J. Soc. Cosmet. Chem., **22**, 349-359 (May 27, 1971)

Some Problems of Predictive Testing

G. W. BATTISTA, Ph.D., and M. M. RIEGER, Ph.D.*

Presented December 1–2, 1970, New York City

Synopsis—TEST PROCEDURES available for the SAFETY EVALUATION of toiletries and COSMETICS are reviewed. Some obvious deficiencies of these tests, possible pitfalls, questionable parameters, and the organization of an evaluation program are discussed. The interpretation of preclinical animal and of clinical human safety data and the importance of in-use tests in determining whether or not a product is safe are also considered. A few clinical experiences are presented to illustrate the discussion.

INTRODUCTION

Product safety is one of the primary concerns of all drug and cosmetic manufacturers. In the case of drugs, the term *safety* becomes relative, and any side effects can be weighed against the benefits derived. The safety of drugs is one of degree, and consumer inconvenience or even risk can be tolerated depending on the product and what it will do for the patient. Drugs can be and often are an essential part of one's health and, in severe circumstances, can be life saving. On the other hand, cosmetics are not life saving, but contribute to general well-being by virtue of beautification, decoration, and camouflage. Within this framework, the manufacturer of cosmetics and toiletries must assume great responsibility for consumer safety, whether it be for reasons of morality or just plain good business sense. Consumer tolerance to irritation or inconvenience resulting from the use of cosmetics is very low; there is no room for side

* Warner Lambert Research Institute, 170 Tabor Rd., Morris Plains, N. J. 07950.

effects; and safety under conditions of use must be as close to 100% as possible.

As a result, the manufacturer of cosmetics needs reliable information on the probability of adverse reactions to his product in advance of its marketing. Cosmetics in general are a fairly innocuous class of products; if they were not, the manufacturers would have had to close their doors long ago. The fact that billions of dollars are spent annually on cosmetics confirms that consumers as a whole do not find cosmetics irritating.

The skin of man has been constantly exposed to a wide variety of different chemicals. The milieu of our modern industrial society further insults our skin with a wide range of chemicals in the form of clothing, industrial finishes, and dusts. Experience has made the elimination of known skin offenders from cosmetics a fairly simple task. Most potential irritants have been identified, and their use in cosmetics is almost nonexistent. Kligman (1) pointed out, on the other hand, that "practically all substances are capable of being contact sensitizers for some persons under some conditions." However, the exclusion of all chemicals which—under one condition or other—could act as irritants or sensitizers would make the formulation of cosmetics impossible.

With this in mind it becomes apparent why an evaluation of currently available predictive test procedures is important. A description of predictive testing techniques, especially as they are applied to cosmetics, has recently been made by Brunner (2). A thoroughly annotated review of this subject has been prepared by Idson (3). In addition, a well-reasoned critique of standard test methods for cutaneous contact allergy has been published by Kligman (1), who concludes that the techniques are "insensitive" to moderately strong known sensitizers. In view of this he developed his so called "maximization test" (4).

In spite of some obvious deficiencies, predictive testing techniques for cosmetics appear to be reasonably reliable. If they were not, the incidence of reaction to newly introduced cosmetics would be expected to be much more common in our industry. In fact, we feel that a well-executed predictive program is most useful in keeping potentially troublesome products off the market.

The purpose of this discussion is not to suggest new tests. Instead, the need for careful selection of predictive testing procedures and for judicious interpretation of the resulting data will be pointed out. In addition, some problems which may be encountered during predictive testing will be explored.

Problems of Predictive Test Procedures

Availability of Test Procedures

The existing techniques are widely used, and all of them can help in predicting the safety of cosmetics if they are wisely selected and if the data obtained are properly applied. These procedures can and should be modified or exaggerated to suit the product; however, the application of results from a small-scale test to the population at large is statistically difficult (5, 6) and leaves much to be desired, especially in view of the artificial conditions which are employed in testing cosmetics.

It is probably worthwhile to mention briefly the various groups of tests that are available.

Animal tests may include the guinea pig immersion procedure (7), the "Draize" rabbit eye test (8), the standard "Draize" dermal irritation procedure (8, 9), various guinea pig sensitization tests (8, 10, 11), and others.

Human patch tests in a variety of forms have been used for many years. These may include the single (primary irritation) patch test (1, 3), the prophetic patch test (1), modification of the patch test to determine possible phototoxicity or photosensitization (3), repeated insult patch tests to determine sensitization (12–14) and, finally, the maximization test procedure of Kligman (1, 4, 15, 16).

In-use tests represent the third group of tests available (17). Although no formal procedures for this type of testing exist, the product is generally used in accordance with the manufacturer's direction after preclinical or clinical patch tests. In some instances, it may be desirable to exaggerate the conditions of use by increasing the amount and frequency of application. Preferably, the period of intensive use is followed by a rest period and then a "challenging" patch test.

Problems of Selection

The protocol for an effective predictive testing program requires a considerable amount of judgment. One of the most important limitations on test procedures is the fact that not all of them are applicable to all products, and the selection of the test is as important as the test itself. To subscribe to the concept that a product must pass a particular test before it can be considered safe for marketing eliminates all opportunity to modify a standard test. It also precludes the need for interpretation of the results obtained by the clinician and, in effect, voids his expert judgment.

In planning a well-organized safety program each preparation must be individually considered, and the tests employed should be appropriate for the product. Knowledge of the formula, familiarity with the intended use of the product, and directions for use of the marketed product are essential requirements for selecting the proper predictive test and any alterations that are needed.

We feel that, by themselves, animal tests are of limited predictive value, whether they are negative or positive (18). It is generally agreed that these tests have their principal value in providing confidence to the investigator to proceed further with tests on human subjects (10, 13). In addition, animal testing procedures often offer useful clues to potential adverse reactions in human subjects. Any reaction in animal tests thus becomes a special alert to the clinician. On the other hand, a strong positive reaction will be a warning to the investigator either to proceed with extreme caution or to abandon further testing.

Animal test procedures are highly exaggerated and rightfully so. They should be carefully interpreted by the investigator so as to be used primarily for guidance to the formulator and not for the purpose of absolute judgment on a go or no-go basis. A typical example is the guinea pig immersion test during which animals are exposed to moderately concentrated solutions of detergents for a prolonged period of time. This test will almost always result in skin reactions. The fact that damage occurs is unimportant, but the degree of damage can be used as a guide for the selection of "safe" detergents for a particular use. Similarly, one would certainly not test a waving lotion or a neutralizing solution *via* the subacute 20-day dermal irritation test with rabbits. The results of such a test would be meaningless and have no relationship to the frequency of use by humans.

Human patch tests too must be selected with great care. For example, any closed patch test technique is exaggerated because both the temperature and the humidity are raised in the occluded area and because evaporation of any volatile materials is precluded. It is also common to use the same test population on more than one occasion. Actually, repeat exposure of the same test population may appear highly desirable because this group might exhibit cumulative irritation and occasional sensitization due to exposure in previous tests. On the other hand, such a population by repeated exposure may have become refractive, thus introducing another limitation. The investigator should be aware of the potential of the test population to react, and his interpretation of positive or

negative results should be tempered by this knowledge. He must, therefore, not only select his test but also his test population.

Not every product should have to be subjected to all known tests. For example, there is no reason to subject a vaginal deodorant, nor any other product not normally exposed to light, to a phototoxicity test. Similarly, it is extremely doubtful whether or not a closed patch test of a shampoo at full strength has any meaning since shampoos normally remain on the scalp only for two to three minutes, are diluted, and are almost completely removed by subsequent rinsing. Finally, a 24- or 48-hour closed patch test of a depilatory would be out of the question and of no value.

Such use-related considerations alone form a solid rationale for the desirability of in-use testing. These procedures are also valuable for the study of topically applied products because protocol automatically includes exposure to the natural elements, such as sun, wind, heat, and cold. The importance of this test will become more apparent when some clinical experiences are discussed later. For the moment, it will suffice to indicate that in-use tests employ the product at near normal frequency in and over the area of usual application and can also take into account the possible misuse or even abuse of a product.

Problems of Interpretation

Let us assume for a moment that the investigator has established in a series of animal safety tests that he can proceed with human testing of a new cosmetic product. He has then carefully selected the proper human tests and has taken into account specific modifications to suit the product and its intended use. The results of these tests are available, and the investigator is now faced with the responsibility of establishing whether the product can be "safely distributed." It is at this point that our inadvertent errors and our lack of complete scientific knowledge can impair clinical judgment.

An important source of error is the grading of reactions which may depend entirely upon the patch test material (14) and on the observer who grades the reaction. Occasionally, reactions may be due to improper application of the patch, i.e., friction. Inadequate tissue contact of the material and improper placement (which may lead to loosening of the occlusive covering) can reduce the severity of reaction to the patch test. Climatic conditions are also believed to have an influence on the reactivity of human skin, and there is some evidence that there is greater reactivity during winter months (19, 20).

Cross sensitization and what has been termed skin fatigue or cumulative reactions add to the problems already facing the investigator. He must now, for example, consider the possible influence of drugs on the reactivity of the test subject or the ultimate user. The age of the patient influences reactivity due to lack of previous exposure or by virtue of the development of the immunological apparatus. Pre-existing skin conditions, such as eczema, may or may not have a direct effect on the final results of the patch test or on the consumer. A further problem is generated by the fact that often patch tests are conducted using more than one and sometimes as many as a dozen products at the same time. The question then arises whether simultaneous patch testing with different materials could have an influence on the reaction to one or more of the products tested. This fact could be an advantage or a disadvantage. Our inability to be more specific is, in fact, a limitation in itself. In this connection it is noted that racial influences and skin color must also be considered, although to date there is relatively little known about the sensitivity of different races to any given chemical (21). It is not surprising that heredity plays an important role in allergic contact dermatitis (22) and this may also apply to reactivity to cosmetics.

Finally, the investigator must take into account that the sensitivity potential of a population to a product can change with time. This type of latent sensitivity was studied by Baer *et al.* (23), who attribute this phenomenon to increased opportunity to exposure to allergic sensitizers. It is conceivable that the safety of chemicals or products established some years ago is no longer applicable today. The relationship of these observations to Agrup's (24) conclusion that patch testing can lead to sensitization is by no means clearly established.

It must be concluded that the interpretation of predictive human patch testing is made difficult by two factors: (a) possible experimental error and (b) the influence of unknown extraneous factors on the test subjects and eventually on the population of potential users. The careful experimenter can usually overcome the limitations due to inadvertent error. He cannot make up for the holes in our knowledge, i.e., the unknown.

Some known and some unknown parameters, both of which may influence the results of patch testing or the response during use, are shown in Tables I and II.

Many of the questions raised above exert an important influence on the results of a well-conducted in-use test. The results from such a test

Table I

Parameters Known to Influence Human Skin Reaction

Parameter	Response	Reference
Site of application (in sensitization test)	1. Back > abdomen > extremities 2. Reapplication to same site > new site	Kligman (21), Magnusson (25) Kligman (1, 21)
Mode of application (in sensitization test)	Epicutaneous > intradermal	Kligman (21)
Type of patch used	Closed > open	Magnusson (14, 25, 26)
Dose	1. Response depends on amount deposited per unit area, not total dose 2. One large dose is more sensitizing than several small doses	Kligman (21) Lowney (27)
Trauma (at site of application)	Chemical trauma > mechanical trauma	Rebello (28), Kligman (21)
Sex	Male > female	Lanman (15)

Table II

Parameters Which May Influence Human Skin Reaction

Parameter	Comment	Reference
Age	Effect in sensitization not clearly established	
Trauma to skin (at a remote site)	Probably increases reaction	Sipos (29)
Race and heredity	Probably important	Kligman (21), Forsbeck (22)
Seasonal variation	1. Response minimum occurs during summer 2. Greater reaction during winter 3. Sweating under patch does *not* increase reaction	Hjorth (30) Justice (20), Kligman (19) Bettley (31)
Pregnancy	Not known	
Diet	Not studied	

are unequivocal: The product causes adverse effects or it does not. The in-use test still has many important limitations, such as: number of subjects; their age and sex; period of use; geographical locale; and possible lack of sensitivity to misuse or deliberate abuse. These problems can sometimes be resolved by careful consideration of the product's intended use and the manufacturer's recommended directions. A major problem is the reliability of the test subjects who must be able to communicate with the investigator and must be depended on to use the product. This problem has been noted by Maibach and Epstein (12), who feel that such a test is frequently not a test at all. This is, of course, a matter of experimental design, but does not detract from the value of properly supervised and controlled in-use tests.

Clinical Experiences

So far, the discussion has been primarily in terms of the possible chance of error or possible problems in the interpretation of results. In order to illustrate these problems and to re-emphasize the need for carefully controlled in-use tests, two examples from the authors' experiences will be discussed.

Approximately six years ago a patch test study of a make-up product elicited no evidence of either primary irritation or of sensitization *via* the repeated insult technique. More recently, this composition was re-examined for primary irritation and sensitivity by the same technique at a different test locality. Surprisingly, the product now caused primary irritation and could not even be subjected to the repeated insult test because of the high incidence of irritation. Our approach to this riddle was a pragmatic one: A series of break-down products, in which one or more ingredients were deleted, was subjected to patch and sensitization tests. After much effort, the offending ingredient was identified, and the product was then reformulated to yield a new, hopefully safer composition.

Although a definitive explanation cannot be offered, several things could have taken place during the intervening five or six years: There may have been subtle changes in the raw material due to a different process of manufacture. The time of year during which the tests were conducted was different. The selection of test subjects could have played a role in causing the second test to show positive results, whereas six years ago the product was considered safe and ready for marketing. Last,

but not least, the change in testing facilities may have been the influencing factor. Interestingly enough, the incidence of complaints from consumers did not change during the intervening years.

What does this all mean? Are our tests too sophisticated or are the artificial test conditions too severe and thus prejudicial to the product? Of course, we do not know the answer, although Baer's (23) report on the changing sensitivity of the population may offer a clue. Regardless, this is a typical example of how the responses to chance patch testing can trigger research activities..

A more interesting example is a fairly recent experience using a translucent facial make-up product. This product was put through the battery of tests using established patch test methods and found to be free of irritation and sensitizing liability. However, in an in-use test, over 75% of the test panelists were unable to tolerate the test product. At first it was believed that the geographical location of the test site might have precipitated the responses. Alternately, exposure to natural sunlight could have caused the reactions. Accordingly, photopatch tests (on the back) were conducted with negative results. Another in-use test was initiated at another site, which confirmed the results of the first in-use test. In view of these adverse results, the product was reformulated in an attempt to eliminate the offending agent or agents. After much work, some minor constituents were eliminated from the formulation, and a third in-use test was conducted. Now, the revised product was tolerated by 100% of the test population. It was concluded that neither light nor occlusion triggered the reaction. In this case it seems evident that the location of the skin, i.e., face *vs.* back, was the primary factor in producing the adverse effects. If the initial patch test results had been used as the sole criteria in judging the product safe, the original product would have been marketed on the basis of established procedures and in good faith. The product would, nevertheless, have caused adverse reactions in many consumers. Based on this example, exaggerated in-use testing would seem to be an important part of any predictive skin testing program.

It should be emphasized that the two examples cited here are unusual. Normally, repetitive patch tests of the same product yield comparable results. Similarly, in-use tests normally confirm the safeness of products established in a patch test series. Our many years of experience indicate that a well-designed and well-conducted predictive testing program is an almost foolproof method of establishing the safety of products before they are marketed.

Conclusion

Potential problems of predictive testing of the safety of a cosmetic product have been pointed out. Particularly troublesome are the following: The selection of the test(s) and of the test population; inadvertent experimental errors; problems of interpretation, which are related primarily to the absence of well-documented scientific information. It is pointed out that in-use testing under clinical supervision is highly desirable and should be seriously considered as an adjunct to the battery of known skin predictive procedures. Two unusual examples from the authors' files illustrate these problems and support these inferences. In the authors' experience, the results of a sound predictive testing program are useful for anticipating the product's safety in the market place.

(Received December 21, 1970)

References

(1) Kligman, A. M., The identification of contact allergens by human assay. I. A critique of standard methods, *J. Invest. Dermatol.*, **47,** 369–74 (1966).
(2) Brunner, M. J., Pitfalls and problems in predictive testing, *J. Soc. Cosmet. Chem.*, **18,** 323–31 (1967).
(3) Idson, B., Topical toxicity and testing, *J. Pharm. Sci.*, **57,** 1–11 (1968).
(4) Kligman, A. M., The identification of contact allergens by human assay. III. The maximization test: a procedure for screening and rating contact sensitizers, *J. Invest. Dermatol.*, **47,** 393–409 (1966).
(5) Wooding, W. M., and Opdyke, D. L., A statistical approach to the evaluation of cutaneous responses to irritants, *J. Soc. Cosmet. Chem.*, **18,** 809–29 (1967).
(6) Barron, B. A., Variations on a theme—The evaluation of necessary and unnecessary risks, *Toxicol. Appl. Pharmacol., Suppl. 3,* 72–5 (1969).
(7) Opdyke, D. L., and Burnett, C. M., Practical problems in the evaluation of the safety of cosmetics, *Proc. Sci. Sect. Toilet Goods Ass.*, **44,** 3–4 (1965).
(8) Food and Drug Administration, *Appraisal of the Safety of Chemicals in Foods, Drugs, and Cosmetics,* Association of Food and Drug Officials of the U. S., Texas State Department of Health, Austin, Tex., 1959.
(9) Wolven, A., and Levenstein, I., Techniques for evaluating dermal irritation, *J. Soc. Cosmet. Chem.*, **18,** 199–203 (1967).
(10) Marzulli, F. N., *et al.*, Delayed contact hypersensitivity studies in man and animals, *Proc. Joint Conf. Cosmet. Sci.*, Washington, D. C., 107–22 (1968).
(11) Magnusson, B., and Kligman, A. M., The identification of contact allergens by animal assay. The guinea pig maximization test, *J. Invest. Dermatol.*, **52,** 268–76 (1969).
(12) Maibach, H. I., and Epstein, W. L., Predictive patch testing for allergic sensitization in man, *Toxicol. Appl. Pharmacol., Suppl. 2,* 39–43 (1965).
(13) Rubenkoenig, H. L., and Quisno, R. A., Use of the patch test in estimating hazards to the skin, *Proc. Sci. Sect. Toilet Goods Ass.*, **28,** 6–8 (1957).
(14) Magnusson, B., and Hersle, K., Patch test methods. I. A comparative study of six different types of patch tests, *Acta Dermato-Venereol.*, **45,** 123–8 (1965).

(15) Lanman, B. M., *et al.*, The role of human patch testing in a product development program, *Proc. Joint Conf. Cosmet. Sci.*, Washington, D. C., 135–45 (1968).
(16) Kligman, A. M., and Wooding, W. M., A method for the measurement and evaluation of irritants on human skin, *J. Invest. Dermatol.*, **49,** 78–94 (1967).
(17) Wolcott, G. L., A plea for in-use testing of cosmetics, *Drug Cosmet. Ind.*, **93,** 155–265 (1963).
(18) Rieger, M. M., and Battista, G. W., Some experiments in the safety testing of cosmetics, *J. Soc. Cosmet. Chem.*, **15,** 161–72 (1964).
(19) Kligman, A. M., Evaluation of cosmetics for irritancy, *Toxicol. Appl. Pharmacol., Suppl. 3,* 30–44 (1969).
(20) Justice, J. D., *et al.*, The correlation between animal and human tests in assessing product mildness, *Proc. Sci. Sect. Toilet Goods Ass.*, **35,** 12 (1961).
(21) Kligman, A. M., The identification of contact allergens by human assay. II. Factors influencing the induction and measurement of allergic contact dermatitis, *J. Invest. Dermatol.*, **47,** 375–92 (1966).
(22) Forsbeck, M., *et al.*, The frequency of allergic diseases among relatives to patients with allergic eczematous contact dermatitis, *Acta Dermato-Venereol.*, **46,** 149–52 (1966).
(23) Baer, R. L., *et al.*, Changing patterns of sensitivity to common contact allergens, *Arch. Dermatol.*, **89,** 3–8 (1964).
(24) Agrup, G., Sensitization induced by patch testing, *Brit. J. Dermatol.*, **80,** 631–4 (1968).
(25) Magnusson, B., and Hersle, K., Patch test methods. II. Regional variations of patch test responses, *Acta Dermato-Venereol.*, **45,** 257–61 (1965).
(26) Magnusson, B., and Hersle, K., Patch test methods. III. Influence of adhesive tape on test response, *Ibid.*, **46,** 275–8 (1966).
(27) Lowney, E. D., Attenuation of contact sensitization in man, *J. Invest. Dermatol.*, **50,** 244–9 (1968).
(28) Rebello, D. J. A., and Suskind, R. R., The effect of common contactants on cutaneous reactivity to sensitizers, *Ibid.*, **41,** 67–80 (1963).
(29) Sipos, K., Chemical hypersensitivity and dermatological diseases, *Dermatologica,* **135,** 421–32 (1967).
(30) Hjorth, N., Seasonal variations in contact dermatitis, *Acta Dermato-Venereol.*, **47,** 409–18 (1967).
(31) Bettley, F. R., and Grice, K. A., The effect of sweating on patch test reactions to soap, *Brit. J. Dermatol.*, **78,** 636–9 (1966).