

Comparing reliabilities of strip and conventional patch testing

Heinrich Dickel¹ , Johannes Geier², Burkhard Kreft³, Wolfgang Pfützner⁴ and Oliver Kuss⁵

¹Department of Dermatology, Venereology and Allergology, Ruhr University Bochum, 44791 Bochum, Germany, ²Information Network of Departments of Dermatology (IVDK), Georg August University Göttingen, 37075 Göttingen, Germany, ³Department of Dermatology and Venereology, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany, ⁴Department of Dermatology and Allergology, Philipps University Marburg, 35043 Marburg, Germany, and ⁵Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), 40225 Düsseldorf, Germany

doi:10.1111/cod.12758

Summary

Background. The standardized protocol for performing the strip patch test has proven to be valid, but evidence on its reliability is still missing.

Objective. To estimate the parallel-test reliability of the strip patch test as compared with the conventional patch test.

Methods. In this multicentre, prospective, randomized, investigator-blinded reliability study, 132 subjects were enrolled. Simultaneous duplicate strip and conventional patch tests were performed with the Finn Chambers® on Scanpor® tape test system and the patch test preparations nickel sulfate 5% pet., potassium dichromate 0.5% pet., and lanolin alcohol 30% pet. Reliability was estimated by the use of Cohen's kappa coefficient.

Results. Parallel-test reliability values of the three standard patch test preparations turned out to be acceptable, with slight advantages for the strip patch test. The differences in reliability were 9% (95%CI: –8% to 26%) for nickel sulfate and 23% (95%CI: –16% to 63%) for potassium dichromate, both favouring the strip patch test.

Conclusions. The standardized strip patch test method for the detection of allergic contact sensitization in patients with suspected allergic contact dermatitis is reliable. Its application in routine clinical practice can be recommended, especially if the conventional patch test result is presumably false negative.

Key words: accuracy; contact dermatitis; kappa coefficient; nickel sulfate; patch test; potassium dichromate; reliability of results; strip patch test.

The strip patch test is a valuable diagnostic option for identifying allergic contact sensitization, especially if false-negative reactions are suspected in a conventional patch test (1). A standardized protocol for performing the strip patch test was introduced on behalf of the German Contact Dermatitis Research Group (DKG) in 2004 (2), and we subsequently verified its validity by use of the standard patch test preparations nickel sulfate,

potassium dichromate, and lanolin alcohol (3). However, validity does not imply reliability. To date, evidence on the reliability of the standardized strip patch test has been lacking.

'Reliability' has many aspects, especially in patch testing, and a number of different terms and definitions have been offered. For this study, we follow the definition of Kraemer (4), with reliability being '(...) the correlation of two (...) independent diagnoses per subject in the population of interest over a span of time during which each individual patient is unlikely to change disorder status'. The aspect of the individual disease status not changing during assessment of reliability is especially relevant for patch testing, because it puts into perspective the value of previous reliability studies in which the first and second tests were months or sometimes even years apart in time

Correspondence: Heinrich Dickel, Department of Dermatology, Venereology and Allergology, St Josef-Hospital, Ruhr University Bochum, Gudrunstrasse 56, 44791 Bochum, Germany. Tel: +49 234 509 3453; Fax: +49 234 509 3451. E-mail: h.dickel@klinikum-bochum.de

Conflict of interest: The authors declare no conflict of interests.

Accepted for publication 15 December 2016

(5, 6). In the following, we used the term 'parallel-test reliability' to refer to our study design of simultaneous duplicate patch testing.

In this multicentre, prospective, randomized, investigator-blinded study, we assessed the parallel-test reliability of the strip patch test, and compared it with the parallel-test reliability of the conventional patch test. For this purpose, simultaneous duplicate patch tests were performed on recalled subjects of the collective of the previous validation study, with the same methodology (3). The values of the reliability estimates were calculated by the use of Cohen's kappa coefficient (7, 8). We could not identify any study within the primary literature dealing with the parallel-test reliability of conventional patch testing that had used a design identical to that of our study. In the following, we avoid the more general term reproducibility (9).

Subjects and Methods

This reliability study took place from 5 December 2013 to 14 April 2015. During this time, former patients of the validation study (3), which had taken place from 4 December 2006 to 31 March 2009, were consecutively recalled and screened in three of the initial seven different patch test units of tertiary-care university-affiliated dermatological clinics. Non-participation of the patch test units in Berlin, Jena, Kiel and Lübeck was attributable to lack of resources following structural or personnel changes since the validation study. The study was first approved by the ethics committee of the Medical Faculty of the Ruhr University Bochum (registration no. 4770-13; 4 September 2013). The study protocol was then subjected to review and approval by the local ethics committees of the different participating dermatological clinics. All procedures were explained in detail to all subjects, who signed informed consent forms prior to participation.

Subjects

The inclusion criterion was participation in the previous validation study (3). In total, 719 former patients were first contacted by telephone (Table 1). Those who could not be contacted by telephone received a written invitation to participate. In cases of no or delayed feedback, the remaining patients were contacted by telephone for a second and final time. Eventually, 132 subjects (overall response rate of 18.4%) were also willing to participate in the reliability study. Reasons for not participating were, in decreasing order, migration, no reply, no interest, job-related inability to participate, death, and corticosteroid medication. The exclusion criteria were pregnancy,

Table 1. Patients of the previous validation study (3) and subjects of the present reliability study stratified into patch test units with respective response rates

	Validation study December 2006 to March 2009		Reliability study December 2013 to April 2015		Response rate
	n	%	n	%	
Bochum	588	81.8	101	76.5	17.2
Halle (Saale)	109	15.2	28	21.2	25.7
Marburg	22	3.0	3	2.3	13.6
<i>Total</i>	<i>719</i>	<i>100</i>	<i>132</i>	<i>100</i>	<i>18.4</i>

lactation, florid eczema, intake of immunosuppressive or immune-modulating drugs such as corticosteroids or cyclosporine, topical use of corticosteroids on the back, and intense ultraviolet exposure on the back within the 4 weeks prior to study enrolment. The discontinuation criteria for individual subjects were severe skin reactions or intense discomfort in the patch test areas. All of these criteria were not met in any case.

Strip patch test and conventional patch test

Strip patch testing was performed according to the standardized protocol described in detail elsewhere (2, 10). Tape stripping was carried out with a 3M™ Blendederm™ surgical tape (3M™ Medica, 3M Deutschland, Neuss, Germany; dimensions 2.5 cm × 4.57 m; CE no. 0493-BSI 0086-class I; batch no. 201808, manufacture date August 2013, expiry date August 2018). Body hair was removed with a 3M™ Clipper before tape stripping. Conventional patch testing was performed according to DKG guidelines (11).

The test sites were on normal back skin that showed no signs of dermatitis. Small Finn Chambers® on Scanpor® tape (Epitest, Tuusula, Finland; inner diameter 8 mm, outer diameter 11 mm, capacity 20 µl) were used. Subjects were tested with the patch test preparations nickel sulfate 5% pet., potassium dichromate 0.5% pet., and lanolin alcohol 30% pet., purchased from Almirall Hermal (Reinbek, Germany) until 2013, and from SmartPractice Europe (Reinbek, Germany) starting in 2014. Additionally, the vehicle white pet. and one blank Finn Chamber® were tested as negative controls (Fig. 1). No other patch test preparations were applied.

The application time of the patch tests was uniformly 1 day (24 h). The readings were routinely performed on day (D) 1, D3, and D7. In a telephone interview on D21, the subjects were finally asked about persistent

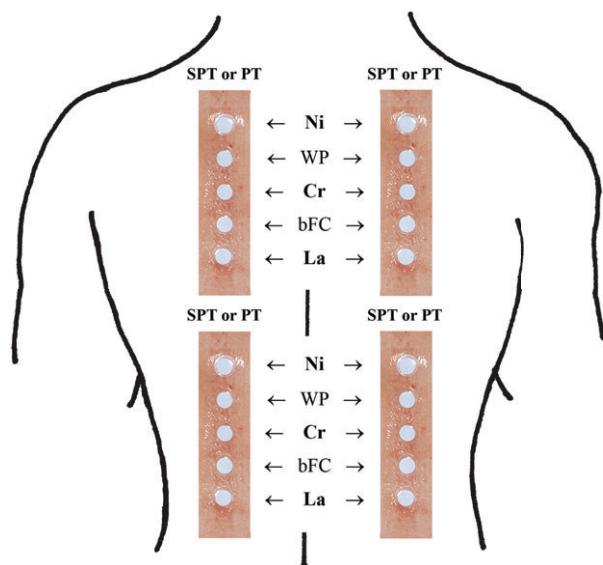


Fig. 1. Positions of the patch test preparations and patch tests; the positions of the patch tests were randomized. bFC, blank Finn Chamber®; Cr, potassium dichromate 0.5% pet.; La, lanolin alcohol 30% pet.; Ni, nickel sulfate 5% pet.; PT, patch test; SPT, strip patch test; WP, white pet.

and/or new reactions on all test sites. Only in the case of an affirmative answer was the subject called in for an additional reading. No new positive reactions were observed to any of the patch test preparations on D21. Documentation of the D7 test reactions was missing for 3 subjects (2.3%) and documentation of the D21 test reactions for 10 subjects (7.6%).

Patch test reactions were graded by use of the DKG scoring system (12), and dichotomized for statistical analysis. All +, ++ and +++ positive results were summarized as positive responses, whereas doubtful (?+) reactions, follicular (f) reactions, irritant reactions (IR) and negative (–) reactions were counted as negative responses. Positive and negative responses are reported on the basis of the reading utilized to assess reliability on D3.

Randomization and blinding

Strip patch tests and conventional patch tests were applied, each in duplicate, to the left and right sides of the upper and lower back of the subject, with two randomly chosen sites being tape-stripped for the strip patch test. Thereby, four test sites were investigated (two for the strip patch test, and two for the conventional patch test), and the allocation of patch tests to the sites was completely randomized (Fig. 1). Randomization was conducted according to an individual randomization list, which was

sent to each of the patch test units at the beginning of the study and monitored at the end of the study (by O.K.).

In each of the three patch test units, all patch tests were read and scored by the same specially trained and experienced dermatologist (H.D., B.K., or W.P.) who had already been involved in the previous validation study (3). Each reading dermatologist was blinded with respect to those sites of the subject's back that were tape-stripped. Unblinding was normally performed after the last reading (D7, $n = 129$, 97.7%; in three cases on D3, 2.3%) for clinical interpretation of test results. Owing to randomization of all duplicate patch tests applied to the four test sites, all readings were blinded, and thus all reliability assessments were also blinded. It is of note that even better blinding could have been achieved by also randomizing the patch test preparations within the Finn Chambers®. However, we deliberately avoided this, because of a perceived high risk of mixing up test preparations when filling the test chambers.

Data management and statistics

The three participating dermatological clinics are active members of the Information Network of Departments of Dermatology (IVDK) (<http://www.ivdk.org/en/>), a public health surveillance system dedicated to the epidemiology of contact allergy (13), where data management of the study was located. Each patch test unit sent study data anonymously to the IVDK for centralized evaluation (performed by J.G.).

For descriptive purposes, we report categorical variables as absolute (relative) frequencies, and continuous variable as quantiles [minimum, 25th percentile (Q1), median, 75th percentile (Q3), maximum]. As our response is binary (positive versus negative responses), we used Cohen's kappa coefficient to measure reliability. Differences in kappa coefficients as the primary outcomes to measure differences in reliability between the strip patch test and the conventional patch test were assessed with the model of Barlow (14). Because the four patch test results within 1 subject have to be considered as dependent, we initially planned to adjust for this dependency by extending the Barlow model with a random subject effect. However, during analysis, these models turned out to be numerically unstable, most probably because of the small absolute number of positive patch test results, so we ignored this dependency in all analyses. It should be noted that this leads to a more conservative analysis. All estimates are given with their 95% confidence intervals (CIs). CIs for the positivity ratio (PR) (15) are computed as exact intervals for binomial proportions, and those for the reaction index (RI) (16) are computed as previously proposed by our group (17).

Data were managed and analysed with the statistical analysis software SAS®, version 9.4 (SAS Institute, Cary, NC, USA). This report is written in close compliance with the proposed Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (18) and with the Guidelines for the descriptive presentation and statistical analysis of contact allergy data (8).

Literature search

A literature search for published patch test studies on parallel-test reliability was carried out with PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>; last accessed August 2016) and by a hand search of three potentially relevant German journals not listed in PubMed: *Allergo Journal*; *Allergologie*; and *Dermatologie in Beruf und Umwelt*. The inclusion criteria for consideration of the literature were as follows: (i) use of the conventional or strip patch test method; (ii) use of the Finn Chambers® on Scanpor® test system; (iii) use of identical patch test preparations from the same commercial source; (iv) use of allergens, except for atopy patch test preparations; (v) use of identical occlusion times (24 or 48 h); and (vi) (re-)evaluation of all positive reactions. According to these criteria, we could identify only nine primary patch test studies reporting parallel-test reliability (5, 6, 19–25). Extending the inclusion criteria in terms of demanding sufficient information to be able to compute Cohen's kappa coefficients for the patch test preparations used finally led to four parallel-test reliability studies (6, 20–22) as a basis for comparison.

Results

One hundred and thirty-two subjects were enrolled in the study. There were no dropouts. No adverse events resulting from performing the patch tests were reported. Hence, all of the subjects were included in further data analysis. Table 2 shows the demographic and clinical characteristics of the subjects. Their characteristics were similar to those of the 787 patients in the previous validation study (3). The time that had elapsed between the two studies ranged from 5.2 to 8.0 years. The median follow-up period was 6.2 years (Q1, 5.8 years; Q3, 6.8 years). Test sites of the strip patch test on the left and right side of the upper and lower back, respectively, were balanced (approximately one-quarter each). In strip patch tests, appearance of the stratum lucidum required a median number of 27 tape strips. After multiplication by the tape-specific correction factor ($cf = 11/26$) (2), the 'median' subject needed 11 tape strips prior to strip patch testing.

Table 2. Demographic and clinical subject characteristics (N = 132, all white-skinned)

Characteristic	Value
Age (years), median (minimum–maximum, Q1, Q3)	56.5 (23–84, 46.5, 64)
Sex, n (%)	
Female	93 (70.5)
Male	39 (29.5)
Fitzpatrick skin phototype, n (%)	
I	2 (1.5)
II	83 (62.9)
III	44 (33.3)
IV	3 (2.3)
Tape-stripped sites for SPT1 and SPT2, n (%)	
Left upper back: SPT1, SPT2	35 (26.5), 35 (26.5)
Left lower back: SPT1, SPT2	33 (25.1), 27 (20.5)
Right upper back: SPT1, SPT2	32 (24.2), 38 (28.8)
Right lower back: SPT1, SPT2	32 (24.2), 32 (24.2)
Tape strips down to the SL, median (minimum–maximum, Q1, Q3)	27 (7–80, 19, 45)
Tape strips for SPT1 and SPT2, median (minimum–maximum, Q1, Q3)	
Calculated ^a	11 (3–34, 8, 19)
Actually carried out	11 (3–34, 8, 19)

Q1, 25th percentile; Q3, 75th percentile; SL, stratum lucidum; SPT, strip patch test (in duplicate).

^aStrips down to the SL $\times 11/26$.

Reaction profiles of the patch test preparations

On the basis of test reactions on D3, Table 3 shows the reaction profiles for nickel sulfate, potassium dichromate and lanolin alcohol applied conventionally and on tape-stripped skin according to the RI and the PR. Overall, the reaction profiles of these test preparations remained largely unaffected by tape stripping of the skin prior to patch testing, as in the previous study (1). In 1 subject, IRs to white pet. and the blank Finn Chamber® were observed on D3 in one of the tape-stripped test sites. In another single subject, a positive reaction (+) to white pet. was observed on D7 in one non-tape-stripped test site. Even though it cannot be proven, this was most likely a mistake in documentation.

Reliability

When we estimated the reliability of simultaneous duplicate patch tests by use of the standard patch test preparations nickel sulfate, potassium dichromate, and lanolin alcohol, all Cohen's kappa values were >0.8 in strip patch testing and between 0.6 and 0.8 in conventional patch testing (Table 4). This resulted in reliability differences of 9 percentage points (95%CI: –8% to 26%) for nickel sulfate and 23 percentage points for potassium dichromate (95%CI: –16% to 63%), favouring the strip patch test for

Table 3. Reactions to patch test preparations at the day 3 reading (N tested = 132)

Patch test preparation	Patch test method	Reaction							IR	RI (95%CI)	PR (%) (95%CI)
		–	?+	f	+	++	+++				
Ni	SPT1	100	2	3	6	16	5	0	0.69 (0.44–0.94)	22 (9–42)	
	SPT2	97	2	5	10	11	5	2	0.49 (0.20–0.78)	38 (20–59)	
	PT1	101	3	4	12	7	5	0	0.55 (0.25–0.84)	50 (29–71)	
	PT2	103	5	0	11	10	3	0	0.66 (0.38–0.93)	46 (26–67)	
Cr	SPT1	125	2	0	3	2	0	0	0.43 (–0.24 to 1.00*)	60 (15–95)	
	SPT2	125	2	0	2	2	0	1	0.14 (–0.59 to 0.88)	50 (7–93)	
	PT1	123	1	0	7	1	0	0	0.78 (0.37–1.00*)	88 (47–100)	
	PT2	125	3	0	2	2	0	0	0.14 (–0.59 to 0.88)	50 (7–93)	
La	SPT1	128	0	0	1	2	0	1	0.50 (–0.35 to 1.00*)	33 (1–91)	
	SPT2	127	1	0	1	2	0	1	0.20 (–0.66 to 1.00*)	33 (1–91)	
	PT1	130	0	0	1	1	0	0	1.00 (ND)	50 (1–99)	
	PT2	130	1	0	1	0	0	0	0.00 (–1.00* to 1.00*)	100 (5–100)	

CI, confidence interval; Cr, potassium dichromate 0.5% pet.; f, follicular reaction (considered to be doubtful); IR, irritant reaction; La, lanolin alcohol 30% pet.; ND, not defined; Ni, nickel sulfate 5% pet.; PR, positivity ratio; PT, conventional patch test in duplicate (PT1, PT2); RI, reaction index; SPT, strip patch test in duplicate (SPT1, SPT2); ?+, doubtful reaction; *, border of the confidence interval was truncated at the minimal or maximal value of the respective scale.

both preparations as a trend. The number of subjects sensitized to lanolin alcohol was too low for coherent reporting of the difference between the estimates of reliability.

Table 5 shows four parallel-test reliability studies of conventional patch testing, which, however, were non-randomized and non-blinded, in contrast to our study. With the exception of the two studies of Gollhausen et al. each with a small sample size, Cohen's kappa values ranged between 0.72 for lanolin alcohol and 0.97 for nickel sulfate, with potassium dichromate in between. These values were consistently greater than the values for conventional patch testing in this study (Table 4). This can be expected in non-blinded studies, where effects are regularly found to be larger than those found in blinded medical studies (27, 28).

Discussion

This study is unique in presenting data on parallel-test reliability not only for the standardized strip patch test (2) but also for the conventional patch test in a multicentre, prospective, randomized, investigator-blinded setting with the use of Cohen's kappa statistic, which is the appropriate method for assessing the reliability of categorical patch test results (7, 8). It cannot be overemphasized that, in patch test studies on reliability, it should nowadays be the gold standard to blind investigators, rigorously apply proper statistical tests and analyse all subjects (29). However, to date, among the patch test studies on parallel-test reliability performed with allergens other than atopy patch test preparations, we could identify only two studies (5, 30) that excluded a reading

bias by means of a randomized blind protocol. Only one methodological study by Uter et al. (26) reported Cohen's kappa values, by reanalysing the data of the parallel-test reliability study by Brasch et al. (20).

A patch test is said to have a high reliability if it produces the same results under consistent conditions. Noteworthy, diagnostic patch testing is a biological assay with inherent pitfalls, that is, variability of results (31). Parallel-test reliability is used to assess the consistency of results of patch tests performed in duplicate on the same patient at the same time. This approach assumes that the technique of patch testing as applied is highly standardized and that both test sides are equivalent, but this is not invariably the case. Possible explanations for discordant patch test reactions are most likely methodological problems, such as differences in the amount of test preparation applied to the Finn Chamber® and unequal adherence of the Scanpor® tape to the skin (24, 32). Differences in skin penetration between the test sites may also contribute to discordant patch test results (33). Parallel-test reliability in our study was quite good, with slight advantages for the strip patch test, but is still not ideal for both patch test methods (Table 4). Only with lanolin alcohol were no discordant reactions observed in strip patch testing. Whether the higher Cohen's kappa values favouring the strip patch test are attributable to stratum corneum reduction (34) or changes in epidermal cell immunology caused by tape stripping of the skin (35) remains a subject of speculation at this point. With reference to the 95% CIs of the reliability differences between the strip patch test and the conventional patch test, we can, however, exclude large advantages of the conventional patch test. However,

Table 4. Parallel-test reliability of the strip patch test and conventional patch test by simultaneous duplicate testing based on reactions on day 3 (N = 132)

Patch test preparation	SPT1 and SPT2					PT1 and PT2					Difference in Cohen's kappa coefficients	
	+/+ n (%)	-/- n (%)	+/- n (%)	-/+ n (%)	Cohen's kappa coefficient (95%CI)	+/+ n (%)	-/- n (%)	+/- n (%)	-/+ n (%)	Cohen's kappa coefficient (95%CI)	SPT-PT (95%CI)	p-Value
Ni	24 (18.2)	103 (78)	3 (2.3)	2 (1.5)	0.88 (0.78–0.98)	20 (15.2)	104 (78.8)	4 (3)	4 (3)	0.80 (0.66–0.93)	0.09 (–0.08 to 0.26)	0.32
Cr	4 (3)	127 (96.2)	1 (0.8)	0 (0)	0.89 (0.66–1.0)	4 (3)	124 (94)	4 (3)	0 (0)	0.65 (0.34–0.97)	0.23 (–0.16 to 0.63)	0.24
La	3 (2.3)	129 (97.7)	0 (0)	0 (0)	1.0 (NE)	1 (0.8)	130 (98.4)	1 (0.8)	0 (0)	0.66 (0.04–1.0)	No convergence	NC

CI, confidence interval; Cr, potassium dichromate 0.5% pet.; La, lanolin alcohol 30% pet.; NC, not computed; NE, non-existent; Ni, nickel sulfate 5% pet.; PT, conventional patch test in duplicate (PT1, PT2); SPT, strip patch test in duplicate (SPT1, SPT2); +/+, positive (+, ++, +++) on both test sites; -/-, negative (-, --, ---) on both test sites; +/-, follicular reaction, irritant reaction, -) on both test sites; +/+, -/+, positive on one test site and negative on the other.

Table 5. Published parallel-test reliability studies of conventional patch testing with the Finn Chambers® on Scanpor® test system and computable kappa coefficients for nickel sulfate 5% pet., potassium dichromate 0.5% pet., or lanolin alcohol 30% pet.

Study	Year of publication	Study design	Number of individuals evaluated	Time interval between the tests	Cohen's kappa coefficient, subsequently calculated (95%CI)				Comments
					Ni	Cr	La		
Brasch et al. (20)	1994	Prospective multicentre study	1285	Simultaneous duplicate testing	0.86 (0.83–0.90)	0.80 (0.74–0.86)	0.72 (0.61–0.84)	No randomization; no blinding; Cohen's kappa calculated in a reanalysis (26)	
Breit and Agathos (6)	1992	Prospective single-centre study	121	Simultaneous duplicate testing	0.97 (0.90–1.0)	0.84 (0.67–1.0)	0.89 (0.74–1.0)	No randomization; no blinding; Cohen's kappa not calculated	
Gollhausen et al. (22)	1989	Prospective single-centre study	35	Simultaneous duplicate testing	0.64 (0.19–1.0)	NC (no positive reactions)	NC (no positive reactions)	No randomization; no blinding; Cohen's kappa not calculated	
Gollhausen et al. (21)	1989	Prospective single-centre study	76	Simultaneous duplicate testing	0.78 (0.54–1.0)	0.66 (0.04–1.0)	NA (not tested)	No randomization; no blinding; Cohen's kappa not calculated	

CI, confidence interval; Cr, potassium dichromate 0.5% pet.; La, lanolin alcohol 30% pet.; NA, not applicable; NC, not computed; Ni, nickel sulfate 5% pet.

reducing the stratum corneum (34) as an alternative to increasing the dose/area of the hapten in a conventional patch test, and inducing danger signals (35) by tape stripping the skin prior to patch testing, might theoretically result in a higher risk of iatrogenic sensitization. So far, our experience with standardized strip patch testing according to the extensively evaluated protocol (2) shows that this risk does not appear to be considerably increased as compared with the generally low risk of conventional patch testing (36). Although it was not observed in the present limited study sample, the 95%CI for no iatrogenic sensitization prevalence in our case amounts to 0–2.3% (37). However, in our previous validation study (3), we also found no iatrogenic sensitization and, owing to the larger sample size, the 95%CI in that study was 0–0.4% (37), that is, considerably smaller and almost completely ruling out a clinically relevant prevalence of iatrogenic sensitization.

In addition to these above-mentioned general aspects, discordance of patch test reactions is allergen-dependent (19). Discordant patch test reactions are particularly seen with allergens that are well-known 'borderline irritants' or have 'weak allergenic potency'. Thus, the inherent marginal irritancy of potassium dichromate (19, 31) and the weak potency of lanolin alcohol (1, 33) probably account for the discordant patch testing with these test preparations. In other words, allergens with a high RI (i.e. low relative proportions of questionable, follicular or irritant reactions) and a low PR (i.e. a low percentage of weak positive reactions among all positive reactions) usually have better patch test reliability (30). Nickel sulfate, for example, is said to be one of the baseline allergens with the highest frequency of concordant patch test reactions (25). In view of these allergen-related differences, it is advisable not to refer to patch test reliability in general, but always to a particular preparation of allergen and test system (30, 38).

Finally, we acknowledge a potential limitation of this parallel-test reliability study. The sample size was rather

small, resulting in large 95%CIs of Cohen's kappa values and their differences, thus hampering the detection of significant differences. However, statistical significance means little when too much potential bias exists in the results (39), and our results with conventional patch testing are largely in accordance with the few comparable parallel-test reliability studies.

Conclusions

First, according to our results, the standardized strip patch test protocol (2, 10) can be recommended for application in routine clinical practice beyond conventional patch testing. Its results are not only valid (3) but also reliable.

Second, we confirm that patch test reliability always needs to be assessed separately for each distinct preparation of allergen and test system (30). In our study, both the strip patch test and the conventional patch test showed acceptable reliability for the patch test preparations nickel sulfate 5% pet., potassium dichromate 0.5% pet., and lanolin alcohol 30% pet., notwithstanding some imprecision of the Cohen's kappa estimates resulting from the small sample size.

Third, we encourage researchers to adhere to the standard principles (e.g. randomization and blinding) of good medical research when performing future studies to assess the performance of the patch test.

Acknowledgements

We thank the participants for their cooperation in this study, and the nurses Sabrina Schemenewitz (Bochum), Dagmar Wieland (Halle/Saale) and Sabine Weimer (Marburg) for skilful technical help.

This multicentre study was supported with expense allowances for the subject group (€100 per subject) from the DKG and with test material (Blenderm™ surgical tapes, Clippers) from 3M™ Medica, 3M Deutschland GmbH, 41460 Neuss, Germany.

References

- Dickel H, Kreft B, Geier J. Strip patch testing does not affect reaction profiles of standard allergens. *Contact Dermatitis* 2015; **73**: 36–43.
- Dickel H, Bruckner T M, Erdmann S M et al. The 'strip' patch test: results of a multicentre study towards a standardization. *Arch Dermatol Res* 2004; **296**: 212–219.
- Dickel H, Kreft B, Kuss O et al. Increased sensitivity of patch testing by standardized tape stripping beforehand: a multicenter diagnostic accuracy study. *Contact Dermatitis* 2010; **62**: 294–302.
- Kraemer H C. The reliability of clinical diagnoses: state of the art. *Annu Rev Clin Psychol* 2014; **10**: 111–130.
- Belsito D V, Storrs F J, Taylor J S et al. Reproducibility of patch tests: a United States multicenter study. *Am J Contact Dermat* 1992; **3**: 193–200.
- Breit R, Agathos M. Qualitätskontrolle der Epikutantestung – Reproduzierbarkeit im Rechts-Links-Vergleich. *Hautarzt* 1992; **43**: 417–421.
- Kwiecien R, Kopp-Schneider A, Blettner M. Concordance analysis – part 16 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2011; **108**: 515–521.
- Uter W, Schnuch A, Gefeller O. Guidelines for the descriptive presentation and statistical analysis of contact allergy data. *Contact Dermatitis* 2004; **51**: 47–56.

- 9 Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015; **349**: 943, aac4716.
- 10 Dickel H, Altmeyer P, Brasch J. 'New' techniques for more sensitive patch testing? *J Dtsch Dermatol Ges* 2011; **9**: 889–896.
- 11 Schnuch A, Aberer W, Agathos M et al. Performing patch testing with contact allergens. *J Dtsch Dermatol Ges* 2008; **6**: 770–775.
- 12 Becker D. Allergic contact dermatitis. *J Dtsch Dermatol Ges* 2013; **11**: 607–622.
- 13 Schnuch A, Geier J, Lessmann H et al. Surveillance of contact allergies: methods and results of the Information Network of Departments of Dermatology (IVDK). *Allergy* 2012; **67**: 847–857.
- 14 Barlow W. Measurement of interrater agreement with adjustment for covariates. *Biometrics* 1996; **52**: 695–702.
- 15 Geier J, Uter W, Lessmann H, Schnuch A. The positivity ratio – another parameter to assess the diagnostic quality of a patch test preparation. *Contact Dermatitis* 2003; **48**: 280–282.
- 16 Brasch J, Henseler T. The reaction index: a parameter to assess the quality of patch test preparations. *Contact Dermatitis* 1992; **27**: 203–204.
- 17 Kuss O, Dickel H. A confidence interval for the reaction index. *Contact Dermatitis* 2010; **62**: 252–253.
- 18 Kottner J, Audigé L, Brorson S et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; **64**: 96–106.
- 19 Bourke J F, Batta K, Prais L et al. The reproducibility of patch tests. *Br J Dermatol* 1999; **140**: 102–105.
- 20 Brasch J, Henseler T, Aberer W et al. Reproducibility of patch tests: a multicenter study of synchronous left-versus right-sided patch tests by the German Contact Dermatitis Research Group. *J Am Acad Dermatol* 1994; **31**: 584–591.
- 21 Gollhausen R, Przybilla B, Ring J. Reproducibility of patch test results: comparison of TRUE Test and Finn Chamber test results. *J Am Acad Dermatol* 1989; **21**: 843–846.
- 22 Gollhausen R, Przybilla B, Ring J. Reproducibility of patch tests. *J Am Acad Dermatol* 1989; **21**: 1196–1202.
- 23 Gollhausen R, Przybilla B, Ring J. High reproducibility of patch tests: reply. *J Am Acad Dermatol* 1991; **24**: 322–323.
- 24 Lindelöf B. A left versus right side comparative study of Finn Chamber patch tests in 220 consecutive patients. *Contact Dermatitis* 1990; **22**: 288–289.
- 25 Memon A A, Friedmann P S. Studies on the reproducibility of allergic contact dermatitis. *Br J Dermatol* 1996; **134**: 208–214.
- 26 Uter W, Pfahlberg A, Brasch J. Zur Reproduzierbarkeit der Epikutantestung – die Bewertung der Konkordanz bei synchroner Applikation. *Allergologie* 2002; **25**: 415–419.
- 27 Bello S, Krogsbøll L T, Gruber J et al. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *J Clin Epidemiol* 2014; **67**: 973–983.
- 28 Savović J, Jones H E, Altman D G et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012; **157**: 429–438.
- 29 Begley C G. Reproducibility: six red flags for suspect work. *Nature* 2013; **497**: 433–434.
- 30 Brasch J, Henseler T, Aberer W et al. Patch testing of nickel sulfate and potassium dichromate with a standardized ready-to-use test system gives highly reproducible results: a double-blind multicentre study. *Acta Derm Venereol (Stockh)* 2001; **81**: 122–124.
- 31 Ale S I, Maibach H I. Reproducibility of patch test results: a concurrent right-versus-left study using TRUE Test™. *Contact Dermatitis* 2004; **50**: 304–312.
- 32 Lachapelle J M, Antoine J L. Problems raised by the simultaneous reproducibility of positive allergic patch test reactions in man. *J Am Acad Dermatol* 1989; **21**: 850–854.
- 33 Carlsen B C, Menné T, Johansen J D. 20 years of standard patch testing in an eczema population with focus on patients with multiple contact allergies. *Contact Dermatitis* 2007; **57**: 76–83.
- 34 Dickel H, Goulioumis A, Gambichler T et al. Standardized tape stripping: a practical and reproducible protocol to uniformly reduce the stratum corneum. *Skin Pharmacol Physiol* 2010; **23**: 259–265.
- 35 Dickel H, Gambichler T, Kamphow J et al. Standardized tape stripping prior to patch testing induces upregulation of Hsp90, Hsp70, IL-33, TNF- α and IL-8/CXCL8 mRNA: new insights into the involvement of 'alarmins'. *Contact Dermatitis* 2010; **63**: 215–222.
- 36 Vigan M, Girardin P, Adessi B, Laurent R. Late reading of patch tests. *Eur J Dermatol* 1997; **7**: 574–576.
- 37 Gefeller O, Pfahlberg A B, Uter W. What can be learnt from nothing? – A statistical perspective. *Contact Dermatitis* 2013; **69**: 350–354.
- 38 Szliska C, Rakoski J. Reproduzierbarkeit von Epikutantestdaten. *Z Hautkr* 1988; **63**: 67–69.
- 39 Mchugh M L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012; **22**: 276–282.