IBRO Reports 6 (2019) 64–73

Contents lists available at ScienceDirect

IBRO Reports

journal homepage: www.elsevier.com/locate/ibror

Research Paper

A machine-learned analysis suggests non-redundant diagnostic information in olfactory subtests

Jörn Lötsch[a,b,*], Thomas Hummel[c]

[a] Institute of Clinical Pharmacology, Goethe - University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany
[b] Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor – Stern - Kai 7, 60590 Frankfurt am Main, Germany
[c] Smell & Taste Clinic, Department of Otorhinolaryngology, TU Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

ARTICLE INFO

Keywords:
Machine-learning
Data science
Human olfaction
Olfactory diagnostics
Patients
Anosmia

ABSTRACT

Background: The functional performance of the human sense of smell can be approached via assessment of the olfactory threshold, the ability to discriminate odors or the ability to identify odors. Contemporary clinical test batteries include all or a selection of these components, with some dissent about the required number and choice.

Methods: Olfactory thresholds, odor discrimination and odor identification scores were available from 10,714 subjects (3662 with anomia, 4299 with hyposmia, and 2752 with normal olfactory function). To assess, whether the olfactory subtests confer the same information or each subtest confers at least partly non-redundant information relevant to the olfactory diagnosis, we compared the diagnostic accuracy of supervised machine learning algorithms trained with the complete information from all three subtests with that obtained when performing the training with the information of only two or one subtests.

Results: The training of machine-learned algorithms with the full information about olfactory thresholds, odor discrimination and odor identification from 2/3 of the cases, resulted in a balanced olfactory diagnostic accuracy of 98% or better in the 1/3 remaining cases. The most pronounced decrease in the balanced accuracy, to approximately 85%, was observed when omitting olfactory thresholds from the training, whereas omitting odor discrimination or identification was associated with smaller decreases (balanced accuracies approximately 90%).

Conclusions: Results support partly non-redundant contributions of each olfactory subtest to the clinical olfactory diagnosis. Olfactory thresholds provided the largest amount of non-redundant information to the olfactory diagnosis.

Introduction

The functional performance of the human sense of smell is often analyzed via (i) assessment of the lowest concentration at which an odor can be smelled, which is the odor threshold (Cain et al., 1988a), (ii) assessment of the ability to differentiate smells, which is the ability of odor discrimination (Cain and Krause, 1979), and (iii) assessment of the ability to name or associate an odor, which is odor identification. While a more comprehensive assessment by inclusion of measurements of olfactory recognition, magnitude estimation and hedonics (Henkin et al., 2016, 2013) has been suggested, most comprehensive psychophysical test batteries aimed at assessing a subject's olfactory performance in clinical practice focus on olfactory thresholds, odor discrimination ability and odor identification performance (Hummel et al., 1997b).

Contemporary test batteries include either all components or a subset of olfactory tests (Cain et al., 1988b; Doty and Agrawal, 1989; Lam et al., 2006; Thomas-Danguin et al., 2003), i.e., they differ in the number of components of olfactory function included. The most frequent choice is the sole assessment of a subject's odor identification performance (Cain, 1979; Doty et al., 1984). This is also pursued in the majority of short tests, developed to provide a quick band easily obtained estimate of a subject's olfactory function (Hummel et al., 2010; Jackman and Doty, 2005; Lötsch et al., 2016b; Mueller and Renner, 2006; Toledano et al., 2009). Tests using a single other olfactory functional dimensions are rarely found in clinical routine, at least when considering the last 20 years (e.g., (Davidson and Murphy, 1997; Yilmaz et al., 2017), for review see (Doty, 2015).

While it has been pointed out that olfactory tests, for all practical purposes, measure a common source of variance (Doty et al., 1994b),

* Corresponding author at: Goethe - University, Theodor - Stern - Kai 7, 60590 Frankfurt am Main, Germany.
E-mail address: j.loetsch@em.uni-frankfurt.de (J. Lötsch).

https://doi.org/10.1016/j.ibror.2019.01.002
Received 17 September 2018; Accepted 5 January 2019

many test batteries use more than one component. A recent position paper on olfactory dysfunction (Hummel et al., 2017) suggested that "Psychophysical assessment tools used in clinical and research settings should include tests of odor threshold, and/or one of odor identification or discrimination. Ideally, however, testing should include two or three of these subcomponents", still many authors use only tests with a single component (Doty, 2015) or even the use of visual analogue scales is discussed for the assessment of olfactory function (Kim et al., 2015). The necessary complexity of olfactory test batteries has therefore remained an active research topic maintained by suggestions that olfactory test measure a common source of variance (Doty et al., 1994b) based on statistics based methods such as correlations (Hedner et al., 2010) or principal component analyses occasionally leading to different conclusions (Doty et al., 1994b; Lötsch et al., 2008).

Considering the ongoing discussion, the present analysis used a novel and alternative approach to assess whether the olfactory subtests confer the same information or whether each subtests at least partly confers non-redundant information relevant to the olfactory diagnosis that is not already conferred by the other subtests. Specifically, the present approach made use of computer-science based machine-learning methods (Murphy, 2012; Shalev-Shwartz and Ben-David, 2014), which aim at optimization of the performance of algorithms to predict a particular outcome such as the olfactory diagnosis. Therefore, supervised machine-leaned algorithms were either trained with the complete information from all three subtests, or with the information of only two or one subtests. By comparing the diagnostic accuracy of the algorithms among different training scenarios, the importance of the subtests for the clinical olfactory diagnosis was assessed. Thus, in the present analysis machine learning algorithms were used for knowledge discovery about the problem of information redundancy in olfactory subtests; a replacement of the simple and practicable sum score based olfactory diagnosis by complex classifiers was not intended.

Methods

Study population

The study followed the Declaration of Helsinki and was approved by the Ethics Committee of the Faculty of Medicine of the TU Dresden (number EK251112006) covering anonymized retrospective and pooled analyses. Informed written consent was obtained from all subjects. Subjects (age: range 6–95 years, mean ± standard deviation: 52.2 ± 17 years; sex: 6004 males, 4710 females) were included since they had presented at the Smell & Taste Clinic, Dept. of ORL, TU Dresden with the symptom "olfactory loss", or they had been tested in the context of a clinical standard check or they were enrolled in research projects as healthy participants. Subjects represented several different etiologies associated with olfactory performance; however, a previous analysis discouraged a strong relationship with the pattern of olfactory subtest results (Lötsch et al., 2016a) and was therefore not pursued again in the present analysis.

Olfactory testing

Olfactory function was assessed using the "Sniffin' Sticks" (Hummel et al., 1997b) test battery (Burghart, Wedel, Germany) that is composed of three sets of felt-tip pens filled with solutions of odors. For olfactory stimulation the pens are placed, with the cap removed, for approximately 3 s at 1–2 cm beneath the nostrils. A first set composed of 16 × 3 pens, of which 16 contain the rose-like odor phenyl ethyl alcohol at 16 successive 1:2 dilution steps starting from a 4% solution in propylene glycol, while the other pens contain just the solvent. During odor thresholds assessment, triplets of one pen with phenyl ethyl alcohol and two blanks were presented employing a three-alternative forced-choice task and a staircase paradigm. The odor threshold was finally estimated as the mean of the last four out of seven staircase reversals. A second set of 16 × 3 pens contained triplets with two similar and one different odor (for names of odors see (Hummel et al., 1997b)). During assessment of odor discrimination performance, the triplets were presented successively, and the subject's task was to identify the pen that smells different from the two others. The third test set comprises 16 pens containing different odors (Hummel et al., 1997b), which for assessment of the subject's odor identification performance had to be recognized in a four-alternative forced-choice task with presentation of a list of four possible descriptors for each pen. Each of the subtests provided scores between 0 (identification, discrimination) or 1 (threshold) as lowest values and 16 as the possible highest value. The final olfactory diagnosis was derived from the sum of the three scores, i.e., [0,...,16] for discrimination, [0,...,16] for identification, and [1,…,16] for thresholds, as follows: Sum scores of TDI (**T**hreshold, **D**iscrimination, **I**dentification) ≤ 16 indicated anosmia. Normosmia was indicated by scores of TDI ≥ 30.5 in females and TDI ≥ 29.5 in males, and the remaining TDI scores indicated hyposmia (Hummel et al., 2007b).

Data analysis

The contribution of each olfactory subtest result to the overall olfactory diagnosis was approached via creating machine-learned classifiers able to obtain the correct diagnosis and subsequently analyzing, for each olfactory subtest, how much the diagnostic accuracy decreased when the respective subtest results were excluded from the diagnostics. In other words, the machine learning was not applied to obtain the diagnosis, which can be obtained straight-forward as the sum of the subtest results. The machine-learning was applied to identify whether the diagnosis can already be obtained by knowing only one of the three numbers from which the sum score is obtained. If this was possible, three subtests would be unnecessary as all the important information was already contained in one single subtest. If, by contrast, the diagnosis not as accurately possible when using only one subtest result than when using the full information of the three subtests, then it can be concluded that each subtest provides relevant diagnostic information not already completely contained in the other subtest.

Data were analyzed using the R software package (version 3.4.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) on an Intel Xeon® computer running on Ubuntu Linux 16.04.3 64-bit. The acquired parameters included the results of the three olfactory subtests quantifying the subject's odor perception threshold, the performance in odor discrimination and the performance in odor identification. Data were incomplete in a single subject who was excluded from the analysis, which was performed in three main steps comprising (i) data preprocessing, (ii) classifier building and performance testing and (iii) estimation of the contribution of each olfactory subtest to the clinical olfactory diagnosis. Quantile-quantile plots suggested a log-transformation of the odor thresholds, which fits with the geometric series of the dilution steps of phenyl ethyl alcohol. Further data preprocessing consisted of standardizing into the [0,…,1] interval, which was obtained by rescaling the data based on their respective minimum and maximum values.

Subsequently to data cleaning and preprocessing, the data were explored using standard statistics consisting of an analysis of variance for repeated measurements (rm-ANOVA) and correlation analyses. Specifically, the rm-ANOVA was designed with "subtest", i.e., olfactory threshold, odor discrimination and odor identification as within-subject factor and "olfactory diagnosis", i.e., anosmia, hyposmia or normosmia, "gender" as between subject factors and "age" as covariate. In addition, correlations between age, the single olfactory subjects, and the TDI sum score, were assessed by means of calculating Spearman's ρ (Spearman, 1904), separately for the three olfactory diagnoses.

Supervised machine-learning was used to create so-called classifiers able to predict the three olfactory diagnoses anosmia, hyposmia or normosmia from the results of the three olfactory subtests. Machine learning addresses the so-called data space

$D = \{(x_i, y_i) x_i \in X,\ y_i \in Y,\ i = 1, \ldots, n\}$ including an input space X comprising vectors $x_i = < \mathrm{x}_{\mathrm{i},1}, \ldots, \mathrm{x}_{\mathrm{i},\mathrm{d}} >$ with $\mathrm{d} > 0$ different parameters (here, the olfactory subtest results) acquired from $\mathrm{n} > 0$ cases belonging to the output classes y_i (here, the three possible olfactory diagnoses). In supervised machine learning, an algorithm is trained on data for which the class labels of the cases are known (training data set), that is able to assign future cases for which this class label information is unknown to the right class (test data set; prediction, generalization (Dhar, 2013)). The necessary independent data sets were obtained by splitting the original data set into independent training (2/3 of the data) and test (1/3 of the data) data subsets, using the R library "sampling" (https://cran.r-project.org/package=sampling).

In the present analysis, the mapping of the input space, given by the selected olfactory subtest results prepared during data preprocessing, to the output space, given by the olfactory diagnoses, was performed using different methods of supervised machine learning. This was implemented to avoid that the results depend on one particular method, thus, for internal validation of the results in a biomedical context rather than for comparative benchmarking of classification performances as often applied in a computer science context. Specifically, supervised machine-learning was implemented as (i) ordinal logistic regression (Walker and Duncan, 1967), (ii) naïve Bayesian classifiers, (iii) classification and regression trees (Breimann et al., 1993), (iv) k-nearest neighbors (Cover and Hart, 1967), (v) random forests (Breiman, 2001), (vi) support vector machines (Cortes and Vapnik, 1995) and (vii) a multilayer perceptron (Rosenblatt, 1958) as specified in the following.

First, ordinal logistic regression provides a method for estimating the probability of the occurrence of an ordered dependent variable, i.e., olfactory diagnosis in improving succession, as a function of independent variables (McCullagh, 1980). This analysis was done using the R library "MASS" (https://cran.r-project.org/package=MASS (Venables and Ripley, 2002)). Second, naïve Bayesian classifiers were used that provide the probability that a data point being assigned to a specific class calculated by application of the Bayes' theorem (Bayes and Price, 1763). The calculations were done using the R package "klaR" (https://cran.r-project.org/package=klaR (Weihs et al., 2005)).

Third, in classification and regression trees, a tree data structure is created with conditions on variables (parameters) as vertices and classes (diagnoses) as leaves. Tree structured rule-based classifiers (Loh, 2014) analyze ordered variables x_i, such as the present results of olfactory subtests, by recursively splitting the data at each node into children nodes, starting at the root node. The split of the variable x takes the form of $x \leq c$ with $\mathrm{n} - 1$ possible splits in the data set (Morgan and Sonquist, 1963). During learning, the splits are modified such that misclassification is minimized. In the present form, the Gini impurity was used to find optimal (local) dichotomic decisions as used for the classification and regression tree method (CART) (Breimann et al., 1993). The calculations were done using the "rpart" function of the similarly named R package (B. Ripley; https://cran.r-project.org/package=rpart).

Forth, the k-nearest neighbor (kNN) classification (Cover and Hart, 1967) provides a non-parametric method that belongs to the most frequently used algorithms in data science although it is one of the basic methods in machine learning. During kNN model building, the entire labeled training dataset is stored while a test case is placed in the feature space in the vicinity of the test cases at the smallest high dimensional distance. The test case receives the class label according to the majority vote of the class labels of the k training cases in its vicinity. The present analyses were performed und $k = 5$ and the Euclidean distance, which also corresponds to the default of the R package "KernelKnn" (Mouselimis L, https://cran.r-project.org/package=KernelKnn).

Fifth, random forests creates sets of different, uncorrelated and often very simple decision trees (Breiman, 2001) with conditions on features as vertices and classes as leaves. In contrast to CART (see above), the splits of the features are random and the classifier relates on the majority vote for class membership provided by a large number of decision trees. In the present analysis, 1000 decision trees were built containing *sqrt(d)* features respectively nucleotide positions as the standard setting implemented in the R library "randomForest" (https://cran.r-project.org/package=randomForest (Liaw and Wiener, 2002)). The number of trees was based on assessing the out-of-bag error rate for up to 1600 trees, which remained at a minimum of 0.02 from 200 trees (Supplementary Fig. 1). As it is known that more trees do not confer a risk of increasing errors (Svetnik et al., 2003), a larger number was considered safe and merely consumed available computation time (see also at the end of the methods section where measures against overfitting are described).

Sixth, support vector machines are supervised learning methods that classify data mainly based on geometrical and statistical approaches employed for finding an optimum decision surface (hyper-plane) that can separate the data points of one class from those belonging to another class in the high-dimensional feature space (Cortes and Vapnik, 1995). Using a kernel function, the hyperplane is frequently selected in a way to obtain a tradeoff between minimizing the misclassification rate and maximizing the distance of the plane to the nearest properly classified data point. In the present analysis, a Gaussian kernel with a radial basis was used. The analyses were done using the R library "kernlab" (https://cran.r-project.org/package=kernlab (Karatzoglou et al., 2004)).

Seventh, a perceptron, which was among the first algorithmically described neural networks (Rosenblatt, 1958) was implemented. The algorithm is built from artificial neurons that are provided with several input channels, a processing level, and an output level that connects a neuron to one or multiple other artificial neurons. Each neuron sums up its weighed inputs plus an offset, or bias, and uses a linear combination according to the input weights to determine its neuron's activation function, usually with a logistic sigmoid shape, which determines the class association of a particular data point. During learning, weights and biases are adapted from initial random values in a way that the activation is shifted toward the desired output, i.e., the learning of a perceptron takes place by adjusting the weighting of each neuron. In the present analyses, a multilayer perceptron was used with 3 input neurons receiving the results of the three olfactory subtests, a single hidden layer composed of four neurons, and the output layer comprising three neurons given by the three output classes (anosmia, hyposmia, normosmia). Experiments using any combination of one or two hidden layers with 2 to 32 neurons each indicated no improvement of classification accuracy beyond four neurons in a single hidden layer. The analyses were done using the R library RNNS (https://cran.r-project.org/package=RSNNS (Bergmeir and Benitez, 2012)).

The analyses were applied on the original fully featured data set, i.e., containing the results of all three olfactory subtests, and subsequently, on reduced-feature data sets from which one or two olfactory subtest results had been omitted. This allowed assessing the classification accuracy of the different machine-learned methods and data set compositions for each olfactory diagnosis, i.e., the fraction of correct hits per olfactory diagnosis, obtained by the single machine-learned classifiers in the original data. In addition, it allowed estimating the classification accuracy obtained by chance using the permutated data, with the expectation that there in the diagnosis-specific classification accuracy was 50% corresponding to flipping a coin. The classification accuracy was primarily assessed as balanced accuracy (Brodersen et al., 2010; Velez et al., 2007), which is the mean of prediction sensitivity and specificity for each olfactory diagnosis and reflects the average of the proportion corrects of each class individually. Further, secondary measures of average classification performance across olfactory diagnoses included test sensitivity, and specificity and negative and positive predictive values calculated using standard equations (Altman and Bland, 1994a; 1994b).

Machine learning is vulnerable to overfitting as discussed previously (Lotsch and Ultsch, 2017), i.e., it may perform rote learning, obtain a perfect classification with a single data set but fail to classify similarly structured new data, or end up in describing noise or irrelevant relationships rather than the true relationship between features and classes. In that case, only the actual data on which the mapping has been learned are successfully classified, but the algorithm fails to classify new data. This issue was addressed in the present analysis fourfold. Firstly, prior to the data analysis, the classification algorithms were tuned with respect to available hyperparameters. For example, the number of *k* in kNN was tested between 3 and 9 and the best performing variant was chosen. Similarly, the number of trees in the random forest was assessed between 100 and 1600 and it was found that the out-of-bag error remained at a minimum of 0.02 from 200 trees (Supplementary Fig. 1); importantly, using more trees did not result in increased error due to, possibly, overfitting. Indeed, random forests achieves the error minimization by variance reduction. Therefore, as stated elsewhere (Svetnik et al., 2003), there is no penalty for having "too many" trees, other than waste in computational resources. Furthermore, several sizes of the multilayer perceptron were tested including one to three hidden layers with up to 20 neurons each. Secondly, analyses were performed in 100 cross-validation runs using random splits of the original data set into training (2/3 of the data) and test (1/3 of the data) data subsets. The reported classification performances are the median of the performances obtained during the 100 runs. Thirdly, a negative control condition was created as described above. A classification better than chance when trained with permuted data would hint at possible overfitting. Fourthly, seven different classifiers were applied to avoid that the analysis relied on a single method in which occasionally overfitting had occurred.

Results

Data of 10,713 subjects was analyzed, of whom 3662 had the olfactory diagnosis of anosmia, 4299 were diagnosed with hyposmia, and 2752 had normal olfactory function (normosmia). Descriptive data visualizations depicting the composition of the data set with respect to olfactory diagnoses, etiologies underling possible olfactory problems and gender, and the distribution of olfactory subtest results are shown in Fig. 1. The analysis of variance (Table 1) resulted in expected significant effects of the olfactory diagnosis, i.e., results obviously differed among subjects with anosmia, hyposmia, or normosmia. The analysis also identified well-known effects of gender and age on olfactory subtest results, as reflected in significant main effects and interactions of the respective factors or covariate, and it also found a significant interaction of "subtest" by "olfactory diagnosis" by "gender", which reflects the fact that the olfactory diagnosis is derived from the sum score of the three tests in a gender-specific manner. Furthermore, correlation analyses provided statistically significant results for among olfactory subtests or age, hover, not always as shown in Fig. 2. For example, lack of correlation was observed between odor identification and the olfactory threshold, or odor discrimination, in normosmic subjects, or between olfactory thresholds and odor discrimination performance in anosmic subjects.

Supervised machine-learning was applied on the three-class problem of predicting the olfactory diagnosis of anosmia, hyposmia or normosmia by mapping the 10,713 × 3 sized feature space to the output space. All machine-learned classifiers could be tuned to provide approximately 98% correct (Table 2), except the naïve Bayesian classifier that performed slightly inferior with approximately 95% correct diagnoses. By contrast, when training the classifiers on permuted data, the prediction was consistently like guessing at a balanced accuracy of approximately 50% (Table 2), making overfitting unlikely.

When omitting one of the olfactory subtests from the analyses (Table 3 and Fig. 3), the largest drop in balanced classification accuracy was observed when the olfactory threshold was excluded (only approximately 85% correct diagnoses in 100 cross-validation runs using random resampling and seven different machine-learned classifiers). Other combinations of two olfactory subtests indicated that when including the olfactory threshold, the diagnosis was equally accurate (approximately 91%). When training the classifiers with results of single olfactory subtests, the best diagnostic performance was obtained with odor thresholds (approximately 84% balanced accuracy) or odor identification scores (approximately 85%), whereas the odor

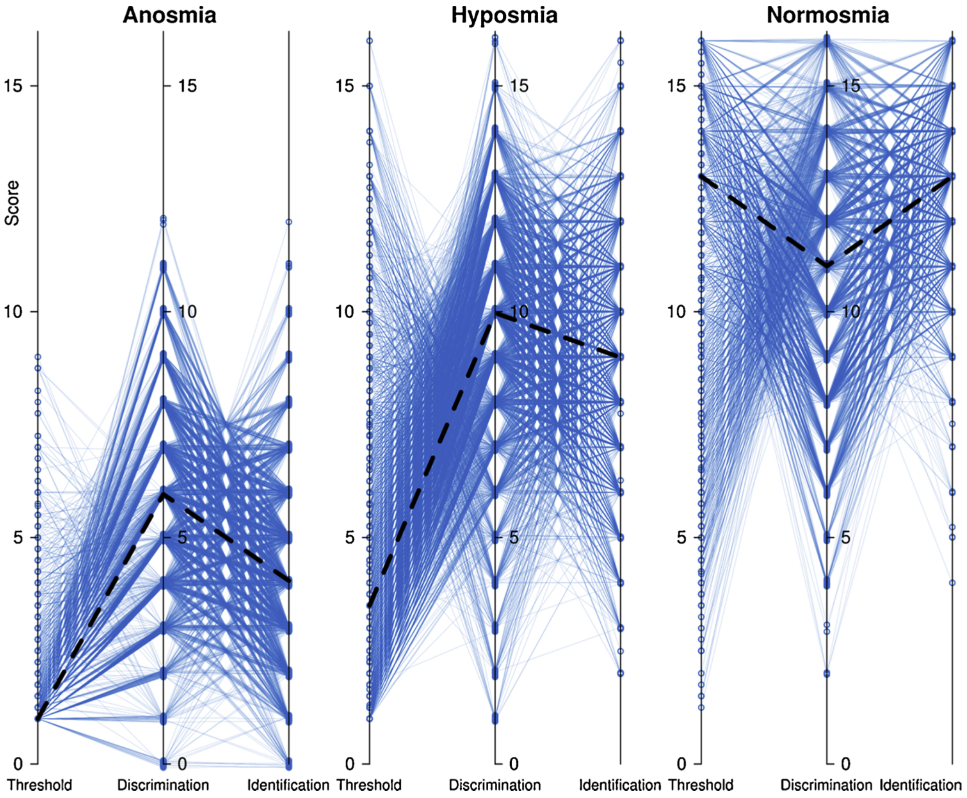

Fig. 1. Original individual data shown in "spaghetti plots", separately for the three olfactory diagnoses for better visibility. The individual values of olfactory threshold, odor discrimination and odor identification are connected by straight lines. The data are slightly jittered to enhance visibility by reducing superimposition of data points. The bold dashed black lines indicate the medians across the whole cohort. The figure has been created using the R software package (version 3.4.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)).

Table 1
Results of the analysis of variance for repeated measurements (rm-ANOVA) and correlation analyses. Specifically, the rm-ANOVA was designed with “subtest”, i.e., olfactory threshold, odor discrimination and odor identification as within-subject factor and “olfactory diagnosis”, i.e., anosmia, hyposmia or normosmia, “gender” as between subject factors and “age as covariate. Degrees of freedom, F-values and p-values are shown for main effects and interactions.

Effect	Degrees of freedom	F-value	p-value
Subtest (threshold, discrimination, identification)	2,21412	70.742	$2.38 \cdot 10^{-31}$
Olfactory diagnosis (anosmia, hyposmia, normosmia))	2,10706	27550.541	$< 10^{-100}$
Gender (male, female)	1,10706	66.355	$4.18 \cdot 10^{-16}$
Age	1,10706	33.169	$8.68 \cdot 10^{-9}$
Subtest * olfactory diagnosis	4,21412	1035.703	$< 10^{-100}$
Subtest * gender	2,21412	1.199	0.301
Subtest * age	2,21412	190.135	$1.41 \cdot 10^{-82}$
Subtest * olfactory diagnosis * gender	4,21412	2.385	0.049
Olfactory diagnosis * age	2,10706	4.386	0.012

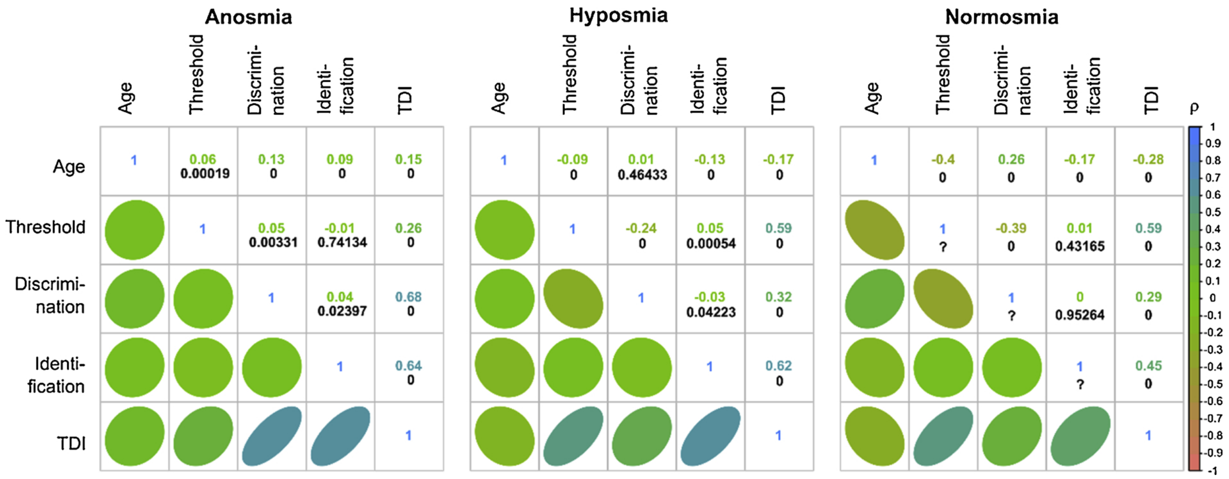

Fig. 2. Explorative analysis of the correlations between age, the single olfactory subjects, and the TDI sum score, separately for the three olfactory diagnoses. At the lower left parts, the correlations are shown as ellipses, with the direction toward positive (upwards) or negative (downwards) correlations, and colored according to the color code of Spearman’s ρ (Spearman, 1904) shown at the bottom of the panels. At the upper right parts, the correlations are provided numerically as values of Spearman’s ρ (colored). The p-values are shown in black numbers below the correlation coefficients; “0″ indicates $p < 1 \cdot 10^{-5}$. The figure has been created using the R software package (version 3.4.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the library “corrplot” (https://cran.r-project.org/package=corrplot (Wei and Simko, 2017)).

discrimination score provided the least information for correct olfactory diagnosis with a classifier performance of approximately 68%. Again, when training the classifiers on permuted data in any combination scenario, the prediction was consistently like guessing at a balanced accuracy of approximately 50% (details not shown).

Finally, the observation was addressed why none of the machine-learned classifiers provides a perfect olfactory diagnosis, even when training the algorithms with the complete olfactory test results. Indeed, the diagnosis is not only obtained as the sum of the scores obtained in the olfactory subtests, but also depends on the subject’s sex, i.e., normosmia is defined as a sum score > 30.5 in females but > 29.5 in males (Hummel et al., 2007b). Letting the classifiers train with the additional parameter sex, scaled as [0,1] for females and males, respectively, raised the accuracy of the olfactory diagnosis up to more 99% (Table 3), and the addition of age (rescaled into [0..1]) raised this further, approaching but not perfectly reaching 100%. Age and sex alone provided already a classification accuracy of around 60%, i.e., increased the correct diagnosis above chance (further details not shown).

Discussion

In a previous analysis of the same data set using unsupervised machine-learning, a high-dimensional structure was shown that was superimposable with the olfactory diagnoses of normosmia, hyposmia or anosmia (Lötsch et al., 2016a). This indicated that results of the olfactory subtests separately, not just by their sum, contain information to establish the clinical olfactory diagnosis. However, a frequently observed high correlation of the three subtests had raised the question of their redundancy in the past (Doty et al., 1994a; Lötsch et al., 2008), which has been answered contrastingly. Indeed, results of statistical analyses of the present data hinted at differences among subtests, such as the significant main effect of the factor “subtest” in the rm-ANOVA or the exceptions from correlation of subtest results. However, lack of some correlations in anosmic or normosmic subjects might be explained by subtest scores at the extremes of the scale. Hence, based on the results of statistical analyses, the question about non-redundant information in the subtests results was still difficult to answer. Therefore, the present study used an alternative analytical approach consisting of the training machine-learned algorithms to establish the olfactory diagnosis with either the complete information for olfactory subjects or with parts of it, and to observe whether the diagnosis accuracy decreases when subtest information was omitted from the training. Thus, applying methods of supervised machine-learning, the decrease in classification accuracy when a subtest was left out from training was taken to assess whether the three common olfactory subtests confer only redundant information or whether there is an additional contribution to the overall olfactory performance conferred by a particular subtest above what is already conferred by any other subtest as well.

The results of the present analyses suggest that each olfactory subtest contributes separate information relevant to the olfactory diagnoses. The information provided by the results of three separate subtests of a comprehensive olfactory test battery suffices to draw the

Table 2

Performance measures of classifiers obtained using different machine-learned methods (ordinal logistic regression, naïve Bayes, classification and regression trees (CART), k-nearest neighbors, random forests, support vector machines, multilayer perceptron) on the data set comprising the results of three olfactory subtests (assessment of olfactory threshold, odor discrimination and odor identification) acquired in 10,713 subjects. Results represent the medians of the test performance measures from 100 model runs using random splits of the data set into training data (2/3 of the data set) and test data (1/3 of the data set). The classifiers were trained on the original data set and again on randomly permuted training data.

Data set	Original data							Permuted data						
Parameter	CART	Random forests	k-nearest neighbors	Support vector machines	Ordinal logistic regression	Multilayer perceptron	Naïve Bayes	CART	Random forests	k-nearest neighbors	Support vector machines	Ordinal logistic regression	Multilayer perceptron	Naïve Bayes
Sensitivity, recall [%]	96.5	97.3	97.2	98	97.2	97.9	93	34.7	15.1	41.2	0.8	0	4.4	0
Specificity [%]	98.5	98.8	98.7	99.1	98.8	99.1	95.6	64.9	77.9	57.7	98.4	100	97.5	100
Positive predictive value, precision [%]	96.5	97.3	97.1	98.1	97	97.8	91	34.3	39.9	34.8	40.5	40.4	40.4	40.6
Negative predictive value [%]	98.6	98.8	98.8	99.1	98.7	99	97.1	65.7	66.7	66.4	73.6	72.8	73.4	73.7
Precision	96.5	97.3	97.1	98.1	97	97.8	91	34.3	39.9	34.8	40.5	40.4	40.4	40.6
Balanced accuracy [%]	97.2	97.8	97.7	98.4	95.4	98.2	94.4	50.2	50.7	50	50	50	50	50

clinical olfactory diagnosis, but when the information of one or two components is omitted, the diagnostic accuracy dropped by different amounts. This has been consistently observed across seven different machine-learned or artificial intelligence algorithms, which remarkably agreed in their results and hinted at odor identification and odor thresholds as providing most relevant information for the diagnosis while odor discrimination contributed to a lower degree although its omission could not be completely substituted for by the information provided by the other two tests. Thus, the present analysis takes the previous assessments (Lötsch et al., 2016a) a step further toward a more detailed analysis of the observed association between patterns of olfactory subtest results and the olfactory diagnosis while the lack of similar associations discouraged further inclusion of the etiologies underlying the olfactory diagnosis in this analysis.

Based on the present data driven approach, which did not require major prior assumptions about relationship or distributions of the included parameters, the present results support non-redundant components of the contributions of each olfactory subtest to the clinical olfactory diagnosis. This contrasts to some degree with suggestions that tests of single components of olfactory function measure a common source of variance (Doty et al., 1994b). That judgment had been based on the Kaiser-Gutman criterion on the results of a principal component analysis of the three olfactory subtest, advising to regard only principal components with an eigenvalue > 1 of the covariance matrix (Guttman, 1954; Kaiser and Dickman, 1959). However, the Kaiser-Gutman criterion was shown to occasionally disregard important factors (Ivanenko et al., 2004; Ultsch and Lötsch, 2015) leading to the advice to retain factors with eigenvalues higher > 0.5, or seeking calculated criteria rather than setting a fixed border (Ultsch and Lötsch, 2015). Indeed, when following this advice, a principal component analysis provided support for independent contributions of olfactory subtests to the olfactory diagnosis (Lötsch et al., 2008). As discussed previously (Doty et al., 1994a), one possibility to interpreted these observations is that olfactory perception is multidimensional and that a variety of olfactory tests tap, to a large degree, elements that are defined by most of the olfactory tests. For this perspective, even the act of detecting an odorant can be viewed as requiring to some degree of ability to remember the odorant and to discriminate it from a blank. By contrast, in the present analysis a particular cut-off criterion of a statistical test was not needed as it used artificial intelligence and machine learning that have developed from computer science (Shalev-Shwartz and Ben-David, 2014; Turing, 1950) while statistics can be regarded as a branch of mathematics. The present analysis was centered on the performance of the algorithms to provide the olfactory diagnosis based on full and reduced sets of olfactory subtest results, rather than on the analysis of the probabilities of the subtest results given a known underlying distribution.

The "Sniffin' Sticks" comprise 3 different subtests, odor threshold, odor discrimination, and odor identification. In numerous studies this concept has been shown for over 2 decades to provide reliable and useful data for the clinical diagnosis of olfactory dysfunction. Normative data have been established in more than 3000 healthy subjects for various age groups (Hummel et al., 2007b), and, importantly, different from other tests, criteria for the interpretation of tests results in terms of clinical improvement have been established (Gudziol et al., 2006). While this is clearly a strength compared to other tests that only look at one dimension of olfactory function, e.g., odor identification, this may also be regarded as weakness because of the differences between the individual subtests of the "Sniffin' Sticks" in terms of reliability, odor components, or task demands. However, such criticism may also apply to other, single-dimensional tests where different odors (single molecules or mixtures) are used in odor identification tests, and it is not exactly known to which degree each single odor contributes to the overall diagnosis of the olfactory dysfunction of an individual patient – because odor identification also depends on numerous factors like verbal abilities, or familiarity with the individual

Table 3
Balanced accuracy of the olfactory diagnosis of different machine-learned classifiers (ordinal logistic regression, naïve Bayes, classification and regression trees (CART), k-nearest neighbors, random forests, support vector machines, multilayer perceptron) trained with data from the full data set comprising the results of three olfactory subtests (assessment of olfactory threshold, odor discrimination and odor identification) acquired in 10,713 subjects, and with reduced data sets consisting of the results of two or one olfactory subtests (olfactory threshold, T, odor discrimination, D, odor identification, I). In addition, the classification accuracy of age or sex or combinations with the full data set has been assessed. Results represent the medians of the test performance measures from 100 model runs using random splits of the data set into training data (2/3 of the data set) and test data (1/3 of the data set).

Subtests	Complete			None	Threshold			Discrimination		Identification
		+ Sex	+ Sex + age	Age and sex	Alone	+ Discrimination	+ Identification	Alone	+ Identification	Alone
	TDI	TDI + sex	TDI + sex + age	none	T	TD	TI	D	DI	I
Methods						w/o identification	w/o discrimination		w/o threshold	
CART	97.2	97.3	96.9	59	83.6	91.1	90.9	68.4	85.2	84.9
Random forests	97.8	98.3	98	59.4	83.6	91	91.1	68.4	85.5	84.9
k-nearest neighbors	97.7	97.7	96.4	55.3	82	90.1	90.4	65.9	84	83.2
Support vector machines	98.4	99.1	98.7	59.9	83.5	91.1	91.5	68.5	85.9	84.9
Ordinal logistic regression	95.4	95.5	95.6	55.5	84.2	89.3	91.3	67.2	83.9	85.1
Multilayer perceptron	98.2	99.1	99.1	59.7	83.5	91.3	91.4	67.5	85.8	85.1
Naïve Bayes	94.4	94.4	93.2	60.4	84.9	90	90.9	67.6	86.5	85.1

odor (Hedner et al., 2010). This is the reason why odor identification tests for clinical diagnostic procedures not only use one odor but several. Having said this, when interpreting the results from the preset study it has to be kept in mind that the “Sniffin Sticks” extended test is a mixture of different concepts.

The present results point at a particular importance of olfactory thresholds. A distinct role of olfactory thresholds had already been suggested by a separate principal component in a previous assessment in a different cohort (Lötsch et al., 2008). In the present analysis, their importance was indicated by the most pronounced decrease in diagnostic accuracy of the machine-learned algorithms when olfactory thresholds were omitted from the parameter set, and on their comparatively good performance as a single predictive parameter equaling that of odor identification. This result supports and encourages the use of olfactory thresholds in unimodal olfactory test batteries, which commonly tend to choose odor identification, including most of the so-called short screening tests, while test batteries based on olfactory threshold assessments are rare (Yilmaz et al., 2017). A possible advantage of olfactory thresholds to odor identification performance tests is the minor role of chance in the test outcome when a staircase paradigm with reversals is used; for example, the present implementation of 16 odors in a four-alternative forced choice assessment of identification implies an average score of 4 due to chance. This requires the addition of sufficient items to reduce the impact of chance on

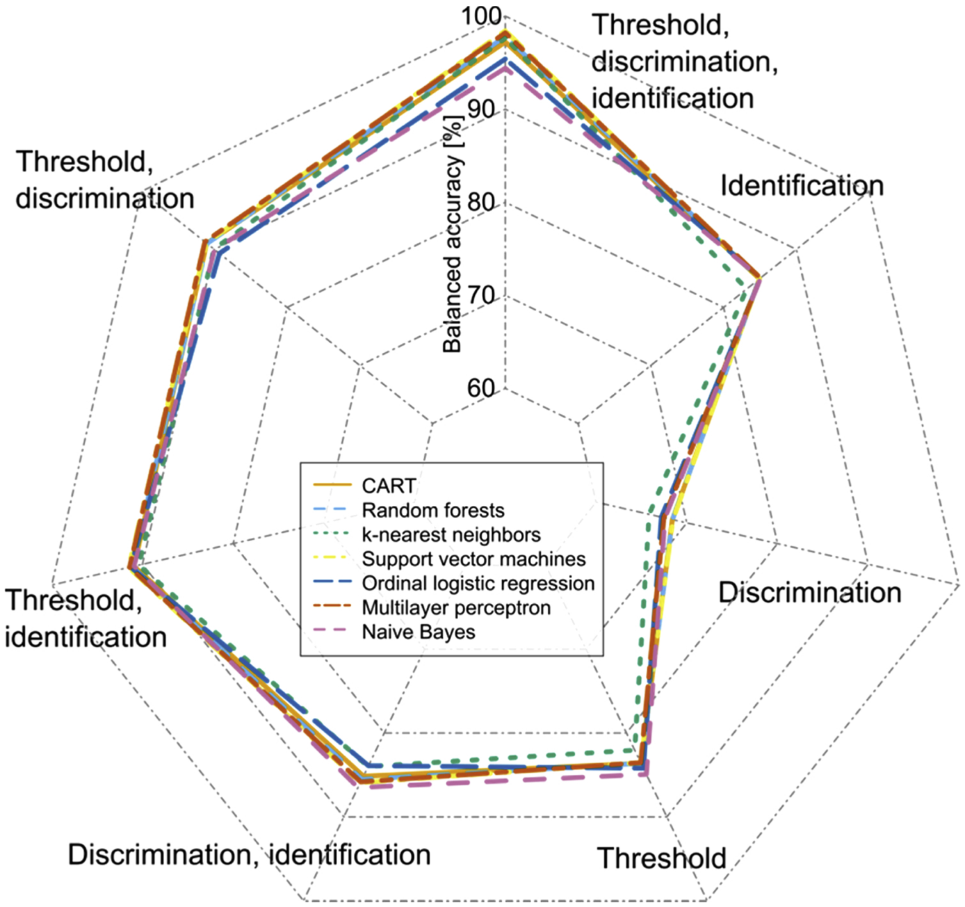

Fig. 3. Radar plot of the balanced accuracy of different classifiers (ordinal logistic regression, naïve Bayes, classification and regression trees (CART), k-nearest neighbors, random forests, support vector machines, multilayer perceptrons) to establish the clinical olfactory diagnosis (anosmia, hyposmia or normosmia) from olfactory subtest results. The classification performance has been assessed in of 100 model runs using random resampling with splits into 2/3 of the data (training data subset) and 1/3 (test data subset). The plot shows the balanced accuracies in a spider web form. Each category, i.e., machine-learning method, has a separate axis, scaled from 60 to 100% balanced accuracy. The axes are arranged in a circle in 360 degrees evenly, and the values of each series are connected with lines indicating the results obtained with either of the three data sets, i.e., the fully featured set of olfactory subtest results comprising the olfactory threshold, T, odor discrimination, D, odor identification, I, or with reduced-feature data sets from which one or two olfactory subtest results had been omitted. The figure has been created using the R software package (version 3.4.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) with the “radarchart” function provided in the library “fmsb” (M. Nakazawa, https://cran.r-project.org/package=fmsb).

the overall test result (Lötsch et al., 2016b).

On a biological level, a distinction of olfactory thresholds from other olfactory subtests hypothetically involves a cognitive component to a lesser degree (Hedner et al., 2010). That is, although formally a three-alternative paradigm is used, the pen containing the odor may be identified immediately without necessary reference to the non-smelling pens. By contrast, during odor discrimination performance testing, the subject must memorize the smell of the other pens before completing the task, and memorizing odors is also required, at least to some degree, for odor identification. For example, this hypothesized difference in the role of cognitive factors for threshold and discrimination/identification testing is in line with the report that odor identification, but not thresholds, were associated with AIDS-related dementia (Hornung et al., 1998) and difficulty in identifying odors was shown to predict subsequent development of mild cognitive impairment in older patients (Wilson et al., 2007). In addition, an advantage of odor threshold testing over testing of odor identification or odor discrimination performance is the relative cultural or social independence of odor thresholds. Both odor discrimination and odor identification strongly depend on the familiarity with odors. This familiarity varies from region to region. For example, odors like "wintergreen" may be known in the USA, but are unknown in most European countries, let alone the Arabian world, so that odor identification tests need to be culturally adapted (Croy et al., 2014; Oleszkiewicz et al., 2016).

However, dismissing odor discrimination and odor identification for olfactory diagnosis due to their more pronounced cognitive and cultural component as compared to olfactory thresholds would result in a different olfactory diagnosis for some subjects. Normal values of olfactory thresholds in the present test battery of > 6 for males and > 6.5 for females (Hummel et al., 2007a) have been exceeded by a n = 24 subjects assigned to the diagnosis of anosmia (Fig. 1). However, a clinical diagnosis of functional anosmia may remain valid when considering the possibility of olfactory agnosia, i.e., dissociation of olfactory acuity and identification abilities, which have been described as a symptom of Korsakoff's amnestic syndrome, lesions of the orbitofrontal cortex (Potter and Butters, 1980) or schizophrenia (Kopala and Clark, 1990), for the latter with disproved relation to task complexity of the olfactory test (Kopala et al., 1995). A more complete regard of olfactory function seems also supported by scenarios were professionals who rely on olfactory function such as chefs or perfumers would certainly be disabled when they lose the ability to identify of discriminate odors regardless of where they can still perceive the presence of an olfactory stimulus.

When accepting that the diagnostic accuracy dropped most when omitting olfactory thresholds from the set of olfactory subtests, it can be hypothesized that odor threshold reflects the basis of odor perception. While alone it preforms as well as basing the diagnosis on odor identification only, almost completely accurate diagnoses depend on it more than on the other two subtests. This suggests an advantage of tests offering sole testing of odor thresholds over alternatives that are based on identification only. However, more importantly the present results encourage more complex tests of olfactory function with separate tests of odor thresholds, identification and maybe also odor discrimination (Hummel et al., 1997a). This thought is also maintained in a recent position paper on olfactory dysfunction (Hummel et al., 2017) where it reads that "…assessing both odor threshold and suprathreshold tasks adds to the diagnostic value of the psychophysical tool.".

The present data analysis included machine-learning approaches; regression and naïve Bayes based assignments to the olfactory diagnoses were added as more classical algorithms. The use of different algorithms aimed at internal validation, which was achieved with the consistency of the results across all algorithms. Moreover, combined with random resampling of disjoint data sets, the machine-learned algorithms did not base the diagnoses solely on prior knowledge but re-established it from raw threshold, discrimination and identification data. While resampling-based cross-validation is not unique to machine-learning but a standard in many mathematical-modeling based data analysis approaches in biomedical research, the as far as possible avoidance of pre-established limits seems to have provided an advantage over a classical statistical approach to the present problem. That is, the focus of establishing normative values in the Sniffn' Sticks test battery was laid on the sum score, and most publications mainly used the sum score as the basis for their analyses. For the subtest, normative values have only been addressed twice. Specifically, limits separating hyposmia from normosmia have been reported as scores of 6.5, 10 and 11 for odor threshold, discrimination and identification, respectively (Hummel et al., 2007b). The respective limits between hyposmia and anosmia were reported earlier and at scores of 1, 8 and 8 (Kobal et al., 2000). Diagnostic accuracies of 78.4, 66.7 and 84.4% were achieved when using these classical limits for establishing the olfactory diagnosis. This was below the best values of balanced accuracies of 84.9, 68.5 and 85.1% reached with machine-leaning, naïve Bayes or regression-based classifiers run with one single subtest result (Table 3). The classical limits had been set at the 10th or 90th percentiles of distributions observed in the data used for establishing normative values. While this has been a sound statistical approach, it nevertheless implied a decision about the percentiles. This kind of decision was attempted to be avoided as much as possible in the present analysis, hence, the use of supervised machine learning where the diagnosis was learned from the training data sets rather than given by pre-established limits. A more objective approach at establishing limits in odor thresholds, discrimination and identification scores would have been, for example, using Bayesian decision borders. Indeed, this has been implicitly implemented in the main analysis. Explicit calculation of these border from the whole data set resulted in new thresholds between anosmia and hyposmia, and between hyposmia and normosmia of 2.8 and 8.6 for odor thresholds, of 7.6 and 12.8 for odor discrimination, and 6.4 and 11.3 for odor identification. Using these data-based limits, the diagnostic accuracy of the single subtest for the clinical olfactory diagnosis raised to 83.9, 67.3 and 85.1 for thresholds, discrimination and identification, respectively. Of note, the differences to the results obtained with Naïve Bayes classifiers may be attributed partly to the resampling strategies used in Table 3, whereas the values above are simply calculated form the whole data set.

However, the present study was not intended to revise published limits; nevertheless, the above demonstrations emphasizes the utility of a data-driven approach to the present problem, conceptually possibly superior to pure statistical calculations. Nevertheless, both approaches came to the same main conclusions, which indicates that despite the criticism expressed above about the establishment of the classical limits for olfactory subtests and the demonstration that it could be easily improved, the historical approach was not completely misleading. Again, it has been an explicit aim of the present analysis to use as few as possible decisions or pre-established rules, to base the diagnosis mainly on the data analysis, and to show that the results are not attributable to a single algorithm but prevail across a variety of implementations.

To the presently observed distinction of odor thresholds adds the observed difficulty of IPD patients to identify or recognize odors (Masaoka et al., 2007), while they are still able to detect odors. This further emphasizes the need to test different aspects of olfaction rather than relying on just one component in neurological diagnosis. As indicated above, such differential approaches to the diagnosis of olfactory loss have been proposed (Henkin, 1971; Henkin et al., 2013). However, in clinical practice the present, possibly simplified, approach using three subtests to establish an olfactory diagnosis, is still a standard. Moreover, a certain interest is focused on even more simple tests as reflected in proposals of several short tests of olfactory function (Hummel et al., 2010; Toledano et al., 2009) (Hummel et al., 2010; Mueller and Renner, 2006).

Conclusions

In a large data set of 10,713 subjects tested with a three-component olfactory test where a previous machine-learned analysis has shown data structures in the olfactory subtest results that agreed with the olfactory diagnosis but not the underlying etiology (Hummel et al., 1997b), the present analysis applied supervised machine learning methods (Murphy, 2012; Shalev-Shwartz and Ben-David, 2014) to take the analysis a step further toward the assessment of the relative importance of subtest results for the olfactory diagnosis. The underling idea was to assess this relative importance by comparing the diagnostic accuracy among classifiers trained either from the full set of olfactory subtest results or from a reduced number of olfactory subtests. The importance of an olfactory subtest was assessed via the drop in diagnostic accuracy when the respective parameter was omitted from the classifier training. This focused on the utility of each single subtest for the complete olfactory diagnosis in possible automatized implementations (Lotsch et al., 2018), rather than on merely calculating the accuracy of one or two olfactory test in providing the diagnosis. In addition to straight-forward statistical approaches, the use of several different algorithms, nonredunant to earlier approaches that also decided to not to rely solely on simple calculations of separate diagnostic accuracies (Doty et al., 1994b), and the massive use of data resampling provided internal validations of the results.

The results of the analyses indicated that the information provided by the results of three separate olfactory subtests suffices to train an artificial intelligence or related machine-learned algorithm that can establish the clinical olfactory diagnosis almost perfectly. Moreover, seven different machine-learned classifiers provided highly consistent results supporting partly non-redundant contributions of each olfactory subtest to the clinical olfactory diagnosis and pointing at olfactory thresholds as providing - within the currently studied diagnostic procedure, disregarding aspects of test-retest reliability - the least negligible information to the overall olfactory diagnosis.

Conflicts of interest

The authors have declared that no competing interests exist.

Funding

This work has been funded by the Landesoffensive zur Entwicklung wissenschaftlich - ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL). We also would like to thank the Deutsche Forschungsgemeinschaft for support (DFG HU 441/18-1 to TH and DFG Lo 612/10-1 to JL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceived and designed the analysis: JL. Analyzed the data: JL. Wrote the paper: JL, TH. Provided data: TH. Critical revision of the manuscript for important intellectual content: TH

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ibror.2019.01.002.

References

Altman, D.G., Bland, J.M., 1994a. Diagnostic tests 2: predictive values. BMJ 309, 102.

Altman, D.G., Bland, J.M., 1994b. Diagnostic tests. 1: sensitivity and specificity. BMJ 308, 1552.

Bayes, M., Price, M., 1763. An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Philos. Trans. 53, 370–418.

Bergmeir, C., Benitez, J.M., 2012. Neural networks in R using the Stuttgart Neural Network Simulator: RSNNS. J. Stat. Softw. 46, 1–26.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breimann, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1993. Classification and Regression Trees. Chapman and Hall, Boca Raton.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Pattern Recognition (ICPR), 2010 20th International Conference. pp. 3121–3124.

Cain, W.S., 1979. To know with the nose: keys to odor identification. Science 203, 467–470.

Cain, W.S., Krause, R.J., 1979. Olfactory testing: rules for odor identification. Neurol. Res. 1, 1–9.

Cain, W.S., Gent, J.F., Goodspeed, R.B., Leonard, G., 1988a. Evaluation of olfactory dysfunction in the Connecticut Chemosensory Clinical Research Center. Laryngoscope 98, 83–88.

Cain, W.S., Gent, J.F., Goodspeed, R.B., Leonard, G., 1988b. Evaluation of olfactory dysfunction in the Connecticut Chemosensory Clinical Research Center (CCCRC). Laryngoscope 98, 83–88.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13, 21–27.

Croy, I., Hoffmann, H., Philpott, C., Rombaux, P., Welge-Luessen, A., Vodicka, J., Konstantinidis, I., Morera, E., Hummel, T., 2014. Retronasal testing of olfactory function: an investigation and comparison in seven countries. Eur. Arch. Otorhinolaryngol. 271, 1087–1095.

Davidson, T.M., Murphy, C., 1997. Rapid clinical evaluation of anosmia. The alcohol sniff test. Arch. Otolaryngol. Head Neck Surg. 123, 591–594.

Dhar, V., 2013. Data science and prediction. Commun. ACM 56, 64–73.

Doty, R.L., 2015. Olfactory dysfunction and its measurement in the clinic. World J. Otorhinolaryngol. Head Neck Surg. 1, 28–33.

Doty, R.L., Agrawal, U., 1989. The shelf life of the University of Pennsylvania Smell Identification Test (UPSIT). Laryngoscope 99, 402–404.

Doty, R.L., Shaman, P., Dann, M., 1984. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiol. Behav. 32, 489–502.

Doty, R.L., Smith, R., McKeown, D.A., Raj, J., 1994a. Tests of human olfactory function: principal components analysis suggests that most measure a common source of variance. Percept. Psychophys. 56, 701–707.

Doty, R.L., Smith, R., McKeown, D.A., Raj, J., 1994b. Tests of human olfactory function: principle component analysis suggests that most measure a common source of variance. Percept. Psychophys. 56, 701–707.

Gudziol, V., Lötsch, J., Hahner, A., Zahnert, T., Hummel, T., 2006. Clinical significance of results from olfactory testing. Laryngoscope 116, 1858–1863.

Guttman, L., 1954. Some necessary conditions for common factor analysis. Psychometrika 19, 149–161.

Hedner, M., Larsson, M., Arnold, N., Zucco, G.M., Hummel, T., 2010. Cognitive factors in odor detection, odor discrimination, and odor identification tasks. J. Clin. Exp. Neuropsychol. 32, 1062–1067.

Henkin, R.I., 1971. Disorders of taste and smell. JAMA 218, 1946.

Henkin, R.I., Levy, L.M., Fordyce, A., 2013. Taste and smell function in chronic disease: a review of clinical and biochemical evaluations of taste and smell dysfunction in over 5000 patients at The Taste and Smell Clinic in Washington, DC. Am. J. Otolaryngol. 34, 477–489.

Henkin, R.I., Abdelmeguid, M., Knoppel, A.B., 2016. On the mechanism of smell loss in patients with Type II congenital hyposmia. Am. J. Otolaryngol. 37, 436–441.

Hornung, D.E., Kurtz, D.B., Bradshaw, C.B., Seipel, D.M., Kent, P.F., Blair, D.C., Emko, P., 1998. The olfactory loss that accompanies an HIV infection. Physiol. Behav. 64, 549–556.

Hummel, T., Sekinger, B., Wolf, S., Pauli, E., Kobal, G., 1997a. "Sniffin' Sticks": olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. Chem. Senses 22, 39–52.

Hummel, T., Sekinger, B., Wolf, S.R., Pauli, E., Kobal, G., 1997b. 'SNiffin' sticks': olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. Chem. Senses 22, 39–52.

Hummel, T., Kobal, G., Gudziol, H., Mackay-Sim, A., 2007a. Normative data for the "Sniffin' Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. Eur. Arch. Otorhinolaryngol. 264, 237–243.

Hummel, T., Kobal, G., Gudziol, H., Mackay-Sim, A., 2007b. Normative data for the "Sniffin' Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. Eur. Arch. Otorhinolaryngol. 264, 237–243.

Hummel, T., Pfetzing, U., Lötsch, J., 2010. A short olfactory test based on the identification of three odors. J. Neurol. 257, 1316–1321.

Hummel, T., Whitcroft, K.L., Andrews, P., Altundag, A., Cinghi, C., Costanzo, R.M., Damm, M., Frasnelli, J., Gudziol, H., Gupta, N., Haehne, A., Holbrook, E., Hong, S.C., Hornung, D., Huttenbrink, K.B., Kamel, R., Kobayashi, M., Konstantinidis, I., Landis, B.N., Leopold, D.A., Macchi, A., Miwa, T., Moesges, R., Mullol, J., Mueller, C.A., Ottaviano, G., Passali, G.C., Philpott, C., Pinto, J.M., Ramakrishnan, V.J., Rombaux, P., Roth, Y., Schlosser, R.A., Shu, B., Soler, G., Stjarne, P., Stuck, B.A., Vodicka, J., Welge-Luessen, A., 2017. Position paper on olfactory dysfunction. Rhinol. Suppl. 54, 1–30.

Ivanenko, Y.P., Poppele, R.E., Lacquaniti, F., 2004. Five basic muscle activation patterns account for muscle activity during human locomotion. J. Physiol. (Paris) 556, 267–282.

Jackman, A.H., Doty, R.L., 2005. Utility of a three-item smell identification test in detecting olfactory dysfunction. Laryngoscope 115, 2209–2212.

Kaiser, H.F., Dickman, K., 1959. Analytic determination of common factors. Am. Psychol.

14, 425.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - an S4 package for Kernel methods in R. J. Stat. Softw. 11, 1–20.

Kim, B.G., Oh, J.H., Choi, H.N., Park, S.Y., 2015. Simple assessment of olfaction in patients with chronic rhinosinusitis. Acta Otolaryngol. 135, 258–263.

Kobal, G., Klimek, L., Wolfensberger, M., Guziol, H., Temmel, A., Owen, C.M., Seeber, H., Pauli, E., Hummel, T., 2000. Multi-center investigation of 1036 subjects using a standardized method for the assessment of olfactory function combining tests of odor identification, odor discrimination, and olfactory thresholds. Eur. Arch. Otorhinolaryngol. 257, 205–211.

Kopala, L., Clark, C., 1990. Implications of olfactory agnosia for understanding sex differences in schizophrenia. Schizophr. Bull. 16, 255–261.

Kopala, L.C., Good, K., Martzke, J., Hurwitz, T., 1995. Olfactory deficits in schizophrenia are not a function of task complexity. Schizophr. Res. 17, 195–199.

Lam, H.C., Sung, J.K., Abdullah, V.J., van Hasselt, C.A., 2006. The combined olfactory test in a Chinese population. J. Laryngol. Otol. 120, 113–116.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Loh, W.-Y., 2014. Fifty years of classification and regression trees. Int. Stat. Rev. 82, 329–348.

Lotsch, J., Ultsch, A., 2017. Machine learning in pain research. Pain 159, 623–630.

Lötsch, J., Reichmann, H., Hummel, T., 2008. Different odor tests contribute differently to the evaluation of olfactory loss. Chem. Senses 33, 17–21.

Lötsch, J., Hummel, T., Ultsch, A., 2016a. Machine-learned pattern identification in olfactory subtest results. Sci. Rep. 6, 35688.

Lötsch, J., Ultsch, A., Hummel, T., 2016b. How many and which odor identification items are needed to establish normal olfactory function? Chem. Senses.

Lotsch, J., Kringel, D., Hummel, T., 2018. Machine learning in human olfactory research. Chem. Senses.

Masaoka, Y., Yoshimura, N., Inoue, M., Kawamura, M., Homma, I., 2007. Impairment of odor recognition in Parkinson's disease caused by weak activations of the orbitofrontal cortex. Neurosci. Lett. Suppl. 412, 45–50.

McCullagh, P., 1980. Regression models for ordinal data. J. R. Stat. Soc. Ser. B (Methodol.) 42, 109–142.

Morgan, J.N., Sonquist, J.A., 1963. Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415–434.

Mueller, C., Renner, B., 2006. A new procedure for the short screening of olfactory function using five items from the "Sniffin' Sticks" identification test kit. Am. J. Rhinol. 20, 113–116.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Oleszkiewicz, A., Walliczek-Dworschak, U., Klotze, P., Gerber, F., Croy, I., Hummel, T., 2016. Developmental changes in adolescents' olfactory performance and significance of olfaction. PLoS One 11, e0157560.

Potter, H., Butters, N., 1980. An assessment of olfactory deficits in patients with damage to prefrontal cortex. Neuropsychologia 18, 621–628.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408.

Shalev-Shwartz, S., Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15, 72–101.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Thomas-Danguin, T., Rouby, C., Sicard, G., Vigouroux, M., Farget, V., Johanson, A., Bengtzon, A., Hall, G., Ormel, W., De Graaf, C., Rousseau, F., Dumont, J.P., 2003. Development of the ETOC: a European test of olfactory capabilities. Rhinology 41, 142–151.

Toledano, A., Ruiz, C., Navas, C., Herráiz, C., González, E., Rodríguez, G., Galindo, A.N., 2009. Development of a short olfactory test based on the Connecticut Test (CCCRC). Rhinology 47, 465–469.

Turing, A., 1950. Computing Machinery and Intelligence. Mind LIX. pp. 433–460.

Ultsch, A., Lötsch, J., 2015. Computed ABC analysis for rational selection of most informative variables in multivariate data. PLoS One 10, e0129767.

Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H., 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. 31, 306–315.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.

Walker, S.H., Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika 54, 167–179.

Wei, T., Simko, V., 2017. R Package "Corrplot": Visualization of a Correlation Matrix.

Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klaR Analyzing German Business Cycles. Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.

Wilson, R.S., Schneider, J.A., Arnold, S.E., Tang, Y., Boyle, P.A., Bennett, D.A., 2007. Olfactory identification and incidence of mild cognitive impairment in older age. Arch. Gen. Psychiatry 64, 802–808.

Yilmaz, Y., Karakas, Z., Uzun, B., Sen, C., Comoglu, S., Orhan, K.S., Aydogdu, S., Karagenc, A.O., Tugcu, D., Karaman, S., Wylie, C., Doty, R.L., 2017. Olfactory dysfunction and quality of life in patients with transfusion-dependent thalassemia. Eur. Arch. Otorhinolaryngol. 274, 3417–3421.