

The validity of rating patch test reactions based on digital images

WOLFGANG UTER¹, DETLEF BECKER², AXEL SCHNUCH³, OLAF GEFELLER¹ AND PETER J. FROSCH⁴

¹Department of Medical Informatics, Biometry and Epidemiology, University of Erlangen-Nürnberg, Erlangen,
²Department of Dermatology, University Hospital of Mainz, ³Information Network of Departments of Dermatology
at the University of Göttingen, and ⁴Department of Dermatology, University of Witten/Herdecke and
Klinikum Dortmund GmbH, Germany

Background: The proper reading of patch test (PT) reactions, based on morphological criteria, is of utmost importance. Digital images are increasingly used for medical training.

Objectives: To assess the diagnostic validity of readings of 20 digital images of various PT reaction grades by congress attendants.

Methods: 122 volunteers took a PT quiz offered during the 8th ESCD meeting, September 2006, Berlin. No information on the respective allergen was given, nor the gold standard grading determined by an European Environmental and Contact Dermatitis Research Group (EECDRG) expert panel was disclosed while the quiz was open.

Results: Overall, 63.5% of all ratings agreed 'exactly' with the gold standard, and 87.8% 'grossly', that is, with regard to the distinction between non-positive [doubtful or irritant (IR)] and positive (+, ++, and +++) reactions. Most images prompted a fair proportion of correct classifications. The distinction between doubtful and + and between ++ and +++, and strong bullous IR reactions proved partly difficult, probably because of blinding of the test substance. The least experienced in particular tended to misclassify IR reactions.

Conclusions: The rating of digital images of PT reactions represents just 1 particular, isolated aspect of PT evaluation. Results were largely valid. Thus, the method could be used for continuing medical education, and for standardization in multicentre networks.

Key words: contact allergy; continuing medical education; digital image; interrater agreement; patch testing; validity. © Blackwell Munksgaard, 2007.

Accepted for publication 15 June 2007

Patch testing is still the main tool for diagnosing contact allergy. Standardization of the patch test (PT) procedure has reached a high level, regarding chemical characterization of allergens, their stability during shelf life in the vehicle and test concentration identified as optimal. International guidelines regarding the application, reading, and interpretation of PT results have been published (1). In this process, the visual evaluation of the PT reaction is an essential step. However, similar to other diagnostic tests which are visually read, different observers may read the same PT reaction differently. Hence, a certain proportion of readings will be incorrect, limiting the validity of the test. While misclassification between ++

and +++ reactions is not important, the distinction between, for example, doubtful (?), and weak positive (+) reactions is more crucial.

The present analysis, based on the results of a computer-based PT quiz, aims at quantifying the amount of variability and divergence from gold standard ratings of 20 different digital images of PT reactions. Moreover, we try to identify factors, which could influence (dis)agreement with the gold standard.

Methods

From a collection of digital images compiled by the EECDRG, 20 images (Fig. 1a–t) were selected



Fig. 1. Figure 1 a–t: patch test reactions evaluated in the quiz: name of the material tested, test concentration (both not disclosed), information on palpable infiltration, gold standard rating (also not disclosed during the quiz).

to be evaluated by participants of the 8th Congress of the European Society of Contact Dermatitis (<http://orgs.dermis.net/content/e01escd/e01aims>) in Berlin, from 14 September to 16 September (<http://www.escd.org>) and the subsequent German satellite meeting. The images were presented in a separate seminar room on liquid crystal display units. Registered participants of the congresses ($n = 480$) were invited to take a quiz on this set of images. Volunteers first filled in anonymous categorical information on the coun-

try of residence, their qualification and their workplace, and their years of PT expertise. Next, the participants rated each of the 20 images as either negative ('negative'), doubtful ('?' or '?+': faint erythema only), weak positive ('+': erythema, infiltration, possibly papules), strong positive ('++': erythema, infiltration, papules, vesicles), extreme positive ('+++': intense erythema and infiltration and coalescing vesicles) or irritant ('IR', different types), following the internationally accepted classification (1, 2). Each image

included information on the presence or absence of palpable infiltration (PI) (Fig. 1a–t).

All user entries were stored electronically. After descriptive analysis of results, the statistical evaluation focussed on the issues of (i) agreement of raters with the gold standard and (ii) those factors which are associated with it. For this purpose, an indicator variable termed 'exact agreement' was conceived which was set to 1 if the individual rating exactly corresponded to the pre-defined gold standard, and to 0 otherwise. Furthermore, to address the classification of a PT reaction as allergic versus non-allergic, another indicator variable was calculated, termed 'gross agreement'. This was set to 1 either if both the gold standard and the individual rating indicated an allergic PT reaction (in the range of '+' to '+++') or if both indicated a non-allergic reaction ('IR' or '?'), and to 0 otherwise. The association between disagreement and years of experience in patch testing (categorized into 3 levels, according to the tertiles of the distribution of the original data: up to 5, 6 to 15 years and > 15 years), being a dermatologist versus a researcher, according to self-classification, and the above mentioned outcome-indicator was quantified in bivariate analyses by the prevalence ratio (PR), accompanied by an asymptotic 95% confidence interval (CI). For data analysis, the statistical software package SAS (version 9.1, SAS Institute, Cary, NC, USA) was used.

Results

A total of 122 persons took part in the PT quiz. Most of them classified themselves as consultant in dermatology ($n = 51$), 27 as researcher (of these, 10 from industry), 24 as dermatologists in private practice, 12 as professors and 8 as registrars in dermatology. Participants came from 22 different countries: Germany ($n = 28$), UK ($n = 12$), France ($n = 9$), Poland, and Spain ($n = 8$ each), Denmark and Sweden ($n = 7$ each), and lesser numbers from other countries, including 11 visitors from outside Europe. The median (and mean) experience in patch testing was 12 years, ranging from 1 year (14 cases) to 42 years (1 case). Experience in patch testing did not differ significantly between dermatologists ($n = 95$) and others ($n = 27$) ($P = 0.67$, Wilcoxon-Mann-Whitney test).

Descriptive results of the readings of each image are presented in Table 1. Evidently, the agreement (in terms of % readings agreeing with the gold standard) varied between different images and reading grades, respectively. Moreover, the consideration of 'gross' agreement greatly increased subsequent concordance, see 2 rightmost columns.

On average, 63.5% of all ratings agreed 'exactly' and 87.8% 'grossly', that is, with regard to the distinction between non-positive and positive. Considering sensitivity (the proportion of true-positive ratings) and specificity (the proportion of true-negative, or rather, true-non positive ratings), the following was observed: Among dermatologists, pooled sensitivity was 87.9%, and specificity 89.1%, while among researchers the respective proportions were 85.9% and 85.2%, respectively (differences not significant, Chi square test). Experience in patch testing partly influenced the ratings: Sensitivity was 85.0% with up to 5 years, 90.0% with 6–15 years and 87.3% with >15 years experience (not significant). Specificity, however, was significantly associated with experience, with proportions increasing from 84.2% to 87.8%, reaching 92.5% in the most experienced category ($P = 0.0004$, Cochran-Armitage Trend Test).

Analysing single images, the risk of disagreeing with the gold standard varied within the categories of experience and being a dermatologist, respectively, which was significant in a few instances (CI not including 1, Table 2). In this context, a PR of greater than 1 indicates a higher frequency of disagreement, compared with the respective reference category (non-dermatologists and 'experienced' raters). In contrast, a PR of less than 1 indicates less disagreement than in the reference category, that is, better diagnostic performance than in the reference category, as measured against the gold standard.

Discussion

The study presented addresses, for the first time, the validity of PT readings based on digital images. These are increasingly used to document and even analyse PT reactions (3). In contrast to the few previous studies in this field (4, 5), a broad range of (slight) IR, doubtful to strong positive reactions was represented, similar to real life conditions. Negative reactions, which should be straightforward to diagnose, had not been considered. Moreover, a gold standard rating was available for the images, enabling us to assess validity, and not merely reliability (agreement independent from a gold standard).

However, before discussing single findings of our analysis, some general aspects should be addressed. First of all, compared with the real life application, the evaluation of digital images of PT reactions was different in several aspects:

- No information as to the allergen, and its concentration and vehicle was given – in case

Table 1. Characteristics of the images used and descriptive univariate results. The gold standard rating is marked with bold numbers. The percentage of exact ('ex') and gross ('gr') concordance between the readings blinded with regard to the test substance and the gold standard is given in the 2 rightmost columns

	PI	Gold standard rating	Reading						Percentage concord	
			–	IR	?	+	++	+++	ex	gr
<i>p</i> -Phenylenediamine 1%	Yes	+++	0	7	0	0	4	111	91	94
Fragrance mix 8%	Yes	+	1	5	7	92	17	0	75	89
Kerosene (as is)	Yes	IR	2	91	0	0	9	20	75	76
Shampoo 1% ^a	No	IR	17	96	9	0	0	0	79	100
Lyrall 5%	Yes	++	0	5	0	48	67	2	55	96
Detergent 10% ^a	Yes	IR	1	90	8	12	6	5	74	81
Fragrance Mix 8%	Yes	+	0	5	4	67	46	0	55	93
Bufexamac 5%	Yes	++	0	41	0	3	32	46	26	66
Cutting oil 4% ^a	Yes	+	2	7	57	54	2	0	44	46
Treemoss 1%	Yes	+	1	7	2	68	44	0	56	92
Disinfectant (as is)	Yes	IR	1	97	4	5	4	11	80	84
NaOH 5% ^a	No	IR	18	69	28	7	0	0	57	94
Potassium dichromate 0.5%	No	?	29	5	77	10	1	0	63	91
MDBGN 0.3%	Yes	++	0	1	0	29	83	9	68	99
Compositae mix 6%	Yes	++	0	3	0	5	106	8	87	98
Thiuram mix 1%	No	?	0	34	70	18	0	0	57	85
<i>p</i> -Phenylenediamine 1%	Yes	++	1	2	0	9	96	14	79	98
Potassium dichromate 0.5%	No	?	14	21	75	11	1	0	61	90
Nickel sulfate 5%	Yes	+	0	9	2	27	77	7	22	91
Benzalkonium chloride 0.1% ^a	No	?	26	5	82	8	1	0	67	93

All allergens in petrolatum, except where otherwise indicated: for example superscript 'a' indicates water; PI, palpable infiltration; and IR, irritant; MDBGN, methyldibromoglutaronitrile.

of several of the PT substances presented on the pictures, this information would most probably have led more raters to choose the gold standard rating, for example, with *p*-phenylenediamine (PPD) 1% (Fig. 1a), the gold

standard reading being +++, the vast majority rated the reaction accordingly (111 of 122), while 7 considered this to be IR. It is very probable that at least some of the 7 raters would have changed their IR rating

Table 2. Association between disagreement with the gold standard ('exact' and 'gross'; see text) and being a dermatologist and years of patch test experience, respectively, quantified as prevalence ratio (PR) with accompanying 95% confidence interval. A PR > 1 indicates more disagreement with the gold standard than in the reference category, a PR < 1 less disagreement, that is, more agreement when reading the reactions blind to the test substance

Substance	'Exact'			'Gross'		
	Derm.	Exp2	Exp1	Derm.	Exp2	Exp1
<i>p</i> -Phenylenediamine: +++	2.84 (0.38–21.2)	2.21 (0.59–8.23)	0.72 (0.13–4.07)	1.71 (0.21–13.56)	1.47 (0.35–6.16)	n.c.1
Fragrance mix +	0.32 (0.18–0.58)	0.60 (0.25–1.47)	1.27 (0.65–2.50)	0.24 (0.09–0.66)	0.55 (0.11–2.85)	1.88 (0.60–5.94)
Kerosene IR	0.69 (0.36–1.33)	1.65 (0.50–5.43)	5.64 (2.12–15.0)	0.75 (0.37–1.49)	1.65 (0.50–5.43)	5.11 (1.90–13.7)
Shampoo IR	1.19 (0.50–2.87)	1.65 (0.50–5.43)	4.30 (1.57–11.8)	n.c.1,2	n.c.1,2	n.c.1,2
Lyrall ++	1.95 (1.00–3.80)	0.65 (0.34–1.24)	1.77 (1.16–2.70)	1.14 (0.13–9.75)	n.c.1,2	n.c.2
Detergent IR	1.02 (0.49–2.09)	2.65 (1.02–6.84)	3.23 (1.29–8.06)	0.65 (0.30–1.42)	2.94 (0.84–10.3)	4.30 (1.31–14.1)
Fragrance mix +	1.28 (0.75–2.18)	0.75 (0.48–1.16)	0.56 (0.33–0.93)	2.27 (0.30–17.4)	1.10 (0.24–5.15)	1.08 (0.23–5.02)
Bufexamac ++	0.78 (0.65–0.94)	0.75 (0.59–0.97)	0.74 (0.57–0.95)	0.78 (0.45–1.33)	0.93 (0.56–1.54)	0.34 (0.15–0.76)
Cutting oil +	0.85 (0.60–1.20)	0.59 (0.37–0.96)	1.16 (0.84–1.59)	0.82 (0.58–1.16)	0.64 (0.39–1.06)	1.25 (0.90–1.75)
Treemoss +	0.99 (0.62–1.61)	0.74 (0.44–1.24)	0.97 (0.62–1.52)	2.56 (0.34–19.3)	0.55 (0.05–5.84)	3.76 (0.83–17.1)
Disinfectant IR	0.90 (0.40–2.03)	5.51 (1.29–23.6)	6.99 (1.68–29.1)	0.85 (0.34–2.13)	9.92 (1.32–74.8)	10.8 (1.44–80.2)
NaOH IR	0.87 (0.55–1.38)	1.22 (0.78–1.90)	0.74 (0.42–1.29)	0.71 (0.15–3.46)	1.10 (0.24–5.15)	0.36 (0.04–3.31)
Potassium dichromate?	0.70 (0.43–1.13)	0.95 (0.50–1.79)	1.46 (0.85–2.50)	0.50 (0.16–1.57)	0.83 (0.20–3.47)	1.08 (0.29–4.01)
MDBGN ++	0.64 (0.38–1.09)	0.55 (0.27–1.14)	1.01 (0.58–1.76)	n.c.1	n.c.1	n.c.1
Compositae mix ++	1.23 (0.38–4.01)	0.55 (0.15–2.06)	1.25 (0.46–3.42)	n.c.1	n.c.2	n.c.2
Thiuram mix?	1.19 (0.69–2.05)	0.83 (0.50–1.38)	0.91 (0.56–1.48)	4.83 (0.67–34.7)	1.76 (0.63–4.94)	1.08 (0.34–3.44)
<i>p</i> -Phenylenediamine: ++	0.77 (0.36–1.64)	0.55 (0.21–1.47)	1.18 (0.56–2.48)	0.57 (0.05–6.03)	n.c.1	2.15 (0.20–22.8)
Potassium dichromate:?	0.93 (0.55–1.57)	0.96 (0.54–1.71)	1.14 (0.67–1.94)	0.28 (0.10–0.81)	0.66 (0.17–2.59)	0.86 (0.25–2.98)
Nickel sulfate +	0.84 (0.70–1.00)	0.94 (0.72–1.21)	1.11 (0.90–1.37)	1.28 (0.29–5.57)	0.74 (0.13–4.17)	2.15 (0.58–8.03)
Benzalkonium chloride:?	0.66 (0.39–1.12)	1.10 (0.58–2.08)	1.16 (0.62–2.15)	0.57 (0.15–2.12)	0.83 (0.20–3.47)	0.54 (0.10–2.78)

Derm., dermatologist (reference: not a dermatologist); Exp., experience in patch testing ('Exp1' up to 5, 'Exp2' 6–15 years, reference: >15 years); n.c.1, not calculable: zero disagreement in category; n.c.2, not calculable: zero disagreement in reference category; n.c.1,2, not calculable: zero disagreement in both categories. MDBGN, methyldibromoglutaronitrile.

to a 'strong allergic' rating, if they had known that the test material was 1% PPD which is definitely not such a strong irritant at this concentration. The same holds true for bufexamac (Fig. 1h) – here, the sharp demarcation of a clearly infiltrated, vesicular reaction may have mislead many raters: 41 considered this to be IR, while in fact this was an allergic reaction of a ++ strength. The opposite can be observed with IR materials: the IR kerosene reaction (Fig. 1c) could also represent a severe bullous allergic reaction. In case of the IR reaction to 10% detergent (Fig. 1f) the reasons for divergence are less evident, and it is indeed somewhat disturbing that 23 rated this as allergic (+, ++ or +++) rather than IR. The lesion is rather sharply demarcated, has no papules or vesicles, and in the center a large pustule is visible. Possibly information on the presence of PI has mislead these readers, similar to Fig. 1k. Nevertheless, if information on the test material had been given, less false ratings could have been expected.

- Although information on the presence of PI (and, implicitly, on its absence) was provided with the images, many physicians probably only rely on their own palpation – otherwise it would be hard to explain why a number of raters judged images as doubtful (by definition erythema only!) where infiltration was present (most stunningly at Fig. 1i).
- Only single reactions, and not the entire back or larger parts of it were shown, excluding comparisons with other reactions on the same patients, or considerations regarding 'angry back' or cross-reactivity. However, while such information will often guide the reading and interpretation of PTs, the quiz was explicitly restricted to evaluating (and training) the correct morphological reading as such.
- This deliberate restriction also applies to other aspects of a full evaluation of PT reactivity: after morphological classification, which was the sole issue here, the interpretation of reactions as either allergic or IR follows, and, after this, the evaluation of clinical relevance if a reaction is deemed allergic (6). Sometimes, a reaction can only be classified as either allergic or IR by further testing with a dilution series or performing a repeated open application test or a suitable provocative use test (7).

Hence, with the PT quiz we evaluated only a part of the process of PT reading, albeit an

essential one and the addition of other pieces of information mentioned above should have improved both accuracy and validity.

Similar restrictions applied to the study published recently, evaluating agreement between 5 experienced dermatologists when rating slides of '55 positive PT reactions' (5). The authors concluded that by collapsing the original International Contact Dermatitis Research Group (ICDRG) classification with the 6 levels used here as well into a simpler scheme with 'positive', 'IR', and 'negative' reactions (the latter including doubtful reactions) increases agreement. We have noted a similar phenomenon comparing 'exact' and 'gross' agreement (see the 2 rightmost columns in Table 1) – the latter representing a dichotomous classification of allergic versus non-allergic, which we consider most important for diagnostic workup. However, for contact allergy research and also in patient care we do not think it advisable to abandon the existing grading system in favour of a simplified system, because the intensity of the test reaction does carry important information and should not be dismissed prematurely.

One previous study examined a set of selected live PT patients, taking partly similar measures to achieve blinding (no information on the allergen – out of a set of altogether 4 allergens – and no communication with patients and other raters, respectively) (4). While the results of this study are difficult to compare with ours because of methodological differences, one general result proved similar: Bruze et al. found agreement in 82% of reactions (in terms of a classification as positive versus negative), and in our analysis 'gross' agreement (in terms of a classification as positive versus non-positive) amounted to 87.8% (88.4% if the analysis was restricted to dermatologists).

No effort had been taken to calibrate colour balance, contrast, and brightness of the images, to render these as similar as possible to the original human skin. This was a deliberate decision, as the images are intended to be used as straightforward teaching material under a variety of viewing conditions as well. However, as 'IR' or '?' reactions had been rated as negative by some, that is, erythema was not recognized as such, it might be difficult to detect (slight) erythema on digital images. It is up to future research to further explore the impact of a possible loss of quality by digitalizing images on the recognition of erythema.

According to the results presented in Table 1, many images were correctly judged by the vast majority of raters – particularly if 'gross' agreement is considered (see above). Only 4 images did

not yield more than 100/122 'grossly agreeing' ratings (Fig. 1c, f, h, and i). In the first 3 pictures it seems to have been difficult to distinguish IR and (stronger) positive reactions. Those first 3 images, where disagreement with the gold standard reading cannot just be attributed to simple ignorance of information on PI as in the last instance, should probably be classified as 'difficult'. They could either be eliminated from teaching material, or deliberately included and extensively explained.

Two facets of the personal profile of raters have been assessed regarding their impact on disagreement with the gold standard:

- Dermatologists tended to rate more often in accordance with the gold standard; however, this was rarely a significant finding. In other instances they tended to rate worse than non-dermatologists (not significant). It should be mentioned, however, that even those who classified themselves as non-dermatologists had, partly extensive, experience in patch testing and were thus not comparable with non-dermatologists in general.
- In contrast, 'little' experience in patch testing was distinctly associated with misclassifying IR reactions: only the IR reaction to NaOH, with its special appearance (punctate crusted erosions without conceivable erythema), did not yield a significant result in this respect. Considering the fact that 'little' experience was defined data adaptively here, based on a highly specialized scientific peer group taking part in the quiz, it is evident that among less specialized colleagues with truly 'little' experience the misclassification of IR reactions may be a common problem.

Conclusion

The present study explicitly addresses just 1 particular 'isolated' step in the diagnostic process of patch testing, namely, the classification of PT reactions (as digital images) according to morphological criteria. Hence, the performance of (a fairly selected group of) raters in terms of agreement with an expert gold standard cannot be directly interpreted as giving an impression of the performance under real life conditions. Nevertheless, the (very) good agreement in case of many of the images indicates that generally the reading of PT reactions is highly valid, even under the somewhat artificial and difficult conditions of the PT quiz. Our analysis identified the distinction between stronger IR and (strong) positive reactions, and the distinction between doubtful and weak posi-

tive reactions as targets for training. Regular PT training, for instance using the digital images employed in the quiz, further supplemented with information on the allergen, on the correct classification and other explanations can be expected to enhance the standardization of patch testing. This appears highly recommendable for observers engaged in multicentre trials or participants of regular contact dermatitis research networks, such as the European Surveillance System on Contact Allergies (<http://www.essca-dc.org>) to reduce between-centre variation induced by differing PT reading standards. Scientific meetings or the Internet could provide a suitable platform for this educational activity. The images presented here are available at <http://www.ivdk.gwdg.de/dkg> (select 'PT Reading Course'), accessible after registration.

Acknowledgement

Figures 1f, 1k, and 1p were kindly provided by Thomas Fuchs, Göttingen; they are reprinted by permission from Dustri Verlag, Deisenhofen, Germany. Figures 1r and 1t were kindly provided by Jürgen Kreusch, Lübeck, Germany. The study set up was generously sponsored by Hermal/Trolab, Reinbek, Germany. The authors thank all participants of the quiz.

References

1. Wahlberg J E, Lindberg M. Patch testing. In: *Contact Dermatitis*, Frosch P J, Menné T, Lepoittevin J P (eds): Berlin, Springer, 2006: 365–390.
2. Fregert S. Manual of contact dermatitis. *Munksgaard, Copenhagen* 1981.
3. Schuster C, Kranke B, Szolar-Platzer C, Aberer W. Macro-video documentation patch tests. *Contact Dermatitis* 2005; 52: 177–183.
4. Bruze M, Isaksson M, Edman B, Bjorkner B, Fregert S, Moller H. A study on expert reading of patch test reactions: inter-individual accordance. *Contact Dermatitis* 1995; 32: 331–337.
5. Ivens U, Serup J, O'goshi K. Allergy patch test reading from photographic images: disagreement on ICDRG grading but agreement on simplified tripartite reading. *Skin Res Technol* 2007; 13: 110–113.
6. Fischer T. The art of patch testing. *Contact Dermatitis* 1990; 23: 221–223.
7. Johansen J D, Frosch P J, Menné T. Allergic contact dermatitis in humans – experimental and quantitative aspects. In: *Contact Dermatitis*, Frosch P J, Menné T, Lepoittevin J P (eds): Berlin – Heidelberg, Springer, 2006: 189–198.

Address:

Wolfgang Uter, MD
Department of Medical Informatics,
Biometry and Epidemiology
University of Erlangen/Nürnberg
Waldstr 6, D-91054 Erlangen
Germany
Tel: +49 9131 8522720
Fax: +49 9131 8522721
e-mail: wolfgang.uter@rzmail.uni-erlangen.de