

The association between size of test chamber and patch test reaction: a statistical reanalysis

OLAF GEFELLER¹, ANNETTE PFAHLBERG¹, JOHANNES GEIER², JOCHEN BRASCH³ AND WOLFGANG UTER⁴

¹Department of Medical Statistics, University of Göttingen, Germany;

²Information Network of Departments of Dermatology, University of Göttingen, Germany;

³Department of Dermatology, University of Kiel, Germany;

⁴Department of Dermatology, Environmental Medicine and Health Theory, University of Osnabrück, Germany

A recent study by Brasch and co-workers reported on the association between size of test chamber and patch test reaction. The investigators interpreted their data on 495 patients as having conclusively shown that standard preparations of fragrance mix, wool wax alcohols, Kathon CG and formaldehyde led to more positive test reactions when large Finn Chambers were used for patch testing. We have scrutinized the statistical aspects of this study and conclude that the authors should have adopted a statistical approach suitable to analyse dependent samples. After explaining the correct methodological way of dealing with quadratic contingency tables formed by 2 dependent samples, we reanalyze the data accordingly and compare the results to those of the original paper. Based on this reanalysis, the conclusions are more complex: the reaction pattern for the fragrance mix and wool wax alcohols is significantly different between small and large test chambers; however, this discrepancy arises primarily from changing weak positive reactions with small chambers to strong positive reactions with large chambers. For formaldehyde, no relationship between chamber size and patch test reaction was found in the data, while for Kathon CG, statistical evidence is borderline that more positive test reactions are yielded by large test chambers than by small ones.

Key words: dependent two-sample design; patch testing technique; reaction pattern; statistics; test chamber size; fragrance mix; wool wax alcohols; Kathon CG; formaldehyde. © Munksgaard, 1999.

Accepted for publication 6 August 1998

Valid conclusions from empirical data require valid statistical procedures for analyzing the data. The appropriate statistical technique has to be selected carefully to reflect precisely the research question to be answered and the study design employed to obtain the data. In a recent paper in this journal, Brasch et al. (1) considered the interesting topic in patch testing as to whether the size of the test chamber affects the test reaction. Based on a study comprising 495 patients who were synchronously tested with 4 standard allergens (fragrance mix, wool wax alcohols, Kathon CG, formaldehyde) with small (8 mm diameter) and large (12 mm diameter) test chambers, they compared the test results. They observed that “for each of the 4 allergens, a significantly higher number of stronger reactions was seen with the large chambers as compared to the small ones” (1, p. 118) and thus concluded “that large test chambers may be useful in the detection of weak sensitization to allergens

that are characterized by a low reproducibility of patch tests with small chambers” (1, p. 120).

We reconsider the same data, paying special attention to the statistical problems raised by the analysis of dependent samples. Particular emphasis is laid here on a detailed description of the methodological aspects of analysing such patch test data, which also apply to other empirical research in patch testing utilizing a similar design. For the data by Brasch et al. (1), reanalysis employing the correct statistical approach for dependent samples yields different conclusions with respect to the effect of the size of the chambers on the test reactions to those published in the original paper.

Materials and Methods

Details of study design, selection of patients, preparation of allergens for the 2 sizes of test chambers

and conduct of the multicentre study are given in the original paper (1). The outcome data for this reanalysis consisted of the reaction pairs from all 495 study subjects (1 component for each chamber size) evaluated 3 days after application of the allergen. Test reactions were grouped into 4 categories: negative (−), questionable or irritant (?/IR), positive (+) and strongly positive (+ +/+ + +).

The earlier data analysis by Brasch et al. (1) used exclusively the standard χ^2 -test for all statistical evaluations (1, p. 118). This well-known asymptotical (i.e., valid only for large samples) test procedure is appropriate for analyzing (a) the differences in the distribution of 1 categorical factor between 2 (or more) independent samples or (b) the deviation from independence of 2 (or more) possibly associated categorical factors in 1 sample. At first glance, it is a surprising feature of the χ^2 -test that it is a valid statistical test for these 2 totally different questions. The derivation of a reasonable test statistic for these 2 situations (a) and (b) leads, however, to algebraically equivalent variants of the χ^2 -test, as can be found in nearly all introductory texts on biostatistics (e.g., 2). Perhaps this perplexing property has led to some confusion among researchers with a non-statistical background, resulting in the impression that the χ^2 -test can be used for all purposes of categorical data analysis, a notion which is definitely not true.

In the present case, we wish to point out that, unlike the previous evaluation, the statistical analysis needs to focus on the discrepancies between the reaction patterns of 2 dependent samples. The dependency of the samples clearly results from the fact that the same patients were tested with small and large test chambers. For this situation, the standard χ^2 -test is inappropriate and will give misleading answers. Here, it could only address the question as to whether test reactions in small chambers are associated with those in large chambers. A significant association indicated by standard χ^2 -test is, however, meaningless, since it is the design of the study that forces the distribution of test reactions for small and large chambers to be highly associated, as the observations consist of pairs of test reactions to the same allergens seen in the same patients.

The correct approach for this study considers the complete 4×4 table showing the joint distribution of patch test results observed under the 2 conditions (small and large chambers) and uses techniques for dependent samples (Table 1).

Based on this quadratic contingency table, the evaluation of differences in the pattern of test reactions between small and large test chambers can be formalized statistically in different ways, depending on the precise research question to be

answered: (i) whether the (marginal) distribution of test reactions is identical for small and large chambers or not (statistically speaking, whether the table fulfils the property of marginal homogeneity); or (ii) whether discrepancies in the joint distribution of test reactions for small and large chambers vary unsystematically over all possible discordant assessments of the test reaction or not (statistically speaking, whether the table fulfils the property of symmetry or not). Only for the case of a dichotomous response variable (i.e., here the patch test reaction when categorized as “not positive” versus “positive”) are the questions (i) and (ii) equivalent, and is the statistical procedure for answering (i) and (ii) the classical McNemar test (3). As, in our case, the response variable has 4 categories, we have to distinguish them. There is a clear hierarchy between the 2 questions: if symmetry holds for a quadratic table, then this implies marginal homogeneity, but not vice versa. Consequently, if a statistical test of marginal homogeneity results in significant evidence ($p < 0.05$) that the dependent samples of test reactions from small and large chambers differ, then there are also automatically significant systematic discrepancies in the joint distribution of test reactions, and a specific statistical test of symmetry becomes obsolete.

For practical statistical testing of marginal homogeneity, a generalization of the Cochran-Mantel-Haenszel test statistic can be used (4). The symmetry can be analysed with the Bowker test (5). Both tests can be viewed as generalizations of the McNemar test; however, only the similarity between Bowker and McNemar test is directly evident in the test statistic (Table 2). In contrast to the McNemar test, where an exact test for small samples can easily be constructed, both generalizations exist only in an asymptotical form. For the Bowker test, the validity of this asymptotical procedure quickly becomes questionable for small samples, whereas the test for marginal homogeneity has better small sample properties. Table 2 summarizes the statistical procedures for analysing dependent samples with categorical response variables typically arising in patch test evaluations. Further statistical details on these methods and a discussion of alternative procedures (e.g., the application of loglinear models or a regression framework for a categorical response variable implementing the generalized estimating equation (GEE) approach) can be found in statistical textbooks (6, 7).

We reanalyzed the data presented by Brasch et al. (1) with the techniques described above, utilizing SAS software (8). Furthermore, we dichotomized the patch test reactions into the 2 response categories “not positive” (−/?/IR) and “positive”

Table 1. Layout of the quadratic contingency table showing joint and marginal distribution of patch test results for small and large test chambers required for statistical evaluation of differences

		Small chambers				Marginal distribution for large chambers
		negative (–)	questionable or irritant (?/IR)	positive (+)	strong positive (+ +/+ + +)	
Large chambers	negative (–)	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
	questionable or irritant (?/IR)	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	positive (+)	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
	strong positive (+ +/+ + +)	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
	Marginal distribution for small chambers	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

Table 2. Summary of properties of statistical tests for analysing dependent patch test data (explanations see text)

Statistical test	No. response categories	Research question to be answered	Formula	Asymptotic distribution	Exact distribution
McNemar	2	(i)/(ii)	$\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$	χ^2_1	binomial
generalized Cochran-Mantel-Haenszel	$k > 2$	(i)	Very complex, see (7,p. 234/5)	$\chi^2_{k(k-1)/2}$	not available
Bowker	$k > 2$	(ii)	$\sum_{i < j} \frac{(n_{ij}-n_{ji})^2}{n_{ij}+n_{ji}}$	$\chi^2_{k(k-1)/2}$	not available

(+ + +/+ + +) and applied the McNemar test to this collapsed fourfold table to analyze a difference in detecting sensitization (of any degree) to the allergens under the 2 conditions of chamber size. In addition, we also assessed the degree of agreement between the reaction patterns of small and large test chambers by estimating the weighted kappa coefficient (κ), the standard measure for chance-corrected agreement on an ordinal scale (9, 10) and its accompanying 95% confidence interval (CI). Such an estimation approach provides an attractive supplement to the earlier analysis focusing completely on statistical testing, as it yields an interpretable quantification of the similarities in reaction patterns that is more informative than mere p -values of statistical tests.

Results

The statistical results of the reanalysis are summarised in Table 3 and are compared to those reported by Brasch et al. (1, p. 119). Whereas the p -values of the standard χ^2 -test (1st column) have been interpreted as indicating significant differences in patch test reactions for small and large chambers for all 4 allergens (though in fact only expressing significant association), the reanalysis (differentiating between testing for marginal homogeneity and

symmetry in the 4×4 table shown in Table 1 and for a difference in the proportions of positive reactions in the correspondingly collapsed 2×2 table) revealed a more complex picture.

The overall agreement of test reactions between small and large chambers is, except for Kathon CG, high, as shown by the chance-corrected κ -values of 0.82 (wool wax alcohols), 0.79 (fragrance mix) and 0.72 (formaldehyde). Differences in patch test results between small and large Finn Chambers were only seen in small subsamples of the study population: for fragrance mix 44 (8.9%), for wool wax alcohols 23 (4.6%), for Kathon CG 27 (5.5%) and for formaldehyde 23 (4.6%) study subjects exhibited discordant pairs of test reactions under the 2 conditions.

Despite the small sample sizes of these subpopulations, the statistical evaluation of discrepancies in reaction patterns yielded significant differences between small and large test chambers for fragrance mix and wool wax alcohols when looking for marginal homogeneity and symmetry of corresponding reaction patterns. The p -values of the Bowker test analyzing symmetry of test reactions (3rd column in Table 3), however, have to be interpreted with caution, as the cell numbers in the 4×4 table are too small for an acceptable approximation of the test statistics to the χ^2 distribution.

Table 3. The relationship of patch test reactions in small and large test chambers for 4 standard allergens: statistical evaluations of reaction patterns in 495 subjects (GCMH=generalized Cochran-Mantel-Haenszel)

Allergen	Association (Brasch et al. (1)) χ^2 -test <i>p</i> -value	Marginal homogeneity (reanalysis) GCMH test <i>p</i> -value	Symmetry (reanalysis) Bowker test <i>p</i> -value	Difference in pos. reactions (reanalysis) McNemar test <i>p</i> -value	Agreement (reanalysis)	
					κ	(95%-CI)
fragrance mix	<0.0001	0.005	0.001	0.115	0.79	(0.73–0.85)
wool wax alcohols	<0.001	0.004	0.016	0.092	0.82	(0.75–0.89)
Kathon CG	<0.01	0.078	0.015	0.049	0.42	(0.32–0.52)
formaldehyde	<0.01	0.154	0.397	0.791	0.72	(0.61–0.83)

A deeper look at the discrepancies for these 2 allergens shows that they arise primarily from changing positive (“+”) reactions obtained in small Finn Chambers to strong positive (“++”/“+++”) reactions seen in the corresponding large Finn Chambers.

Statistical evidence for differences in patch test reactions under the 2 conditions is borderline for Kathon CG ($p=0.078$ for marginal homogeneity, $p=0.049$ for a difference in the proportions of positive reactions). The results suggest that here large test chambers may sometimes be helpful in the detection of sensitization to Kathon CG not found with small test chambers. For formaldehyde, there is no evidence for a relationship between size of chamber and patch test reaction.

Discussion

The primary objective of this reanalysis was to clarify the statistical aspects of analyzing patch test data obtained by synchronously observing test reactions in the same patients under 2 different conditions. The implications of a design statistically characterized by dependent samples of a categorical response variable are often overlooked in practical data analysis. In another study recently published in this journal, Trattner et al. (11) made a similar mistake when analyzing the effect of different formaldehyde concentrations (1% versus 2%) on patch test reactions. Investigators planning future studies of this kind are encouraged to pay more attention to the statistical problems of the analysis along the lines delineated in this paper.

Reanalysis of the data of Brasch et al. (1) with statistical methods appropriated for dependent samples yielded a more complex picture of results than the earlier publication described. Instead of confirming the conclusion that larger test chambers lead to more positive reactions for all 4 allergens, we now have to distinguish between the different allergens in our conclusions. For formaldehyde, no relationship between chamber size and patch test reaction was found, whereas for Ka-

thon CG statistical evidence is borderline that more positive test reactions are yielded by large test chambers than by small ones. The reaction pattern for fragrance mix and wool wax alcohols is significantly different between small and large test chambers; however, the discrepancy arises primarily from changing positive reactions with small chambers to strong positive reactions with large chambers. Overall, the agreement between test reactions with small and large test chambers is high. Consequently, the subsamples of the study populations with different reactions under the 2 conditions become rather small and do not allow a definite conclusion for all 4 allergens. A future study on this topic would need to recruit a larger patient population to increase the statistically relevant subsample of study subjects showing different test reactions for small and large test chambers. In such a design, yielding pairs of 2 dependent observations of a categorical response variable, the number of discordant pairs is statistically much more important than the total sample size.

References

1. Brasch B, Szliska C, Grabbe J. More positive patch test reaction with larger test chambers? *Contact Dermatitis* 1997; 37: 118–120.
2. Bland M. *An introduction to medical statistics*, 2nd edition. Oxford: Oxford University Press: 1995.
3. McNemar Q. Note on sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; 12: 153–157.
4. Mantel N, Byar D P. Marginal homogeneity, symmetry and independence. *Comm Statist* 1978; A7: 953–976.
5. Bowker A H. A test for symmetry in contingency tables. *J Amer Statist Assoc* 1948; 43: 572–574.
6. Bishop Y V V, Fienberg S E, Holland P W. *Discrete multivariate analysis*. Cambridge, MA: MIT Press: 1975.
7. Agresti A. *Categorical data analysis*. New York: John Wiley & Sons Inc, 1990.
8. Stokes M E, Davis C S, Koch G C. *Categorical data analysis using the SAS® system*. Cary NC: SAS Institute Inc: 1995.
9. Spitzer R L, Cohen J, Fleiss J L, Endicott J. Quantification of agreement in psychiatric diagnosis. *Arch Gen Psychiatry* 1967; 17: 83–87.

10. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–220.
11. Trattner A, Johansen J D, Menné T. Formaldehyde concentration in diagnostic patch testing: comparison of 1% with 2%. *Contact Dermatitis* 1998; 38: 9–13.

Address:

*Olaf Gefeller
Department of Medical Statistics
Georg-August-Universität
Humboldtallee 32
D-37073 Göttingen
Germany*