

Letter to the Editor

Reproducibility of patch tests

Dear Sir,

In a recent article, Schiessl et al. (1) address the important question of reproducibility of patch test readings in terms of concordance between synchronous testing of the same allergens from 2 different manufacturers. On the basis of their clinical study, the authors conclude that patch test materials from these 2 manufacturers essentially yield the same results. While most clinicians and researchers will be only too glad to read this, we have to point out that severe methodological problems cast doubt on the validity of these conclusions:

- (1) Although the authors report on a large number of single patch tests, the study comprised only 71 patients for the analysis. This small sample size renders its results imprecise. For instance, an estimated proportion of 5% (3 discordant of 60 positive reactions) has an exact 95% confidence interval (CI) of 1.0–13.9%. This puts the notion of a seemingly reliable performance in perspective, especially in relation to other, larger studies discussed.
- (2) Moreover, information such as these 5%, or other proportions of concordantly negative and positive results or the 4 literature-based standardization factors (2) are not the adequate quantitative components of information required for the evaluation of agreement between the 2 tests. Some decades ago, it was recognized that in such situations due to chance alone, a certain degree of concordance has to be expected, the magnitude of which depends on the specific structural framework of the data. Thus, concordance beyond chance, most commonly quantified using Cohen's kappa coefficient (3), is the most meaningful concept in this situation (4). We re-analysed data on selected allergens from (1) using Cohen's kappa (κ) with accompanying 95% CIs, utilizing a method for deriving exact bootstrap CIs which has recently been suggested (5). For cobalt chloride and 2-bromo-2-nitropropane-1,3-diol, for instance, the bootstrap kappa value is 0.62, and concordance can be regarded as substan-

tial (6) at first glance. However, a look at the precision of this estimate (95% CI: 0.00–1.00) shows again the very limited power of this study, compared, for example, to the previous multicentre study quoted (4): the CI includes everything from 'no agreement beyond chance' ($\kappa = 0$) to 'perfect agreement' ($\kappa = 1$), which renders the result non-informative. In fact, the study is so small that for many allergens (11 of the 30 studied in this investigation) no positive reactions at all were observed, and κ cannot be calculated.

In those allergens where perfect agreement was observed, and hence $\kappa = 1$ (like nickel sulfate), previous evidence, namely, the suitably sized study already quoted (2, 4) points to the fact that, by virtue of the small sample size, chance may have led to such an extremely favourable pattern. Under the assumption that concordance with synchronous testing of the same allergen of the same manufacturer (2, 4) will be at least as good as in the present situation, one could extrapolate the concordance found in that study, e.g. for nickel sulfate ($\kappa = 0.86$, 95% CI: 0.83–0.90) (4) to correspond to 12 concordant positive and 56 concordant negative and 3 discordant reactions, respectively, in the study by Schiessl, yielding a probably overoptimistic $\kappa = 0.86$, 95% CI: 0.68–1.0, i.e. a moderately inflated precision estimate. Clearly, this extrapolation is a speculative approach and not a valid substitute for a proper study.

- (3) The observation of manufacturer-associated differences in the grading of positive reactions is interesting; in a suitably powered study, it would be worthwhile to verify possible systematic differences by appropriate statistical testing (7).
- (4) 14 of 85 patients initially enrolled were later excluded due to 'angry back' reactions or questionable results (details of the definitions are not given). While the application of strict criteria in research studies is often well justified, the exclusion of a considerable proportion of patients impairs external validity, i.e. the very transfer of study results to clinical routine the authors suggest.

In conclusion, we agree with the authors that addressing reproducibility of patch test results constitutes an important issue. It will be extremely relevant for multicentre studies incorporating test results obtained with materials from different suppliers to elucidate further the degree of methodological variation introduced by this factor. The results presented can, however, only be regarded as preliminary and have to be confirmed in a larger study to be accepted as valid.

Yours sincerely,

Wolfgang Uter, Annette Pfahlberg,
Olaf Gefeller

Department of Medical Informatics, Biometry
and Epidemiology,
University of Erlangen – Nürnberg,
Erlangen, Germany

References

1. Schiessl C, Wolber C, Strohal R. Reproducibility of patch tests: comparison of identical test allergens from different commercial sources. *Contact Dermatitis* 2004; 50: 27–30.
2. Brasch J, Henseler T, Aberer W et al. Reproducibility of patch tests. *J Am Acad Dermatol* 1994; 31: 584–591.
3. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37–46.
4. Uter W, Pfahlberg A, Brasch J. Zur Reproduzierbarkeit der Epikutantestung – die Bewertung der Konkordanz bei synchroner Applikation. *Allergologie* 2002; 25: 415–419.
5. Klar N, Lipsitz S R, Parzen M, Leong T. An exact bootstrap confidence interval for κ in small samples. *Statistician* 2002; 51: 467–478.
6. Landis J R, Koch G G. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
7. Gefeller O, Pfahlberg A, Geier J, Brasch J, Uter W. The association between size of test chamber and patch test reaction: a statistical reanalysis. *Contact Dermatitis* 1999; 40: 14–18.

Address:

Wolfgang Uter
Department of Medical Informatics, Biometry and
Epidemiology
University of Erlangen – Nürnberg
Waldstr. 6
91054 Erlangen
Germany
Tel: +49 9131 8522750
Fax: +49 9131 8522721
e-mail: wolfgang.uter@rzmail.uni-erlangen.de