Check for updates

Contact Dermatitis

WILEY

CONTACT DERMATITIS

ORIGINAL ARTICLE OPEN ACCESS

Reproducibility of Patch Test Reactions and Interpretation of Doubtful Results: A Retrospective Study of Identical Haptens Tested Simultaneously in Different Series

Norbertus A. Ipenburg | Anton C. de Groot | Thomas Rustemeyer

Dermato-Allergology and Occupational Dermatology, Amsterdam University Medical Center, Amsterdam, the Netherlands

Correspondence: Anton C. de Groot (antondegroot@planet.nl)

Received: 8 July 2026 | **Revised:** 2 September 2026 | **Accepted:** 9 September 2026

Keywords: Amerchol L-101 | balsam of Peru | concordance | discordance | HEMA | hydroxyethyl methacrylate | hydroxyisohexyl 3-cyclohexene carboxaldehyde | lanolin alcohol | mercaptobenzothiazole | reproducibility of patch tests

ABSTRACT

Background: Reproducibility of patch test reactions remains debated, and the clinical significance of doubtful (?+) reactions is unclear.

Objectives: To evaluate reproducibility of positive reactions to identical haptens tested simultaneously in different series and analyse doubtful reactions using paired co-reactivity patterns.

Materials and Methods: Patch test data from consecutive patients tested during 2015–2025 were analysed retrospectively. Six haptens were evaluated at standard concentrations in petrolatum: hydroxyisohexyl 3-cyclohexene carboxaldehyde (Lyral, HICC) 5% pet., Amerchol L-101 50% pet., lanolin alcohol 30% pet., hydroxyethyl methacrylate (HEMA) 2% pet., mercaptobenzothiazole (MBT) 2% pet. and *Myroxylon pereirae* resin (balsam of Peru, MPR) 25% pet.

Results: Reproducibility was hapten-dependent. HEMA and MBT showed substantial to almost perfect agreement. HICC showed substantial agreement but a significantly higher positivity rate in the perfume series than in the baseline series. Amerchol L-101, lanolin alcohol and MPR showed lower positive agreement and fair-to-moderate kappa values. Most doubtful reactions were isolated, particularly for Amerchol L-101 and lanolin alcohol, whereas 43.8% of doubtful HICC reactions in the baseline series were supported by perfume-series reactivity.

Conclusions: Reproducibility was strongly hapten-dependent. In this dataset, paired co-reactivity provided clinical context for interpreting discordant positive and doubtful reactions.

1 | Introduction

Many studies have investigated the reproducibility of patch test reactions, defined as the ability of patch testing to produce consistent results when performed under comparable conditions, for example, [1–16]. Most of these studies used synchronous duplicate testing of selected identical haptens in patients undergoing diagnostic patch testing. Reactions are considered concordant when both test sites are positive and discordant when one test site is positive and the other is negative. Previous studies have reported reproducibility mostly as the proportion of concordant positive reactions among all patients with at least one positive reaction. Reported concordance rates have varied widely, ranging from 19.7% for cocamidopropyl betaine 1% water [13] to more than 90% [6–10] and even 99.2% with nickel sulfate in the TRUE-test [15]. Most studies were performed 20–35 years ago,

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *Contact Dermatitis* published by John Wiley & Sons Ltd.

and the reproducibility of patch test reactions remains a matter of debate.

In our department, several haptens were historically present in more than one patch test series. This overlap resulted from the composition of independently developed routine and supplementary series and did not reflect an intentional strategy of duplicate testing. Some specialized series, which are tested only when indicated by the patient's history, contained haptens that were also present in or were later added to the European baseline series or in our additional routine series, both of which are tested in all patients undergoing diagnostic patch testing in our department. This provided an opportunity to evaluate the reproducibility of patch test reactions under routine clinical conditions by comparing simultaneous reactions to identical haptens tested in two different series in the same patient.

The aims of this study were: (1) to evaluate the reproducibility of positive patch test reactions to identical haptens tested simultaneously in different patch test series; and (2) to analyse doubtful (?+) reactions, which may be either irritant or allergic [17], by examining their co-reactivity patterns in the corresponding test series, including positive, doubtful and negative reactions to the same hapten.

2 | Materials and Methods

2.1 | Study Population and Patch Testing

This retrospective study was conducted in a cohort of 9088 consecutive patients who underwent diagnostic patch testing at the Department of Dermato-Allergology and Occupational Dermatology of Amsterdam University Medical Center between January 1, 2015 and December 31, 2025. All patients were tested with the European baseline series and the additional routine series of 27 haptens used in our department. Selected supplementary series and patients' personal products and occupational contact materials were tested when indicated by the clinical history.

All haptens included in the present analysis were obtained from SmartPractice Europe GmbH (Barsbüttel, Schleswig-Holstein, Germany; www.smartpracticeeurope.com; brand allergEAZE). Petrolatum-based hapten preparations were applied as a single, uninterrupted ribbon approximately 9 mm in length, placed diagonally across each Van der Bend chamber. This amount corresponded approximately to the recommended dose of 25 mg per chamber [18]. Individual doses were not weighed in routine practice; consistency was sought by applying ribbons of approximately equal length and thickness. All patch tests were applied to the back. When multiple series were tested simultaneously, their relative positions on the back depended on the available space and were not standardized or systematically documented. Patch testing was performed with Van der Bend patch test chambers (Van der Bend, Brielle, The Netherlands), fixed with Omnifix elastic (Paul Hartmann BV, Nijmegen, The Netherlands). The occlusion time was 48 h. Reactions were read by one of three dermatologists on day (D)2 with a second reading and final scoring on D3 or D4, according to the ICDRG criteria as described in the 2015 ESCD guideline [18]. According to these criteria, a doubtful reaction (?+) was defined as faint erythema only, whereas a weak positive reaction (+) required erythema and infiltration, with or without discrete papules; no fixed percentage of the chamber area was used as a separate criterion. D7 readings were not routinely performed, but all patients were instructed to contact the department if new reactions were observed after the final reading.

2.2 | Selection of Allergens and Paired Comparisons

The study focused on identical allergens that were present in two different patch test series and had been tested simultaneously in the same patients. Identical allergens were defined as the same hapten, concentration, vehicle and manufacturer. For each eligible allergen, paired results from the two corresponding test sites were compared within patients.

All haptens meeting the definition of an identical preparation were initially considered for inclusion. For the formal analysis, sufficient data were operationally defined as at least 100 same-date paired observations and at least 10 positive reactions (+, ++ or +++) in each of the two series. These criteria were used to avoid unstable agreement estimates based on sparse contingency tables and were applied independently of the magnitude or direction of the observed agreement. All six haptens included in the formal analysis met both criteria.

Ten candidate haptens did not meet these criteria: *p-tert*-butylphenol formaldehyde resin, lauryl glucoside, propyl gallate, ethylenediamine dihydrochloride, tosylamide/formaldehyde resin, benzyl benzoate, musk xylene, eugenol, cetearyl alcohol and IPPD.

2.3 | Classification of Patch Test Reactions

Patch test reactions scored as +, ++ or +++ were classified as positive. Reactions scored as negative (–), doubtful (?+) or irritant (IR) were classified as non-positive for the primary agreement-discordance analysis, in accordance with routine clinical interpretation of patch test results. Thus, the primary analysis compared positive with non-positive reactions rather than strictly positive with unequivocally negative reactions.

For the primary analysis, a positive reaction in both series was classified as concordant positive, and a non-positive reaction in both series was classified as concordant non-positive. A positive reaction in one series with a non-positive reaction in the other series was classified as discordant. Discordant reactions were recorded separately in both directions.

2.4 | Statistical Analysis

Agreement between paired determinations was assessed using raw (overall) agreement, positive agreement and Cohen's kappa (κ). Raw agreement was defined as the proportion of paired

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

observations assigned to the same dichotomised category, either positive in both series or non-positive in both series. Positive agreement was calculated as $2a/(2a+b+c)$, where a represents concordant positive pairs and b and c represent the two directions of discordant positive/non-positive pairs. Because raw agreement may be dominated by concordant non-positive pairs when positive reactions are uncommon (which is the case with patch test results), Cohen's kappa was calculated to account for agreement expected by chance.

Kappa values were interpreted as follows: ≤ 0.20 poor agreement, >0.20–0.40 fair agreement, >0.40–0.60 moderate agreement, >0.60–0.80 substantial agreement and >0.80 almost perfect agreement. Ninety-five percent confidence intervals (95% CI) based on the standard error of κ were calculated.

A two-sided exact McNemar test was used to test marginal homogeneity, that is, whether the proportion of positive reactions was equal in the two paired determinations. In a paired 2×2 table, marginal homogeneity is equivalent to symmetry of the two discordant cells (positive/non-positive versus non-positive/positive). Consequently, the test also evaluates whether discordant pairs occur equally often in both directions. For the additional reaction-strength analysis, pairs containing at least one positive reaction were classified according to the strongest positive reaction observed within the pair: maximum grade +, maximum grade ++ or maximum grade +++. Within each mutually exclusive stratum, pairwise concordance was calculated as the number of pairs in which both reactions were positive divided by the number of pairs with at least one positive reaction. Because only six haptens were available and several of the ++ and +++ strata were sparse, no formal correlation analysis between reaction strength and hapten-level reproducibility was performed. All p-values were unadjusted, and a $p<0.05$ was considered statistically significant. Statistical analyses were performed using R version 4.6.1 (R Foundation for Statistical Computing, Vienna, Austria).

2.5 | Descriptive Analysis of Doubtful Reactions

Because doubtful (?+) reactions may represent both weak allergic reactions [17] and irritant reactions, they were additionally analysed descriptively. This analysis was intended to assess the co-reactivity associated with doubtful reactions and was not used as an independent confirmatory statistical analysis.

For each allergen, all doubtful reactions in one test series were identified, and the corresponding reaction to the same allergen in the other test series was recorded. Corresponding reactions were categorized as positive (+ to +++), doubtful (?+), negative (–) or irritant (IR). Percentages were calculated using the number of doubtful reactions in the index series as the denominator. This descriptive analysis was performed separately in both directions for each allergen.

For clinical interpretation, doubtful reactions accompanied by a positive reaction in the corresponding series were considered suggestive of weak allergic reactivity. Doubtful reactions accompanied by another doubtful reaction were considered potentially compatible with weak allergic reactivity. Isolated doubtful reactions accompanied by a negative corresponding reaction were considered unlikely to represent contact allergy and more compatible with irritant reactivity. As this analysis was exploratory and clinically interpretative, no formal statistical testing was applied.

3 | Results

3.1 | Agreement and Discordance Between Identical Haptens

Six allergens tested simultaneously in two different patch test series yielded sufficient numbers of paired observations and positive reactions to allow meaningful statistical analysis. All six allergens had been tested in a routine series: five in the European baseline series (EBS) and one in the additional routine series (ARR). The corresponding supplementary series were the perfume series (PER), wool alcohol series (WOL), (meth)acrylate series (ACR), essential oil and flavour series (EOS) and rubber series (RUB). The six allergens included were hydroxyisohexyl 3-cyclohexene carboxaldehyde (Lyral, HICC) 5% pet., Amerchol L-101 50% pet., lanolin alcohol 30% pet., hydroxyethyl methacrylate (HEMA) 2% pet., mercaptobenzothiazole 2% pet. (MBT) and *Myroxylon pereirae* resin (balsam of Peru, MPR) 25% pet.

Table 1 shows, for each allergen, the number of paired tests, the prevalence of positive reactions in both series, the numbers of concordant positive and discordant positive/non-positive results, overall agreement, positive agreement, Cohen's kappa (κ) with 95% confidence intervals, the kappa-based interpretation of agreement quality and exact McNemar p-values.

The number of patients tested with identical haptens in two series ranged from 148 for MPR to 3726 for HICC. The lowest frequencies of positive reactions were observed for HICC (2.1% and 2.9%) and MBT (2.8% and 3.5%). In all six comparisons, the percentage of positive reactions was numerically higher in the supplementary series than in the routine series; this difference in positivity rates was statistically significant for HICC, Amerchol L-101 and lanolin alcohol in the exact McNemar test. Positive agreement was lowest for MPR (46.8%), lanolin alcohol (51.1%), and Amerchol L-101 (54.6%), higher for HICC (70.3%), and highest for MBT (80.0%) and HEMA (87.3%). Cohen's kappa indicated fair agreement for MPR, moderate agreement for Amerchol L-101 and lanolin alcohol, substantial agreement for HICC and MBT and almost perfect agreement for HEMA.

3.2 | Concordance According to Reaction Strength

The concordance of positive reactions stratified according to the strongest reaction within the pair is shown in Table S1. Pooled descriptively across the six haptens, concordance was 40.1% (255/636) when the strongest reaction was +, 90.4% (47/52) when the strongest reaction was ++ and 85.7% (6/7) when the strongest reaction was +++. The +++ category contained only seven pairs and should therefore be interpreted cautiously.

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Agreement and discordance between identical allergens tested in different patch test series.

Allergen[a]	Series compared	*n* tested	Positive reactions, Series 1	Positive reactions, Series 2	1/2, +/+	1/2, +/np	1/2, np/+	Agreement[b]	Positive agreement[c]	Cohen's κ (95% CI)	Kappa agreement	Exact McNemar, p[d]
HICC 5% pet.	1 EBS/2 PER	3726	2.1%	2.9%	65	13	42	98.5%	70.3%	0.695 (0.619–0.772)	Substantial	**< 0.001**
Amerchol L-101 50% pet.	1 ARR/2 WOL	1399	10.6%	12.7%	89	59	89	89.4%	54.6%	0.487 (0.416–0.558)	Moderate	**0.017**
Lanolin alcohol 30% pet.	1 EBS/2 WOL	1399	8.2%	11.7%	71	44	92	90.3%	51.1%	0.459 (0.381–0.536)	Moderate	**< 0.0001**
HEMA 2% pet.	1 EBS/2 ACR	533	12.8%	13.9%	62	6	12	96.6%	87.3%	0.854 (0.788–0.920)	Almost perfect	0.238
MBT 2% pet.	1 EBS/2 RUB	395	2.8%	3.5%	10	1	4	98.7%	80%	0.794 (0.617–0.970)	Substantial	0.375
MPR 25% pet.	1 EBS/2 EOS	148	13.5%	18.2%	11	9	16	83.1%	46.8%	0.370 (0.173–0.567)	Fair	0.230

Note: Kappa values were interpreted as follows: < 0.20 poor agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and > 0.80 almost perfect agreement. np: non-positive; this is the sum of negative, dubious positive (?+) and irritant (IR) patch test reactions. +: positive patch test reactions (+ to +++); +/+ = concordant positive reactions; +/np = series 1 positive/series 2 non-positive; np/+ = series 1 non-positive/series 2 positive.

Abbreviations: ACR, acrylate series; ARR, additional routine series; EBS, European baseline series; EOS, essential oil and flavour series; HEMA, hydroxyethyl methacrylate; HICC, hydroxyisohexyl 3-cyclohexene carboxaldehyde (Lyral); MBT, mercaptobenzothiazole; MPR, *Myroxylon pereirae* resin (Balsam of Peru); PER, perfume series; RUB, rubber series; WOL, wool alcohol series.

[a]All haptens in petrolatum and of the Brand Allergeaze (www.smartpracticeeurope.com).

[b]Overall/raw agreement: sum of concordant positive and concordant non-positive results divided by the number of all paired tests.

[c]Positive agreement = $2a/(2a+b+c)$, where a = +/+, b = +/np, and c = np/+; this represents the proportion of positive test results that were paired with a positive result in the corresponding series.

[d]Significant exact McNemar p-values (< 0.05) are shown in bold.

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Reaction strength did not explain all differences between haptens. Within pairs in which neither reaction was stronger than +, concordance was 67.9% for HEMA, 58.3% for MBT, 47.1% for HICC, 36.5% for Amerchol L-101, 34.0% for lanolin alcohol and 25.8% for MPR.

3.3 | Doubtful (?+) Reactions and Co-Reactivity in the Corresponding Series

Doubtful (?+) reactions were analysed separately to explore whether these reactions were supported by reactivity to the same hapten in the corresponding test series; the results are shown in Table 2. Of the 3 haptens with many doubtful (?+) reactions, percentages of positive (+ to +++) co-reactions were 1.6% and 18.8% for HICC, 22.4% and 16.3% for Amerchol L-101 and 25% and 14.3% for lanolin alcohol. When combined with doubtful co-reactivity, these proportions rise to 14.8% and 43.8% for HICC, 44.8% and 30.8% for Amerchol L-101 and 35.4% and 19.8% for lanolin alcohol. Accordingly, most doubtful reactions in the index series were negative in the corresponding series, proportions ranging from 55.2% for Amerchol L-101 in ARR to 85.2% for HICC in PER.

For HICC, doubtful reactions showed a marked directional pattern. Among 61 doubtful reactions in PER, 1 (1.6%) had a positive co-reaction in EBS, whereas among 32 doubtful reactions in EBS, 6 (18.8%) had a positive co-reaction in PER. Thus, a doubtful HICC reaction in EBS was more often supported by reactivity in PER than the other way around.

For HEMA, MBT and MPR, the numbers of doubtful reactions were small. HEMA had relatively frequent co-reactivity: three of 4 (75.0%) doubtful reactions in EBS were accompanied by a positive or doubtful reaction in ACR, and 4 of 8 (50.0%) doubtful reactions in ACR were accompanied by a positive or doubtful reaction in EBS. For MBT, doubtful reactions were accompanied by a negative co-reaction in 11 of 12 cases (91.7%).

MPR had 6 doubtful reactions in both EBS and EOS, of which 2 (33.3%) had a positive co-reaction in the corresponding series, while 4 of 6 (66.7%) were isolated doubtful reactions with a negative corresponding test.

An overview of the frequencies of all paired reaction combinations, with positive, doubtful, negative and irritant reactions shown separately, is provided in Table S2.

4 | Discussion

4.1 | Reproducibility of Positive Reactions to Identical Haptens

In this study, identical haptens tested simultaneously in different patch test series showed marked differences in reproducibility of positive patch test reactions. Although overall raw agreement was high for most haptens, this was mainly driven by the large number of concordant negative reactions. More clinically informative differences were observed in the number and direction of discordant positive/non-positive reactions and in the proportion of positive reactions that were concordant between series.

4.1.1 | HEMA and MBT

HEMA may serve as an example of a highly reproducible hapten in this dataset. Positive reactions to HEMA were usually present in both the baseline and acrylate series, while discordant reactions were infrequent. Positivity rates in the two series were very similar, and the exact McNemar test showed no statistically significant difference between them. These observations were reflected by high positive agreement and an almost perfect Cohen's kappa value. Clinically, this supports the view that most positive HEMA reactions represented true allergic sensitisation rather than irritant reactivity. MBT also showed high reproducibility, with few discordant reactions and substantial agreement. However, the number of positive MBT reactions was small, and conclusions should therefore be more cautious than for HEMA.

4.1.2 | Amerchol L-101 and Lanolin Alcohol

The findings for lanolin alcohol and Amerchol L-101 were different. For both haptens, positive reactions were much less consistently concordant between the two series, resulting in relatively low positive agreement and only moderate kappa values. In other words, a substantial proportion (45%–49%) of patients with a positive reaction in one series did not have a corresponding positive reaction in the other series. These findings fit with clinical experience that lanolin alcohol and especially Amerchol L-101 frequently produce weak positive, doubtful, irritant or otherwise difficult-to-interpret reactions [19].

It is therefore plausible that a proportion of discordant positive reactions to these haptens represented irritant responses rather than true allergic reactions. The present data cannot prove this for individual patients, because no independent gold standard for allergic sensitisation was available. Nevertheless, the high number of discordant reactions, together with the lower positive agreement, supports the view that positive reactions to these haptens require careful clinical interpretation.

4.1.3 | Hydroxyisohexyl 3-Cyclohexene Carboxaldehyde

HICC showed a different pattern. Overall agreement was very high and positive agreement was substantial (70.3%), but the positivity rate was significantly higher in the perfume series than in the baseline series (2.9% vs. 2.1%; exact McNemar $p < 0.001$). This difference corresponded to an excess of discordant pairs that were positive in the perfume series and non-positive in the baseline series compared with the reverse direction (42 vs. 13).

A possible explanation is immune-stimulation by the presence of neighbouring related allergens [20]. In this scenario, HICC reactions may be facilitated by the local immunological environment of the perfume series. In patients sensitized to fragrances, neighbouring reactions to other fragrance haptens may enhance local inflammatory or immune activation, allowing weak HICC allergy to become clinically visible in the

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Doubtful (?+) reactions: co-reactivity in the corresponding test series.

Allergen	Index series with ?+	Corresponding series	*n* ?+ in index series	Positive in corresponding series (+ to +++)	Doubtful in corresponding series (?+)	Positive or doubtful co-reactivity (+ to +++ or ?+)	Negative in corresponding series (–)
HICC	EBS	PER	32	6 (18.8%)	8 (25%)	14 (43.8%)	18 (56.2%)
	PER	EBS	61	1 (1.6%)	8 (13.1%)	9 (14.8%)	52 (85.2%)
Amerchol L-101	ARR	WOL	67	15 (22.4%)	15 (22.4%)	30 (44.8%)	37 (55.2%)
	WOL	ARR	104	17 (16.3%)	15 (14.4%)	32 (30.8%)	72 (69.2%)[a]
Lanolin alcohol	EBS	WOL	48	12 (25%)	5 (10.4%)	17 (35.4%)	31 (64.6%)
	WOL	EBS	91	13 (14.3%)	5 (5.5%)	18 (19.8%)	73 (80.2%)
HEMA	EBS	ACR	4	2 (50%)	1 (25%)	3 (75%)	1 (25%)
	ACR	EBS	8	3 (37.5%)	1 (12.5%)	4 (50%)	4 (50%)
MBT	EBS	RUB	5	1 (20%)	0 (0%)	1 (20%)	4 (80%)
	RUB	EBS	7	0 (0%)	0 (0%)	0 (0%)	7 (100%)
MPR	EBS	EOS	6	2 (33.3%)	0 (0%)	2 (33.3%)	4 (66.7%)
	EOS	EBS	6	2 (33.3%)	0 (0%)	2 (33.3%)	4 (66.7%)

Note: For each allergen, the table starts from all doubtful (?+) reactions in one series and shows the reaction observed to the same allergen in the corresponding series. Percentages are row percentages, using the number of doubtful reactions in the index series as denominator. Interpretation guide: positive co-reactivity (?+/+) is suggestive of weak allergic reactivity; doubtful co-reactivity (?+/?+) is potentially compatible with weak allergic reactivity; isolated doubtful reactions with a negative corresponding test (?+/–) are unlikely to represent contact allergy.

Abbreviations: ACR, acrylate series; ARR, additional routine series; EBS, European baseline series; EOS, essential oil and flavour series; HEMA, hydroxyethyl methacrylate; HICC, hydroxyisohexyl 3-cyclohexene carboxaldehyde (Lyral); MBT, mercaptobenzothiazole; MPR, *Myroxylon pereirae* resin (Balsam of Peru); PER, perfume series; RUB, rubber series; WOL, wool alcohol series.

[a]Includes 1 irritant reaction.

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

perfume series while remaining negative (or ?+) in the baseline series. This interpretation would fit with the excess of perfume-series-positive/baseline-series-non-positive discordant reactions. It also suggests that discordance is not always best explained by technical error or poor test quality; in some cases, it may reflect context-dependent elicitation of weak allergic reactivity.

At the same time, it cannot be excluded that some of the additional perfume-series-positive reactions were false-positive, resulting from non-specific inflammatory or spreading effects from nearby positive patch test reactions [4, 20]. As the occurrence and proximity of other positive reactions were not systematically analysed, the contribution of this mechanism could not be determined.

4.1.4 | *Myroxylon pereirae* Resin

MPR showed the lowest raw agreement (83.1%) and the lowest Cohen's kappa value ($\kappa = 0.370$) among the six haptens analysed. Positive agreement was also low (46.8%), and discordant reactions were relatively frequent. However, this comparison included only 148 patients because MPR 25% in petrolatum had only been included in the essential oil and flavour series from March 2022 onward.

The positivity rate was numerically higher in the essential oil and flavour series than in the EBS (18.2% versus 13.5%), but this difference was not statistically significant in the exact McNemar test ($p = 0.230$). Thus, in this dataset, positive reactions to MPR were less consistently reproduced than positive reactions to HEMA, MBT and HICC. However, the relatively small numbers of paired observations and positive reactions resulted in considerable statistical uncertainty and limits firm conclusions.

4.1.5 | Influence of Reaction Strength

The descriptive stratified analysis suggested that stronger positive reactions were generally more reproducible. Across all six haptens, pairwise concordance was 40.1% (255/636) among pairs with a maximum reaction grade of +, 90.4% (47/52) among pairs with a maximum grade of ++ and 85.7% (6/7) among pairs with a maximum grade of +++. Thus, concordance was considerably higher for ++ than for + reactions. However, the estimate for +++ reactions was based on only seven pairs and should therefore be interpreted cautiously.

Reaction strength did not fully account for the observed differences between haptens. Within the maximum-grade + stratum, pairwise concordance ranged from 25.8% for MPR to 67.9% for HEMA. This suggests that reaction strength may contribute to reproducibility, but that hapten-, preparation-, patient- and test-context-related factors may also be involved. Because only six haptens were included, their paired populations differed substantially and several hapten-specific ++ and +++ strata were sparse, no formal hapten-level correlation analysis was performed.

4.2 | Interpretation of Doubtful (?+) Reactions According to Paired Co-Reactivity

In our opinion, the separate analysis of doubtful (?+) reactions adds an important dimension to the interpretation of paired patch test results. In routine dichotomised analyses, doubtful reactions are commonly classified as non-positive. This is practical and consistent with conventional scoring for statistical analyses and clinical interpretation of patch test reactions, but it may obscure clinically relevant differences within this heterogeneous category. A doubtful reaction accompanied by a positive or doubtful reaction to the same hapten in the corresponding series may be clinically different from an isolated doubtful reaction with a negative corresponding test.

We therefore explored whether doubtful reactions were supported by co-reactivity in the paired series. This approach is clinically relevant because doubtful reactions may represent either weak allergic reactions or irritant reactions. We considered doubtful reactions accompanied by a positive co-reaction to be suggestive of weak allergic reactivity. Doubtful/doubtful combinations may also be compatible with weak allergic reactivity, although less convincingly. In contrast, isolated doubtful reactions with a negative corresponding reaction were considered unlikely to represent contact allergy and more compatible with irritant reactivity.

4.2.1 | Hydroxyisohexyl 3-Cyclohexene Carboxaldehyde

The results showed that the presumed clinical meaning of doubtful reactions differed between haptens and also depended on the direction of comparison. HICC showed the most striking asymmetry. Doubtful reactions in the perfume series were usually isolated, with only 14.8% accompanied by positive or doubtful co-reactivity in the baseline series. Conversely, 43.8% of doubtful HICC reactions in the baseline series were accompanied by positive or doubtful co-reactivity in the perfume series. This suggests that a doubtful HICC reaction in the baseline series may carry more clinical weight than an isolated doubtful reaction in the perfume series. One possible explanation for the high proportion of isolated doubtful HICC reactions in the perfume series is that they may have resulted from non-specific inflammatory effects from nearby positive patch test reactions [4, 20]. In accordance with the interpretative framework used in this study, these isolated doubtful reactions were considered unlikely to represent contact allergy and more compatible with irritant reactivity. As the occurrence and proximity of other positive reactions were not systematically analysed, the contribution of this mechanism could not be determined.

4.2.2 | Amerchol L-101 and Lanolin Alcohol

For Amerchol L-101 and lanolin alcohol, the pattern was different. Both haptens showed substantial numbers of doubtful reactions, but many of these were isolated, particularly when the doubtful reaction occurred in the wool alcohol series. With Amerchol L-101, positive or doubtful co-reactivity was present in 30.8% of doubtful reactions in the wool alcohol series, whereas

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

68.3% were negative and 1.0% showed an irritant reaction in the corresponding routine series. For lanolin alcohol, this pattern was even more pronounced: only 19.8% of doubtful reactions in the wool alcohol series showed positive or doubtful co-reactivity in the baseline series, while 80.2% were isolated.

These findings are consistent with the clinical impression that doubtful reactions to Amerchol L-101 and lanolin alcohol often represent poorly reproducible, likely irritant reactions [19]. The data do not prove that isolated doubtful reactions are irritant, because no independent gold standard was available. However, the high proportion of isolated doubtful reactions, together with the limited reproducibility of positive reactions observed in the main analysis, supports a cautious interpretation of weak and doubtful reactions to these haptens.

4.2.3 | HEMA, MBT and MPR

HEMA showed a contrasting pattern. Doubtful HEMA reactions were relatively often accompanied by positive or doubtful co-reactivity in the corresponding series. This is in line with the high reproducibility of unequivocal positive HEMA reactions in the main analysis and suggests that doubtful HEMA reactions may deserve careful clinical consideration. However, because only a few doubtful HEMA reactions were present, this observation should be interpreted cautiously.

For MBT and MPR, the numbers of doubtful reactions were also small, limiting firm conclusions. Most doubtful MBT reactions were isolated, particularly in the rubber series. For MPR, positive co-reactivity occurred in one third of doubtful reactions in both directions, but the total number of doubtful reactions was very low. These findings should therefore be regarded as descriptive only.

4.2.4 | Evaluation

Overall, this analysis suggests that doubtful patch test reactions should not be treated as a uniform category. Their interpretation may be improved by considering paired co-reactivity to the same hapten in another test series. Doubtful reactions accompanied by positive or doubtful co-reactivity may indicate weak allergic reactivity, whereas isolated doubtful reactions are less supportive of allergy and may more often reflect irritant responses. This contextual approach may be particularly useful for haptens known to produce weak, irritant, or difficult-to-interpret reactions.

This analysis remains exploratory and clinically interpretative. It was not designed to establish the true allergic or irritant nature of individual doubtful reactions. Nevertheless, it illustrates that paired testing of identical haptens can provide additional information about the possible significance of doubtful reactions beyond their conventional classification as non-positive.

4.3 | Comparing Results With Previous Studies

Comparison of our results with previous studies on the reproducibility of patch test reactions is difficult for several reasons. First, the study design differed substantially. Whereas most previous studies used prospective duplicate testing of selected haptens in consecutive patients, our study was based on a retrospective review of routine patch test data from more than 9000 consecutive patients undergoing diagnostic patch testing over an 11-year period.

Second, three of the haptens included in the present study (HICC, Amerchol L-101 and HEMA) have, to our knowledge, not been included in previous reproducibility studies. Third, our patch test materials (haptens, patch test chambers, fixation) differed from those used in the other studies. Most previous investigations used investigator-loaded patch test chambers, as in our study, whereas others used ready-to-use systems such as TRUE Test [2, 12, 15] or Epiquick [9, 10]. These methodological differences may affect reproducibility.

In addition, lanolin alcohol, MBT and MPR were each evaluated in only a limited number of previous studies. Some of these included fewer than 50 patients or only very few positive reactions, which limits the reliability of estimates of concordance and discordance in paired testing [6, 11, 12].

For lanolin alcohol, previous studies reported positive reaction-concordance rates of 56% [3], 57% [4], 70% [2] and 100% [7], whereas positive agreement in our study was 51% (Table 1). The highest reported concordance [7] should be interpreted cautiously because it was based on only 10 positive reactions. The study reporting 70% concordance used the TRUE Test system, in which a standardized amount of hapten is applied, which may have contributed to higher reproducibility [2]. For MBT, one previous study reported concordance in 100% of 14 positive reactions, broadly consistent with the high positive agreement (80%) observed in our study, although our estimate was based on a larger paired dataset (Table 1). For MPR, previous studies reported positive concordance rates of 70% [4], 75% [3], 94% [7] and 95% [2], which are considerably higher than the positive agreement observed in our study.

Despite these differences, our findings are broadly in line with earlier observations that reproducibility of patch test reactions is strongly hapten-dependent. Discordance appears to occur more frequently with haptens that have irritant or marginal irritant properties, or that produce weak, difficult-to-interpret reactions [1, 3, 4]. These include haptens such as formaldehyde, potassium dichromate, benzalkonium chloride [16] and cocamidopropyl betaine [13]. It also applies to haptens with relatively weak allergenic potency, such as lanolin alcohol [1, 3, 4]. Our findings support the view that lanolin alcohol belongs to this group of haptens with limited reproducibility and a propensity for weak or irritant reactions.

4.4 | Duplicate Patch Testing in Clinical Practice

The duplicate testing evaluated in this study arose from historical overlap between routine and clinically indicated supplementary series and was not an intentional diagnostic strategy. Our findings should therefore not be interpreted as supporting routine duplicate testing. Testing an identical hapten twice generally adds to the test burden and may generate discordant

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

results that complicate clinical interpretation and occasionally prompt retesting. Although paired co-reactivity provided useful research context for interpreting doubtful reactions in the present study, this incidental advantage does not, in our view, justify deliberate duplicate testing in routine clinical practice. We are therefore revising our supplementary series to remove unnecessary exact duplicates.

4.5 | Limitations

This study has several limitations. It was retrospective and based on routine clinical patch testing. Because supplementary series were tested only when clinically indicated, the paired populations were not randomly selected and differed between haptens, which may limit generalisability to unselected patch-tested populations. The batches of the test haptens in the routine series and the specialized series may not have been identical. The number of paired observations differed considerably between allergens, and some comparisons included relatively small numbers of positive reactions. The analyses were based on test results recorded in clinical practice, without an independent gold standard for allergic sensitisation. Therefore, discordant reactions cannot be definitively classified as false-positive, false-negative, irritant, or weak allergic reactions. In addition, the classification of doubtful reactions remains partly interpretative. Local effects of neighbouring allergens and differences in the position of test chambers could not be evaluated retrospectively. Although all reactions were scored according to the ICDRG criteria as described in the ESCD guideline, interobserver variability between the three dermatologists, particularly in distinguishing doubtful, weak positive and irritant reactions, could not be assessed because no formal interobserver evaluation or standardized photographic documentation was available. Finally, D7 readings were not routinely performed, so late reactions may have been missed. Clinical relevance was not systematically assessed.

4.6 | Conclusions

Identical haptens tested simultaneously in different patch test series showed marked hapten-dependent differences in reproducibility of positive patch test reactions. HEMA (2% pet.) and MBT (2% pet.) showed high concordance of positive reactions, supporting their reliability as patch test allergens. By contrast, lanolin alcohol (30% pet.), Amerchol L-101 (50% pet.) and MPR (25% pet.) showed more frequent discordance, suggesting that positive reactions to these test preparations require more cautious clinical interpretation. HICC (5% pet.) showed high overall agreement but a significantly higher positivity rate in the perfume series than in the baseline series, possibly reflecting context-dependent elicitation of weak fragrance allergy.

Doubtful (?+) reactions should not be regarded as a uniform non-positive category. Their clinical significance appears to depend on paired co-reactivity. Doubtful reactions accompanied by a positive co-reaction are suggestive of weak allergic reactivity, and doubtful/doubtful combinations may potentially also be compatible with weak allergic reactivity. By contrast, isolated doubtful reactions with a negative corresponding reaction are likely compatible with irritant reactivity. In this retrospective dataset, paired observations provided additional information on the reproducibility and interpretation of positive and doubtful patch test reactions. This does not imply that routine duplicate testing is necessary or recommended.

Author Contributions

Norbertus A. Ipenburg: conceptualization, investigation, data curation, methodology, writing – review and editing. **Anton C. de Groot:** conceptualization, visualization, writing – original draft, writing – review and editing. **Thomas Rustemeyer:** supervision, writing – review and editing.

Acknowledgements

AI disclosure: The authors used ChatGPT model 5.5 (OpenAI, USA), between 21 May 2026 and 8 June 2026. The tool was used to assist with English language editing, organization and streamlining of the manuscript, statistical analyses based on author-provided data, generation of tables and preparation of the graphical image/graphical abstract.

AI assistance was used in the manuscript text, tables and Supporting Information tables, the statistical summaries reported in the Results section, the graphical image/graphical abstract and responses to reviewers' comments. All AI-generated output, including text, tables, statistical calculations and the graphical image, was critically reviewed, checked, edited where necessary and approved by the authors. Statistical outputs were verified against the original datasets by the authors, and all scientific interpretations, conclusions and clinical relevance assessments were made by the authors.

To maintain privacy and compliance requirements, no directly identifiable human subject data were entered into the AI tool. The data provided to the tool were fully anonymized before use and contained no names, initials, dates of birth, addresses, patient numbers or other direct identifiers. The use of AI did not replace author responsibility for the accuracy, integrity and originality of the manuscript.

Funding

The authors have nothing to report.

Consent

Informed consent was obtained from all participants.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

References

1. H. Dickel, J. Geier, B. Kreft, W. Pfützner, and O. Kuss, "Comparing Reliabilities of Strip and Conventional Patch Testing," *Contact Dermatitis* 76, no. 6 (2017): 342–349, https://doi.org/10.1111/cod.12758.

2. S. I. Ale and H. I. Maibach, "Reproducibility of Patch Test Results: A Concurrent Right-Versus-Left Study Using TRUE Test," *Contact Dermatitis* 50, no. 5 (2004): 304–312, https://doi.org/10.1111/j.0105-1873.2004.00351.x.

3. J. F. Bourke, K. Batta, L. Prais, A. Abdullah, and I. S. Foulds, "The Reproducibility of Patch Tests," *British Journal of Dermatology* 140, no. 1 (1999): 102–105, https://doi.org/10.1046/j.1365-2133.1999.02615.x.

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4. J. Brasch, T. Henseler, W. Aberer, et al., "Reproducibility of Patch Tests. A Multicenter Study of Synchronous Left-Versus Right-Sided Patch Tests by the German Contact Dermatitis Research Group," *Journal of the American Academy of Dermatology* 31, no. 4 (1994): 584–591, https://doi.org/10.1016/s0190-9622(94)70220-9.

5. R. Breit and M. Agathos, "Qualitätskontrolle der Epikutantestung – Reproduzierbarkeit Im Rechts-Links-Vergleich [Quality Control in Epicutaneous Testing – Reproducibility in the Right-Left Comparison]," *Der Hautarzt* 43, no. 7 (1992): 417–421.

6. M. T. Bousema, A. M. Geursen, and T. van Joost, "High Reproducibility of Patch Tests," *Journal of the American Academy of Dermatology* 24, no. 2 Pt 1 (1991): 322–323, https://doi.org/10.1016/s0190-9622(08)80640-8.

7. J. Macháčková and O. Seda, "Reproducibility of Patch Tests," *Journal of the American Academy of Dermatology* 25, no. 4 (1991): 732–733, https://doi.org/10.1016/s0190-9622(08)80686-x.

8. B. Lindelöf, "A Left Versus Right Side Comparative Study of Finn Chamber Patch Tests in 220 Consecutive Patients," *Contact Dermatitis* 22, no. 5 (1990): 288–289, https://doi.org/10.1111/j.1600-0536.1990.tb01598.x.

9. J. M. Lachapelle and J. L. Antoine, "Problems Raised by the Simultaneous Reproducibility of Positive Allergic Patch Test Reactions in Man," *Journal of the American Academy of Dermatology* 21, no. 4 Pt 2 (1989): 850–854, https://doi.org/10.1016/s0190-9622(89)70265-6.

10. J. M. Lachapelle, "A Left Versus Right Side Comparative Study of Epiquick Patch Test Results in 100 Consecutive Patients," *Contact Dermatitis* 20, no. 1 (1989): 51–56, https://doi.org/10.1111/j.1600-0536.1989.tb03095.x.

11. R. Gollhausen, B. Przybilla, and J. Ring, "Reproducibility of Patch Tests," *Journal of the American Academy of Dermatology* 21, no. 6 (1989): 1196–1202, https://doi.org/10.1016/s0190-9622(89)70329-7.

12. R. Gollhausen, B. Przybilla, and J. Ring, "Reproducibility of Patch Test Results: Comparison of TRUE Test and Finn Chamber Test Results," *Journal of the American Academy of Dermatology* 21, no. 4 Pt 2 (1989): 843–846, https://doi.org/10.1016/s0190-9622(89)70263-2.

13. A. Schnuch, H. Lessmann, J. Geier, and W. Uter, "Is Cocamidopropyl Betaine a Contact Allergen? Analysis of Network Data and Short Review of the Literature," *Contact Dermatitis* 64, no. 4 (2011): 203–211, https://doi.org/10.1111/j.1600-0536.2010.01863.x.

14. A. A. Memon and P. S. Friedmann, "Studies on the Reproducibility of Allergic Contact Dermatitis," *British Journal of Dermatology* 134, no. 2 (1996): 208–214.

15. J. Brasch, T. Henseler, W. Aberer, 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 Dermato-Venereologica* 81, no. 2 (2001): 122–124, https://doi.org/10.1080/00015550152384263.

16. W. Uter, H. Lessmann, J. Geier, and A. Schnuch, "Is the Irritant Benzalkonium Chloride a Contact Allergen? A Contribution to the Ongoing Debate From a Clinical Perspective," *Contact Dermatitis* 58, no. 6 (2008): 359–363, https://doi.org/10.1111/j.1600-0536.2008.01327.x.

17. P. Arora, C. Brumley, and S. Hylwa, "Clinical Relevance of Doubtful Reactions in Patch Testing: A Single-Centre Retrospective Study," *Contact Dermatitis* 90, no. 6 (2024): 607–612, https://doi.org/10.1111/cod.14526.

18. J. D. Johansen, K. Aalto-Korte, T. Agner, et al., "European Society of Contact Dermatitis Guideline for Diagnostic Patch Testing – Recommendations on Best Practice," *Contact Dermatitis* 73, no. 4 (2015): 195–221, https://doi.org/10.1111/cod.12432.

19. J. Knijp, D. P. Bruynzeel, and T. Rustemeyer, "Diagnosing Lanolin Contact Allergy With Lanolin Alcohol and Amerchol L101," *Contact Dermatitis* 80, no. 5 (2019): 298–303, https://doi.org/10.1111/cod.13210.

20. I. Duarte, R. Lazzarini, and R. Buense, "Interference of the Position of Substances in an Epicutaneous Patch Test Battery With the Occurrence of False-Positive Results," *American Journal of Contact Dermatitis* 13, no. 3 (2002): 125–132.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Concordance of paired positive reactions according to the strongest reaction grade within the pair. **Table S2:** Overall frequencies of paired reaction patterns for identical haptens tested simultaneously in two series.

16000536, 2026, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cod.70251, Wiley Online Library on [04/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License