

Sensitivity, specificity and positive predictive value of patch testing: the more you test, the more you get?

THOMAS L. DIEPGEN¹ AND PIETER JAN COENRAADS²

¹Department of Social Medicine, Center of Occupational and Environmental Dermatology, Ruprecht-Karls-University Heidelberg, Germany

²Occupational & Environmental Dermatology Unit, University Hospital Groningen, The Netherlands

On behalf of the ESCD Working Party on Epidemiology

Pathophysiological variability affects the results of patch testing. In addition, even a minimal degree of test-imprecision due to this variability has a number of important statistical consequences for the analysis and interpretation of any patch test data set. One such statistical phenomenon that is often overlooked is the dependance of the positive predictive value (i.e., the predictive value of a positive patch test) on sensitivity and specificity, the impact of which is heavily dependent on the proportion of truly allergic subjects that are studied. A 2nd important issue is the fact that patch testing is performed in series, which means multiple tests. If we assume, for example, a patch test series of only 10 allergens, then it can be demonstrated that there is a random probability of over 40% to find, simply by chance, for at least 1 allergen, a statistically significant difference between 2 groups of patients. Comparison of the results of series between patients calls for statistical adjustments in order to prevent erroneously positive differences and/or associations.

Key words: allergic contact dermatitis; sensitivity; specificity; false positive; diagnostic test; statistics; misclassification. © Munksgaard, 2000.

Accepted for publication 11 January 2000

Patch testing is a well-established method to determine whether sensitization to certain agents has occurred. The question as to whether agents that are positive are causally linked to contact dermatitis has several pitfalls: patch testing carries the risk of irritant reactions, false-positive results, and difficulties in interpretation of the clinical relevance of a positive finding. The accuracy of this diagnostic procedure depends on the experience, knowledge and skill level of the physician who performs the test. Even when an allergic reaction is found, it is sometimes not certain whether the contact dermatitis is of allergic origin or not. Clinically irrelevant reactions must be distinguished from false-positive patch test results. The magnitude of the problem of false-positive (or false-negative) patch test reactions is unknown and scarcely mentioned in the literature. In this paper we will describe some statistical aspects of patch testing as a diagnostic tool, which, like most laboratory tests, has a certain degree of imprecision. These well-known statistical phenomena are frequently over-

looked in studies on series of patients with contact dermatitis.

Prevalence Influences the Positive Predictive Value

It was reported by Nethercott (1) that the sensitivity and specificity of patch testing are approximately 70%. The relevance of a positive test was reported to be 50%, which means that in only 1/2 the cases can the substance inducing a patch test response be established as the cause of the patient's skin disease. To ascertain the validity of a diagnostic procedure such as patch testing, the terms "sensitivity", "specificity" and "predictive value" are used. The sensitivity indicates the *probability* that cases with allergic contact dermatitis (ACD) are correctly diagnosed, the specificity that the non-ACD cases are correctly classified. From a clinical point of view, however, it is more appropriate to calculate the positive predictive value (PPV), which is the proportion of individuals who *actually have*

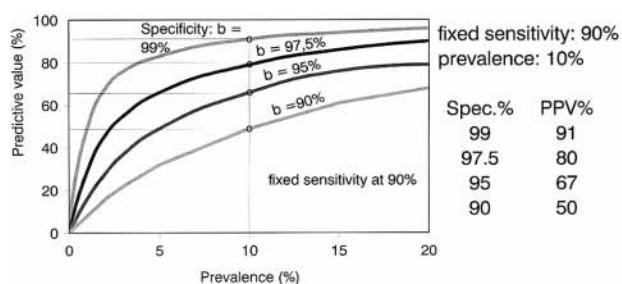


Fig. 1. The positive predictive value of a diagnosis of allergic contact dermatitis (ACD) is a function of the true prevalence of ACD, the patch test specificity and the patch test sensitivity.

ACD among those diagnosed as such by the diagnostic test that was used.

The PPV is a function of the prevalence of ACD in the population, and of the sensitivity and specificity of the patch test. In Fig. 1, the positive predictive value is shown as a function of the prevalence, assuming a fixed sensitivity of 90% and 4 different specificities: 99%, 97.5%, 95%, and 90% respectively. If, for example, the prevalence of ACD due to an allergen were 10%, and the sensitivity 90%, then the PPV would be 91% if the specificity were 99%. The PPV would decrease to 80% if the specificity were 97.5%, 67% if the specificity were 95%, and 50% if the specificity were 90%.

An example

If we assume that the prevalence of being sensitive to nickel is 10%, and the sensitivity and specificity of patch testing to be 90%, then, from a statistical point of view (Fig. 1), a positive reaction results in only 50% of the cases being diagnosed correctly. Therefore, almost invariably individuals are missed who do have an ACD, while others are wrongly diagnosed as cases of ACD. In Table 1 the magnitude of this problem is shown on the assumption that the sensitivity and specificity of patch testing were 90% and that the prevalence of sensitization to various allergens ranges from 1% to 50% in a sample of 1000 patients tested during a specific time period. The resulting false-positive patch test results would misdiagnose between 99 and 50 patients. If the true prevalence were 1% (not uncommon for many allergies), then out of 10 truly sensitized patients you will correctly diagnose 9, but out of 990 in the truly non-sensitized patients you will have 99 false-positive patch test results. Even if we assume a true prevalence of 50%, the expected number of false positives would be 50 out of 500 non-sensitized patients.

This example demonstrates how important it is to achieve a high prevalence rate of truly sensitized patients in the clinical setting in order to reduce

Table 1. The number of true-positive and false-positive patch test results according to different prevalence rates of sensitization assuming a sensitivity and specificity of 90%; numbers are given on the assumption that 1000 patients were patch tested

	Prevalence of ACD				
	1%	5%	10%	20%	50%
patients with ACD	10	50	100	200	500
patients without ACD	990	950	900	800	500
total no. of patients	1000	1000	1000	1000	1000
no. of true positives	9	45	90	180	450
no. of false positives	99	95	90	80	50

Table 2. Statistical significance can be influenced by sensitivity and specificity; the upper table shows a statistically significant association between vesicular hand eczema and nickel allergy; the lower table shows the effects of misclassification, assuming a sensitivity and specificity of 90% for patch testing (worst case scenario)

	Vesicular hand eczema		Total
	yes	no	
nickel allergy yes	11	9	20
nickel allergy no	19	61	80
total	30	70	100

χ^2 test: $\chi^2=7.44$; $p=0.006$

Sensitivity and specificity 90%

	Vesicular hand eczema		Total
	yes	no	
nickel allergy yes	9	17	26
nickel allergy no	21	53	74
total	30	70	100

χ^2 test: $\chi^2=0.39$; $p=0.549$ indicating no significant association.

the number of false-positive patch test results. Thus, simply from a statistical point of view, it is crucial to explore the patient's history carefully and exactly before performing patch testing; indiscriminate testing of many patients with a doubtful allergic origin of their skin problem (i.e., a low prevalence of true allergies) will lead to many cases of wrongly diagnosed contact dermatitis.

The Effect of Sensitivity and Specificity on p -values

A 2nd example demonstrates other possible pitfalls associated with sensitivity and specificity. Suppose that we would like to study whether there is a statistically significant association between vesicular hand eczema and sensitization to nickel. We perform patch tests in the next 100 patients with hand dermatoses coming to our outpatient clinic. In 30% of these patients we diagnose a vesicular type of hand dermatitis, and find a positive patch test in 20%. The results, shown in Table 2, demonstrate a statically highly significant association be-

tween vesicular hand eczema and nickel allergy. But, if we assume that the sensitivity and specificity of patch testing is 90%, then we will expect 10% false-positives and 10% false-negative patch test results. This would mean that 2 out of the 20 patients with a positive patch test to nickel are in reality not sensitized, and 8 out of the 20 patients with negative patch test results in fact having a nickel allergy. In the worst case, if the 2 false positives had vesicular hand eczema, and the 8 false negatives did not have dyshidrotic hand eczema, this would result in a chi-square test statistic of 0.36. With a corresponding p -value of 0.549, this would mean no statistically significant association between vesicular hand eczema and nickel allergy.

In the example described above, we are dealing with a common measurement error of categorical data, which is well known in the statistical literature as a phenomenon of misclassification (2): the occurrence of this is the rule rather than the exception. The example shows the consequences when representing our contact dermatitis data in a 2×2 table. The possible consequences are widely discussed in the statistical literature (3–5). Various techniques to deal with this problem, other than just increasing the sample size, have been developed over recent decades. Nevertheless, insight into the degree of false-positive and false-negative results of patch tested allergens is desirable. A sensitivity of 90%, as assumed in the above-mentioned example, is not unrealistic. In a recent publication on the reproducibility of patch tests, a negative test on one side with a positive test on the opposite side were recorded in 8% of patients (6). Discordance, which is obviously not the same as sensitivity but only indicative, up to 15% has been reported (7).

The Problem of Multiple Tests and Comparisons

Patch testing is always a matter of multiple tests, whereby each individual patch test in the series has a certain degree of imprecision. Normally, we want to know which patch test results are associated with a 3rd factor such as a disease, sex, occupation, etc. Therefore, we have to do some pairwise comparisons of the frequencies of the patch test results in the 2 groups of patients (e.g., between males and females, or between carpenters and bricklayers). The simplest approach is to do a series of chi-square-tests, 1 on every allergen patch tested. Unfortunately, this procedure increases the probability of a false rejection of the null hypothesis. In our case, this would lead to an erroneous claim of difference in sensitization rates between the 2 groups, while there is in fact no difference. If we assume a patch test series of only 10 allergens, and a p -value of 0.05 being statistically significant, then there is a random

probability of over 40% of finding, simply by chance, for at least 1 allergen a statistically significant difference between the 2 groups.

A variety of methods has been proposed to handle this problem. The most simple, though also the least powerful among these, is the Bonferroni correction. In this procedure the desired level of significance for the comparison of the overall patch-test series results is divided by the number of (patch) tests carried out in the experiment. Pairwise χ^2 tests on the individual patchtests are calculated with the corrected significance levels. More powerful methods exist, with increased statistical complexity (8).

This complexity is not of trivial importance because the individual patch tests that make up the series are not completely independent of each other; between some there is, in statistical terms, a high degree of collinearity (e.g., chromium and nickel).

Conclusions

Although this overview of statistical problems with the analysis of patch test data is far from complete, the above-mentioned points might, in addition to pathophysiological considerations, be helpful in interpreting and analysing the results of routinely or experimentally-collected patch-testing data.

References

1. Nethercott J R. Practical problems in the use of patch testing in the evaluation of patients with contact dermatitis. *Curr Probl Dermatol* 1990; 2: 4.
2. Agresti A. *Categorical data analysis*. Wiley, New York 1990.
3. Newell D J. Misclassification in 2×2 tables. *Biometrics* 1963; 19: 187–188.
4. Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. *Journal of Clinical Epidemiology* 1988; 41: 1167–1174.
5. Brenner H. Notes of the assessment of trend in the presence of nondifferential exposure misclassification. *Epidemiology* 1992; 3: 420–427.
6. Bourke J F, Batta K, Prais L, Abdullah A, Foulds I S. The reproducibility of patch tests. *Br J Dermatol* 1999; 140: 102–105.
7. Brasch J, Henseler T, Aberer W et al. Reproducibility of patch tests. *J Am Acad Dermatol* 1994; 31: 584–591.
8. Kuss O, Diepgen T L. Proper statistical analysis of trans-epidermal water loss (TEWL) measurements in bioengineering studies. *Contact Dermatitis* 1998; 39: 64–67.

Address:

Thomas L. Diepgen
University Hospital Heidelberg
Department of Social Medicine, Occup. & Environmental Dermatology
Bergheimer Str. 58
D-69115 Heidelberg
Germany