

Positive reactions to pairs of allergens associated with polysensitization: analysis of IVDK data with machine-learning techniques

Werner Adler, Olaf Gefeller and Wolfgang Uter

Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University of Erlangen/Nürnberg, 91054 Erlangen, Germany

doi:10.1111/cod.12706

Key words: contact allergy; epidemiology; machine-learning; polysensitization.

Correspondence: Wolfgang Uter, Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Waldstr. 4-6, D-91054 Erlangen, Germany. Tel: +49 9131 8522750; Fax: +49 9131 8522721. E-mail: wolfgang.uter@imbe.med.uni-erlangen.de

Conflicts of interest: W. Uter has accepted honoraria for presentations or travel reimbursement from cosmetic industry associations, and a lecture fee from Almirall/Hermal. The other authors do not declare a conflict of interest.

A varying proportion of patients has at least one positive reaction to allergens of the (European) baseline series, for example between 33.3% and 61.7% [5% and 95% percentiles of European Surveillance System on Contact Allergies departments, 2009–2012 (1)]. Often, only one single positive reaction is observed (2). However, it is not entirely uncommon to observe multiple positive reactions to the baseline series, even to three or more different allergens. The latter phenomenon has been

termed 'polysensitization' (PS) (3), and is considered to be a phenotypic marker of increased susceptibility to contact sensitization. Previous analyses have indicated which single allergens are associated with PS (2); the present analysis addresses the novel question of whether certain 'pairs' of positive reactions to allergens may be particularly associated with PS. As the number of such pairs is high, even considering the baseline series with n allergens, namely, $n \times (n - 1)/2$, modern biostatistical techniques of variable selection rooted in machine-learning methodology have to be employed to identify a set of explanatory factors (allergen pairs, but also clinical and demographic information), which can then be used in a regression model to quantify the association of these factors with PS.

Methods

The Information Network of Departments of Dermatology (IVDK), dedicated to the clinical surveillance of contact allergy, has been described elsewhere (4). Briefly, the participating departments are predominantly located in Germany ($n = 59$), but also in Switzerland ($n = 4$) and Austria ($n = 1$), are members of the German contact dermatitis research group, and adhere to international patch test guidelines (5). The members transmit anonymized patch test data along with clinical data and items of the patient's history twice yearly to the data centre in Göttingen. Between 1992 and 2014, 230 316 consultations were documented; in cases of multiple consultations by 1 patient, only the first consultation was considered. For the present analysis, only patients patch tested with the baseline series were included. However, the composition of the baseline series had changed several times during the study period. For the present analysis, a 'common denominator' set of allergens was chosen ($n = 24$) that was regarded as a good compromise in terms of including a sufficient number of cases to achieve sufficient statistical power, and addressing an adequate number of allergens that can be considered to be fairly representative of a baseline series (Table S1). For the purpose of the present analysis, reactions to the following allergens were considered to be equivalent: (i) fragrance mix II and hydroxyisohexyl 3-cyclohexene carboxaldehyde, (ii) mercapto mix (three or four components) with 2-mercaptobenzothiazole, and (iii) methyldibromo glutaronitrile (MDBGN, 0.2 or 0.3% pet.) or MDBGN with 2-phenoxyethanol (Euxyl™ K 400, 1% pet.). Patch test reactions were dichotomized into positive [if +, ++ or +++ on day (D) 3, or, exceptionally, if no D3 reading was available, on D4] versus non-positive (negative, ?+, or irritant reaction). Anatomical sites were categorized into five groups as trunk, hand(s)/arm(s),

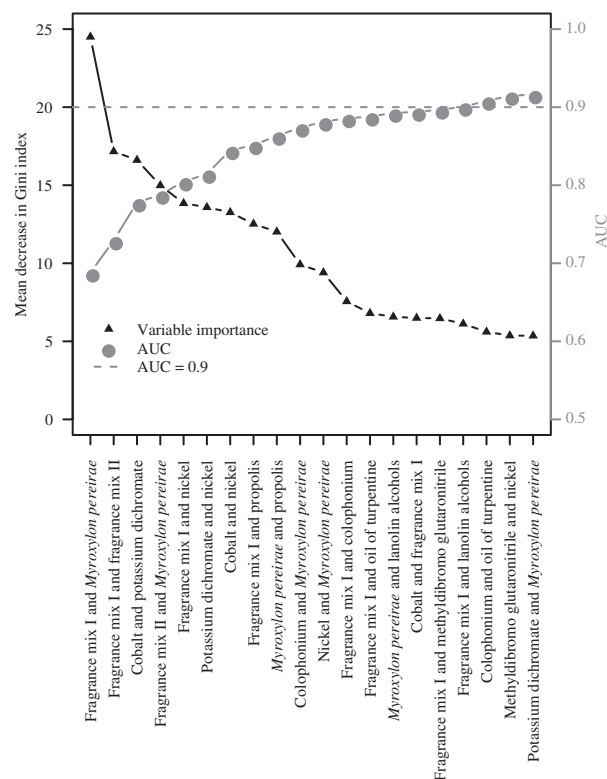


Fig. 1. The 20 most important pairs of allergens associated with polysensitization, ordered by importance from left to right. The mean decrease in the Gini index, a measure for the importance of the variables in producing purer splits in the random forest, and thus for a variable's importance, is shown as black triangles, decreasing from the left to the right variable (left y-axis). To determine the final logistic regression model, areas under the receiver operating characteristic curve (AUCs) for a growing number of allergen pairs were calculated, leading to increasing AUC values from only one variable ('fragrance mix I and *Myroxylon pereirae*') to 20 allergen pairs, including 'potassium dichromate and *Myroxylon pereirae*', shown in this plot. The AUCs are indicated by grey circles (right y-axis). For an explanation of the statistical methods, see the text and Table S1.

foot/feet/leg(s), face/head, and remaining categories or missing. Eventually, 105 325 patients in whom all 24 allergens had been tested were included in the analyses.

Statistical methods were used first to address the selection of meaningful, important variables. To this end, the complete dataset consisting of 105 325 observations was split into a tuning dataset ($n = 35\,294$) and a validation dataset ($n = 70\,031$). The importance of allergens for PS was estimated by first training a random forest (6) on the tuning dataset. A random forest is a classification ensemble consisting of several hundred classification trees, in our case 500. The trees are trained by use of a subset of the dataset, so that different trees are constructed.

Typically, bootstrap samples are drawn to obtain different subsets. As the tuning dataset was comparatively large, to speed up the training process, subsets consisting of a randomly selected 10% of the available observations were constructed for training of the classification trees. To increase the difference between trees, and thus to increase ensemble diversity, which is deemed to be important for the performance of the ensemble (6), at each split of a single tree only a random subset of all available variables is considered. As split criterion, the so-called Gini index was used. This index, which generally quantifies inequality of a distribution, is a measure of the purity of the split in the present context, that is, the proportions of classes after the split: a pure split leads to observations belonging to only one class of the outcome (PS) in one node and to observations belonging to the other class in the other node, and thereby to a Gini index close to 0. A way to determine the importance of a specific variable for the performance of the random forest is to sum up the decrease of impurity at each split where this variable is used. The estimation of the importance for each variable in the dataset allows for a ranking of these variables by their importance. The quality of the random forest regarding classification accuracy is given by the so-called out-of-bag estimation (Appendix S1).

The number and list of variables (allergen pairs) for a logistic regression model were determined by sequential assessment of the performance of several models trained with 1–50 of the most important allergens as determined by random forest variable importance, together with the clinical variables sex, age and site, by 10-fold cross-validation. The most parsimonious model achieving an area under the receiver operating characteristic curve (AUC) of >0.90, as a stop criterion, determined the optimal number of allergens. For a more detailed description, see Appendix S1.

The pairs of allergens identified in the previous step, along with age, sex and sites as described above, were used in a logistic regression analysis on the validation dataset as explanatory factors for the dichotomous outcome PS, that is, positive reactions to three or more allergens versus no, one or two positive reactions to the baseline series. The association was quantified with odds ratios accompanied by 95% confidence intervals calculated with the profile likelihood method. For statistical analyses, the software package R was used (version 3.2.2, www.r-project.org/, package randomForest).

Results

With the sequential approach used, 17 pairs of allergens, in addition to the basic variables, sufficed to achieve an

AUC of >90%. The mean decrease in the Gini coefficient and the increase in the AUC associated with adding the additional pairs are shown in Fig. 1; it can be seen that the addition of further variables beyond the 17 (e.g. the rightmost three pairs) would further increase the accuracy of prediction, but at the expense of making the model less parsimonious.

The selected variables were used in a logistic regression analysis in a second step (Table 1). Whereas sex was not associated with PS in this analysis, a weak but significant association with age was found, and anatomical sites previously identified as being associated with PS were confirmed, namely, hand/arm and foot/leg dermatitis. However, the novel approach of this analysis was to incorporate pairs of allergens as predictors of PS. Pairs of fragrances (if propolis and colophonium are considered to be at least 'fragrance-related'), metals and combinations of these two groups of allergens were the main driving force of PS; only lanolin alcohols, on the 14th rank, and MDBGN, on the 16th rank, are included as important other predictors.

Discussion

The present analysis unifies two different perspectives: primarily, the interest in factors associated with PS, similarly to previous analyses (2, 3), but also the role of associations between pairs of allergens, as first examined in (7). Evidently, only pairs that are associated to some extent and are thus also relatively common as a pair, and furthermore in which at least one component is associated with another contact allergy, or other contact allergies, will have sufficient leverage to be associated with PS, that is, the occurrence of at least one further reaction, beyond those two defining the pair.

The canopy of mutual associations, which is particularly strong in the case of fragrance mix I (7), is evident from the fact that this allergen occurs in nine of 17 selected pairs (Table 1). This is in contrast to nickel, which is more of a solitary allergen (7), and which is involved in just four pairs (Table 1), even though it is the most common allergen, with a prevalence of 14.3% in univariate analysis (Table S1). As a consequence, the importance of fragrance mix I for PS is much higher than that of nickel.

In conclusion, the present analysis provides a unique view on PS, which is conventionally defined as the occurrence of at least three positive reactions to the (European) baseline series (3), based on state-of-the-art statistical methodology concerning variable selection. Thereby, it became apparent that (combinations of/with) fragrances are the main driving forces of multiple sensitization, possibly because exposure to fragrances is both ubiquitous and multiple (8).

Table 1. Results of a logistic regression analysis with polysensitization as outcome (prevalence, 8.6%) (reference: no, one or two positive reactions to baseline series allergens) and the variables shown in the table as explanatory factors

Variable	%	OR	95%CI
Basic clinical/demographic data			
Sex: female (ref.: male)	62.3	0.99	0.91–1.09
Age (per year); mean, 50.7; standard deviation, 18.3	–	1.01	1.01–1.01
Site: trunk	16.4	1.00	Reference
Site: hand(s) or arm(s)	31.6	1.60	1.40–1.82
Site: foot/feet or leg(s)	16.0	1.37	1.18–1.59
Site: face/head	20.6	0.90	0.77–1.05
Site: 'other' or missing	15.4	0.72	0.61–0.85
Pairs of allergens			
Fragrance mix I and <i>Myroxylon pereirae</i>	2.7	17.3	14.2–21.1
Fragrance mix I and fragrance mix II	2.1	25.4	21.1–30.5
Cobalt chloride and potassium dichromate	1.5	39.7	32.5–48.5
Fragrance mix II and <i>Myroxylon pereirae</i>	1.1	34.2	24.0–48.8
Fragrance mix I and nickel sulfate	1.5	18.1	14.0–23.4
Potassium dichromate and nickel sulfate	1.3	34.0	26.8–43.0
Cobalt chloride and nickel sulfate	3.0	19.4	16.7–22.4
Fragrance mix I and propolis	0.96	38.9	26.4–57.4
<i>Myroxylon pereirae</i> and propolis	1.2	34.2	26.0–45.0
Colophonium and <i>Myroxylon pereirae</i>	1.1	46.2	34.9–61.4
Nickel sulfate and <i>Myroxylon pereirae</i>	1.3	23.8	18.6–30.3
Fragrance mix I and colophonium	0.90	36.9	24.2–56.3
Fragrance mix I and oil of turpentine	0.61	44.7	27.8–73.0
<i>Myroxylon pereirae</i> and lanolin alcohols	0.74	42.8	30.3–60.7
Cobalt chloride and fragrance mix I	0.74	35.8	22.2–58.1
Fragrance mix I and methylidibromo glutaronitrile	0.74	44.6	30.4–65.5
Fragrance mix I and lanolin alcohols	0.74	43.5	29.6–64.4

CI, confidence interval; OR, odds ratio.

The variables in the lower part are ordered according to their importance (see Fig. 1).

Acknowledgements

The following members of the IVDK contributed data to this analysis. Aachen: H. Dickel, S. Erdmann, C. Schröder, and J. M. Baron. Aarau: J. Grabbe. Augsburg: A. Ludwig. Basel: A. Bircher. Berlin B.-Franklin: R. Treudler, B. Tebbe, and M. Worm. Berlin BWK: A. Köhler. Berlin Charité: B. Laubstein, T. Zuberbier, and M. Worm. Bern: D. Simon. Bielefeld: I. Effendy. Bochum: H. Dickel. Bochum BGFA: M. Fartasch. Bremen: D. Meyersburg. Buxtehude: P. Große-Hüttmann, P. Hausenblas, and E. Wrage-Brors. Dermatologikum Hamburg: K. Reich, V. Martin, and K. Breuer. Dessau: U. Lippert and A. Jung. Dortmund: B. Mydlach, B. Pilz, P. J. Frosch, C. Pirker, R. Herbst, K. Kügler, and U. Beiteke. Dresden: G. Richter, R. Aschoff, P. Spornraft-Ragaller, and A. Bauer. Dresden Friedrichst.: A. Koch. Erlangen: K.-P. Peters, T. L. Diepgen, M. Fartasch, M. Hertl, and V. Mahler. Essen: J. Schaller, H.-M. Ockenfels, and U. Hillen. Falkenstein: H. Schwantes. Freudenberg: Ch. Szliska. Gera: J. Meyer, H. Grunwald-Delitz, and M. Kaatz. Göttingen: Th. Fuchs and J. Geier. Graz: B. Kränke and W. Aberer. Greifswald: M. Jünger. Halle: G. Gaber,

D. Lübke, and B. Kreft. Hamburg, M. Kiehn, D. Vieluf, R. Weißbecher, E. Coors, and J. Witte. Hamburg BUK: C. Schröder-Kraft, K. Breuer, and U. Seemann. Hannover: Th. Werfel and T. Schaefer. Heidelberg: A. Schulze-Dirks, M. Hartmann, U. Jappe, and K. Schäkel. Heidelberg AKS: H. Dickel, T. L. Diepgen, and E. Weisshaar. Heilbronn: H. Löffler. Homburg/Saar: F. A. Bahmer, P. Koch, and C. Pföhler. Jena: M. Gebhardt, A. Bauer, W. Wigger-Alberti, M. Kaatz, and S. Schliemann. Kiel: J. Brasch. Krefeld: S. Wassilew, M. Lilie, and A. Wallerand. Lausanne: C. Curdin, P. Spring, and J. di Lucca. Leipzig: R. Treudler. Lippe Detmold: St. Nestoris. Lübeck: K. P. Wilhelm, J. Kreusch, J. Grabbe, I. Shimanovich, and U. Jappe. Mainz: D. Becker. Mannheim: Ch. Bayerl, D. Booken, H. Kurzen, C.-D. Klemke, and W. Ludwig-Peitsch. Marburg: I. Effendy, H. Löffler, M. Hertl, and W. Pfützner. Minden: J. Hoffmann and R. Stadler. München LMU: P. Thomas, B. Przybilla, T. Oppel, R. Eben, T. Schuh, and S. Molin. München Schwabing: M. Agathos, K. Ramrath, M. Georgi, and G. Isbary. München TU: J. Rakoski and U. Darsow. Münster: R. Brehler, B. Hellweg, and M. Behring. Nürnberg: I.

Müller, D. Debus, A. Bachtler, and K. Ertner. Oldenburg: M. Padeken. Osnabrück: W. Uter, S. M. John, N. Schürer, H. J. Schwanitz (†), H. Dickel, and Ch. Skudlik. Rostock: H. Heise (†), Ch. Schmitz, J. Trcka, and M. A. Ebisch. Stuttgart: T. Plaza, G. Nist, and J. Rieker-Schwiebacher. Tübingen: G. Lischka, M. Röcken, T. Biedermann, and J. Fischer. Ulm: H. Gall (†), G. Staib, and P. Gottlöber. Ulm BWK: H. Pillekamp. Ulm Univ.-Klinik: R. Hinrichs, G. Staib, and J. Weiss. Würzburg: J. Arnold and A. Trautmann. Zürich: B. Ballmer-Weber. Zwickau: B. Knopf, D. Teubner, and D. Mechtel.

References

- 1 Uter W, Gefeller O, Giménez-Arnau A et al. Characteristics of patients patch tested in the European Surveillance System on Contact Allergies (ESSCA) network, 2009–2012. *Contact Dermatitis* 2015; **73**: 82–90.
- 2 Schwitulla J, Gefeller O, Schnuch A, Uter W. Risk factors of polysensitization to contact allergens. *Br J Dermatol* 2013; **169**: 611–617.
- 3 Carlsen B C, Andersen K E, Menné T, Johansen J D. Characterization of the polysensitized patient: a matched case–control study. *Contact Dermatitis* 2009; **61**: 22–30.
- 4 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.
- 5 Johansen J D, Aalto-Korte K, Agner T et al. European Society of Contact Dermatitis guideline for diagnostic patch testing – recommendations on best practice. *Contact Dermatitis* 2015; **73**: 195–221.
- 6 Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
- 7 Brasch J, Uter W, Geier J, Schnuch A. Associated positive patch test reactions to standard contact allergens. *Am J Contact Dermat* 2001; **12**: 197–202.
- 8 Uter W, Yazar K, Kratz E-M et al. Coupled exposure to ingredients of cosmetic products: I. Fragrances. *Contact Dermatitis* 2013; **69**: 335–341.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Set of allergens included in the German baseline series during the study period, simultaneously patch tested in 105 325 patients. All allergens in pet., except where indicated otherwise.

Appendix S1. Description of the approach to variable selection.