard patch test, daily thirty-minute repetitive applications of the test substance are made to the forearm using saturated 7.5 × 7.5 cm gauze squares on three different sites in rotation. After an initial five-day application period, there is a rest period of one week followed by another ten-day application period. A standard patch test is repeated at the end of the series. With this method primary irritant reactions are easily distinguished from sensitization since irritant reactions fade by the day following application, when the test sites are inspected. Sensitization reactions consist of persistent redness, sometimes with blistering and swelling, at the application sites. In many cases spontaneous flares occur in the sites of previous applications at the time sensitization is induced. Standard patch tests also become positive in the subjects developing sensitization. Groups of 12 subjects are tested consecutively; if more than one reaction occurs in the

first group, the test is generally not continued. Usually three to five groups are tested. With this technic it has been possible to show lack of allergenicity in substances such as ammonium thioglycolate and ammonium bisulfite, which produce few allergic reactions in consumer use, while 50–100% of subjects reacted with potent sensitizers such as mercaptohydrazides, mercaptoethane sulfonic acid, etc. The latter agents produced a few sensitization reactions in laboratory personnel who had only casual contact with them during testing. Thioglycerol, which has produced a moderate number of sensitization reactions in consumer use, sensitized about 10% of the test subjects.

This test is based on the following concepts: First, for any given substance there is a certain upper limit in the number of individuals who will become sensitized to it. Secondly, that one of the factors which determine the completeness with which this potential index of sensitization is exploited is the number of contacts. It is not necessary to keep up applications indefinitely, since after a certain number of contacts there is a diminishing return of newly sensitized individuals. Calnan, Epstein, and Kligman (5), for example, report that in animals, ten injections are optimal for sensitization; fifteen are no better; and five give fewer reactions. In experimental sensitization of humans with Krameria and monobenzone (monobenzylether of hydroquinone), they found four weekly exposures gave more reactions than three. They suggest that with very weak sensitizers more exposures may be needed.

Subsequent to the report by Brunner and Smiljanic, other predictive tests for sensitization were described by Shelanski (6), Draize (7), and Traub and his co-workers (8). As in the Brunner-Smiljanic procedure, all these tests involve multiple contacts, but the factors of occlusion, site, duration and area of contact are varied. The pertinent factors in each test are shown in Table I. The Traub test obviously has only limited predictive value since it consists of three weeks of use application in a group of only 200 subjects without intensification of exposure conditions. Also, one cannot repetitively apply many substances, such as cold wave formulations or hair dyes exactly according to use in the three-week test period. One the other hand, in the Shelanski test occlusive repetitive applications to the same site give a higher yield of sensitized subjects but tend to magnify the irritant effects of the test formulation. The resulting so-called "skin fatigue," due to summation of irritations under the test conditions, may make the differentiation between sensitization and irritancy more difficult. The Draize test also uses occlusion to increase the yield of sensitized individuals but varies the application site so that

Table I
Comparison of Various Predictive Tests for Sensitization

	Number of Subjects	Site	Repetition	Occlusion	Duration of Test	Duration of Application	Area	Evidence of Sensitization
Schwartz	200	Same	2	+	14–21 days	48 hours	2.5 cm × 2.5 cm	> Redness than original
Brunner	Multiple of 12	Varied, same arm	Up to 15	0	21 days	30 minutes for hair waving and hair dyeing agents	7.5 cm × 7.5 cm forearm	Reaction >24 hours. Distant flare. Patch test at conclusion
Draize	200	Varied	Up to 10–15	+	30 days	24 hours	0.5 g or 0.5 cc	Patch test after 2 weeks
Shelanski	200	Same site	Up to 10–15	+	30 days	24 hours	2.5 cm × 2.5 cm	Patch test after 2 weeks
Traub	200	According to usage	Up to 21	0	21 days	According to use	According to use	Patch test after 2 weeks

"skin fatigue" is not a factor. However, one may question the advisability of using occlusion or other methods of facilitating penetration, such as skin damage with sodium lauryl sulfate or freezing, in predictive testing. One of the reasons that a formulation does not sensitize in actual use may be that it does not penetrate, and a false impression of allergenicity may result from artificially making the substance pass through the skin in the test procedure.

Thus, alterations in any of the test conditions can change the incidence of sensitizations. Of major interest, however, is not the rather academic question of what the potential sensitizing power of a substance or formulation may be when exploited to its limits by the most stringent conditions but how it will behave under use conditions. Can one forecast, for example, what the incidence of sensitization will be in consumer use of a hair dye (used every six weeks on the scalp, without occlusion, with exposure to the dye intermediates for a minute or two) when 1 in 50 test subjects reacts after 10 or 15 daily applications to the arm under occlusion? There is, as yet, no predictable relationship between the two situations. If there is a high yield of true reactors (for example, more than 20%) on any of the predictive tests, one may conclude that the substance is probably going to be troublesome regardless of the conditions of contact, but especially if severe auxiliary damaging stimuli have not been used. Low sensitization rates, in the range of 1–5%, are generally acceptable, particularly if this incidence is no greater than that of a known control which has already had extensive consumer use, with a tolerable level of sensitization reactions. The control must be used in the same way as the new agent and have similar features as regards penetrability, irritancy, etc. The level of sensitivity of the test should probably be fixed at the point at which the control agent (which is safe for consumer use) produces at least an occasional reaction in the test. Otherwise, the sensitivity of the test may be too low to detect potentially troublesome materials. For different classes of cosmetics it may be necessary to modify test conditions, such as frequency and duration of application, occlusion, etc. The Brunner-Smiljanic test, for example, was designed for use with mercaptan-containing hair waving formulations which are applied unoccluded for 30–60 minutes in consumer use. Alteration in time and type of contact is needed for some other types of products, such as facial creams.

With new classes of compounds animal tests may be useful for preliminary screening before any human sensitization testing. The procedure described by Voss (9) appears satisfactory for this purpose. Based on

his results, one may conclude that, if a substance produces a significant number of reactions in the guinea pig, it is also likely to produce sensitization in human testing. However, some agents produced no reactions in the guinea pig but did sensitize humans. A rational program may, therefore, involve consecutively the preliminary animal screening test, followed by an applicable human sensitization test, and finally a supervised use test by a panel of subjects. With consumer usage panels, dependable confirmatory answers on sensitization may be expected only if the panel is large, the contacts frequent, and the observation close and prolonged. Valuable information can also be obtained from market tests, which are the final step before unrestricted sales, but there must be careful follow-up of users, and the trial sale period must be long enough to allow for the repetition of contacts which is a requisite for sensitization. A trial sale period which is only long enough to allow for one or two uses of a product is inadequate for a study of sensitization.

The general principles of predictive testing for sensitization can be recapitulated as follows: To obtain a positive control an agent is applied to the skin in such a manner that it will produce allergic sensitization reactions in one or more subjects in a small test group, whereas in actual use it produces only a negligible number of reactions. If other agents applied in the same way give no more reactions than this control, it is assumed that they will be no worse in actual use than the control; they are, therefore, acceptable as regards sensitizing potential. It is the difference between test conditions and usage conditions which changes the sensitization rate. Aside from occlusion, which has an obvious effect on penetration, the main differences are amount of antigen applied, duration of contact, frequency of contact, spacing between contacts, and total duration of exposure. It is generally held that on contact the antigen combines with some component of skin protein, is absorbed and reaches the reticulo-endothelial system, where it causes certain cells to produce antibodies. The lymph nodes draining the area of skin to which the substance is applied are the important reticulo-endothelial tissues in this type of sensitization. In experimental sensitization, Calnan and co-workers (5) showed that varying the site of application reduces sensitizing potential, since this apparently reduces transport of antigen to a single group of lymph nodes to a level below the threshold required to cause modification of lymphoid cells and the formation of antibodies. Why lengthening of the interval between applications appears to decrease the percentage of reactors is not clear. Perhaps, closely spaced repetition results in an accumulation of antigen in regional lymph nodes to that

critical level necessary to induce immunologic modification in the mesenchymal cells. Or repeated delivery of antigen to the nodes will be more likely to catch the cells at that point in their cycle of growth and division at which they are most susceptible to this modification by the antigenic substance.

Returning to the practical aspects of the problem, perhaps the situation can best be summarized as follows. Results of predictive tests as now constituted serve as guides rather than absolute criteria of sensitization and irritancy. The tests function best when the test substance is used on a comparative basis against control formulations with known behavior in actual usage. Further study of the basic mechanisms of sensitization is required before the tests can be significantly improved. Alterations in current test procedures can increase or decrease the number of reactors but do not help in answering the crucial question of the relationship between sensitization results in laboratory tests and in consumer usage. Properly supervised consumer use tests are still required to supplement the laboratory studies.

(Received September 21, 1966)

REFERENCES

(1) Draize, J. H., Woodard, G., and Calvery, H. O., Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes, *J. Pharmacol. Exptl. Therap.*, **82,** 377–390 (December, 1944).

(2) Finkelstein, P., Laden, K., and Miechowski, W., New methods for evaluating cosmetic irritancy, *J. Invest. Dermatol.*, **40,** 11–14 (January, 1963).

(3) Schwartz, L., and Peck, S. M., The patch test in contact dermatitis, *Public Health Rept.*, **59,** 546–557 (Apr. 28, 1944).

(4) Brunner, M. J., and Smiljanic, A., Procedure for evaluating skin-sensitizing power of new materials, *Arch. Dermatol.*, **66,** 703–705 (December, 1952).

(5) Calnan, C. D., Epstein, W. L., and Kligman, A. M., *Methods of Evaluating Contact Sensitizers*, in Sternberg, T. H., and Newcomer, F. C., *Evaluation of Therapeutic Agents and Cosmetics*, McGraw-Hill Book Co., New York, N. Y., 1964, pp. 157–170.

(6) Shelanski, H. A., and Shelanski, M. V., A new technique of human patch tests, *Proc. Sci. Sect. Toilet Goods Assoc.*, **20,** 46–49 (1953).

(7) Draize, J. H., *Dermal Toxicity*, The Association of Food and Drug Officials of the United States, P. O. Box 1494, Topeka, Kan., 1959, p. 52.

(8) Traub, E. F., Tusing, T. W., and Spoor, H. J., Evaluation of dermal sensitivity: Animal and human tests compared, *Arch. Dermatol.*, **69,** 399–409 (1954).

(9) Voss, J. G., Skin sensitization by mercaptans of low molecular weight, *J. Invest. Dermatol.*, **31,** 273–279 (1958).