JOURNAL OF FORENSIC SCIENCES

EVALUATION OF DISTANCE-BASED APPROACHES FOR FORENSIC COMPARISON: APPLICATION TO HAND ODOR EVIDENCE

Journal:	*Journal of Forensic Sciences*
Manuscript ID	Draft
Manuscript Type:	Paper
Keywords:	Bayesian inference, dissimilarity measure, forensic science, human hand odor, likelihood ratio, logistic regression

EVALUATION OF DISTANCE-BASED APPROACHES FOR FORENSIC COMPARISON: APPLICATION TO HAND ODOR EVIDENCE

ABSTRACT

The issue of distinguishing between the same-source and different-source hypotheses based on various types of traces is a generic problem in forensic science. This problem is often tackled with Bayesian approaches, which are able to provide a likelihood ratio that quantifies the relative strengths of evidence supporting the two competing hypotheses. Here, we focus on distance-based approaches, whose robustness and especially capacity to deal with high-dimensional evidence are very different, and need to be evaluated and optimized.

A unified framework for direct methods based on estimating the likelihoods of the distance between traces under the two competing hypotheses, and indirect methods using logistic regression to discriminate between same-source and different-source distance distributions, is presented. They are compared in terms of sensitivity, specificity and robustness, with and without dimensionality reduction, with and without feature selection, on the example of hand odor evidence. Empirical evaluations on a large panel of 534 subjects show the significant superiority of the indirect methods, especially without dimensionality reduction.

KEYWORDS

Bayesian inference; dissimilarity measure; forensic science; human hand odor; likelihood ratio; logistic regression.

HIGHLIGHTS

- Direct and indirect distance-based likelihood ratio estimation methods for forensic comparison are investigated
- They are applied to high-dimensional evidence consisting of hand odor traces
- Their robustness, AUC, sensitivity and specificity are evaluated on a panel of 534 subjects
- Indirect methods based on logistic regression outperform direct ones and are more robust
- Indirect methods using a vectorial distance outperform those using a scalar one, with and without feature selection

1. INTRODUCTION

A generic problem in forensic science is to decide whether a trace of an unknown source, often drawn from a crime scene, and a trace from a known source, stem from the same source, a person or a firearm for example. If the source is a person, the traces might be biometric such as a DNA profile [Aitken and Taroni 2004] [Puch-Solis et al. 2012], fingerprints [Neuman et al. 2012], a voice [Morrison 2011], an olfactory profile [Cuzuel et al. 2017], or they might consist of footwear impressions [Tang and Srihari 2014bis], handwriting [Tang and Srihari 2014], etc. If the source is a firearm, the traces may be features such as striations and impressions of a bullet or of a cartridge case [Mattijssen et al. 2020].

The most common approach to this problem is to estimate a likelihood ratio (LR), i.e. the ratio of the joint probability of occurrence of the two traces under the hypothesis that they arose from the same source and under the hypothesis that they arose from different sources. A convenient solution is to replace the joint probability of the traces by the probability of a distance between the two traces quantifying their dissimilarity [Neuman et al. 2007, Riva and Champod 2014, Tang and Srihari 2014bis, Ali et al. 2015, Muehlethaler et al. 2016, Riva et al. 2020, Vergeer et al. 2020]. If, as is most often the case, the distance is scalar, there is an important loss of information. Thus, we choose to focus on distance-based methods, but with the possibility to use a vectorial distance between traces.

Furthermore, the distance-based LR estimate can be obtained either directly, by estimating the distance likelihoods under the two hypotheses, or indirectly, by first using logistic regression to discriminate between same-source and different source distance distributions, and then Bayes' formula to infer the LR. Whilst the direct method is more flexible, the indirect method is more robust and quite natural in machine learning [Bishop 2006, Hastie et al. 2009]. It is sometimes advocated for in the forensic context, for the same reasons and also because it enables score calibration and fusion with minimal mathematical complexity [Enzinger et al. 2016, Morrison 2013]. Here, we show that the indirect method also enables the use of a vectorial distance, thus preventing the severe information loss suffered by scalar distance approaches. We discuss the direct and indirect methods in terms of robustness and ability to handle high dimensional evidence, with or without dimensionality reduction, and with or without feature selection. We evaluate them in terms of sensitivity, specificity and robustness on the example of traces consisting in a hand odor profile.

2. MATERIALS AND METHODS

2.1 Problem statement

The aim is, given the evidence consisting of a pair of traces (e.g. two olfactory profiles), to decide whether these traces have the same source (e.g. the same person) or not. In the following, H_{ss} refers to the hypothesis that the two traces stem from the same source, and H_{ds} to the alternative hypothesis that they stem from different sources. Given the *a priori* probabilities $P(H_{ss})$ and $P(H_{ds})$, the Bayesian formula yields the posterior probability of H_{ss} given the evidence E:

$$P(H_{ss} \mid E) = \frac{f(E \mid H_{ss})P(H_{ss})}{f(E|H_{ss})P(H_{ss}) + f(E|H_{ds})P(H_{ds})} \quad (1)$$

where $f(E|H_{ss})$ and $f(E|H_{ds})$ are the distributions of the evidence under H_{ss} and H_{ds}, or likelihoods. Jeffreys developed an absolute scale to evaluate the degree of confidence in the same-source hypothesis outside a decisional framework based on the posterior probability of H_{ss} using the LR [Jeffreys 1939, Robert 2001] defined as:

$$LR(E) = \frac{f(E \mid H_{ss})}{f(E \mid H_{ds})} \quad (2)$$

which is independent from the *a priori* probabilities. In fact, the observed evidence E consists of the two traces which are represented by n-dimensional vectors (whose components are the amounts of each odor compound). Since we focus on distance-based methods, the information contained in these two vectors is transformed into a distance or dissimilarity measure, which can be either a scalar or a n-vector (a distance for each feature of the trace, here for each odor compound). In the following, this distance between the two traces will be denoted by *d*.

2.2 Candidate methods

The LR can be obtained either directly, i.e. by estimating the likelihoods of the distance under the two competing hypotheses, or indirectly, i.e. using first logistic regression to discriminate between same-source and different source distance distributions, and then formulas (1) and (2) to infer the LR.

2.2.1 Direct methods

For direct methods, we need to be able to evaluate $f(d|H_{ss})$ and $f(d|H_{ds})$ whatever the value of d. For this purpose, part of the available dataset can be used to build pairs of traces of the two types, same-source and different-source pairs. If d is scalar, or of dimension 2 or 3 at most, the two empirical distributions can be fitted, with Gaussian mixtures, for example, leading to parametric estimates of $f(d|H_{ss})$ and $f(d|H_{ds})$ [Mattijssen et al. 2020, Riva and Champod 2014, Riva et al. 2020]. In the case of many features, a fit of each component of d can be performed in the same way, and the overall likelihoods can be approximated through the product of the likelihoods in each dimension, leading to the naïve Bayes classifier. To be successful, the latter approach necessitates however that the features are not overly correlated.

2.2.2. Indirect methods

The aim of these method is to build a discriminative model of the boundary between the two categories of pairs (same-source and different-source pairs) rather than a generative model explicitly parameterizing the distributions in the two categories. It is well known that, under the hypothesis of single-Gaussian distributions with the same variance under H_{ss} and H_{ds} in the scalar case, or same covariance matrix in the multidimensional case, the posterior probability of H_{ss} takes the form of a sigmoidal curve [Bishop 2006, Hastie et al. 2009], hence the motivation for a logistic regression approach. Despite its result being a discriminative model, it enables to calculate the posterior probability of Equation (1) as well as the LR of Equation (2). As a matter of fact, the logistic regression model with parameters $\theta = [a^T\ b]^T$ has output:

$$r(d,\theta) = \frac{1}{1+\exp\left(-(a^T d + b)\right)} \tag{3}$$

where *b* is a scalar, and *a* is either a scalar in the case of a scalar distance, or otherwise a n-vector (of the dimension of the evidence). If the proportions of the same-source and different-source categories in the calibration set are denoted by f_{ss} and f_{ds}, $r(d, \theta)$ approximates:

$$\frac{f(d\,|\,H_{ss})f_{ss}}{f(d\,|\,H_{ss})f_{ss} + f(d\,|\,H_{ds})f_{ds}} \tag{4}$$

Thus, the posterior probability for *a priori* probabilities $P(H_{ss})$ and $P(H_{ds})$ can be retrieved with:

$$P(H_{ss} \mid d) = \frac{1}{1 + \exp\left(-(a^T d + b)\right) \dfrac{f_{ss}}{f_{ds}} \dfrac{P(H_{ds})}{P(H_{ss})}} \tag{5}$$

and the likelihood ratio with:

$$LR(d) = \exp\left(a^T d + b\right) \frac{f_{ds}}{f_{ss}} \tag{6}$$

2.3. Pros and cons

The indirect method offers several advantages:

- it spares the necessity to fit the likelihoods,
- in the multi-dimensional case, contrary to the naïve Bayes classifier, the independence assumption is not necessary, because the logistic regression automatically takes care of the correlation between features,
- by construction, the log LR is defined by a hyperplane, and thus robust with respect to the equal variance assumption, and to outliers or sparse data far from the boundary,
- in the forensic context, since the log LR is directly proportional to $a^T d + b$, the logistic allows a convenient and interpretable calibration of the dissimilarity score d, and a fusion of scores in the multidimensional case [Morrison 2013].

On the other hand, the indirect method might suffer from:

- a reduced flexibility since it amounts to assume single-Gaussian distributions,
- a possibly important computation time in the case of high-dimensional evidence and of a distance of the same dimension.

These advantages and disadvantages will be examined and discussed on the example of hand odor evidence using a large panel of subjects.

2.4. Dataset description

A panel of 534 volunteers was set up which gathers 218 men and 316 women aged 7 to 94 years (median 28, interquartile interval [22 ; 48]), see Table 1 for the detailed composition in terms of sex and age. Note that this composition does not aim at reflecting that of a precise target population, such as one which is more likely to commit a crime, but to be as representative of the diversity of odors as possible. As a matter of fact, criminal investigations also often necessitate to look for victims, or to discriminate between traces from different people present at a crime scene, including those of victims or witnesses, who might be women as well as men, children or seniors as well as middle-aged adults. All data were completely anonymized prior to analysis, and no personal information was stored.

The goal was here to identify the subjects by their hand odor, whose volatile profile was shown to display a between-subject variability which is sufficient for differentiation [Curran et al. 2010, Cuzuel et al. 2017]. Also, in the forensic context, the hands have the advantage to be more likely to be directly in contact with objects at a crime scene, and to be easier to sample during a police interrogation.

The volatile profiles were obtained by a direct sampling procedure using identical sample collection kits of 4 small polymer strands that the subjects were asked to rub together in their hands for 15 minutes. The polymer strands were thermodesorbed, and the concentrated substances were separated by comprehensive bidimensional gas chromatography (GCxGC) coupled with mass spectrometry (MS). The sampling method and the optimization of the GCxGC-MS analysis were extensively described in [Cuzuel et al. 2017bis, Cuzuel et al. 2018]. Data were acquired, converted to .mzXML files with GC Real Time Analysis 4.20 (Shimadzu software), and then processed with MatlabTM (Natick, MA, USA) version 9.6.0.1150989 (R2019a), its Statistics and Machine Learning Toolbox version 11.5 and its Bioinformatics Toolbox version 4.12.

Using a “home-made” Matlab script [Cuzuel 2017], the preliminary manual processing of 25 chromatograms obtained on 3 subjects between 23 and 26 years old of both genders sampled several times at different time instants enabled us to draw up a first list of several hundreds of peaks. A library was built to store their retention times, their linear retention index, their mass spectrum, and the name of the corresponding compound when it could be identified using the NIST library. Indeed, if the availability of its mass spectrum is compulsory, a compound does not need to be formally identified for the comparison of chromatograms. We also checked whether compounds described in the literature as constituents of the human hand odor were present in this library, otherwise they were included. The library was then continuously enriched as the panel was increased with compounds potentially relevant to human hand odor because of their empirical frequency in new samples. This work led us to a customized library of 741 compounds, which were looked for in each chromatogram. As a result, each sample was characterized by the peak area of 741 compounds.

In order to compensate for uncontrolled variations of the total area of the chromatograms, the sum of these areas was normalized in logarithmic scale to unit value, see Table 2 for a comparison of the reproducibility of the data without normalization, and normalization in scalar and logarithmic scales. Not knowing whether all 741 compounds are really relevant for identification (the median frequency of presence of the compounds across all the samples is of 40.2 %, quartiles [12.8 %; 77.6 %]), such a normalization might be questionable. Thus, the possibility to avoid the problem by working on the binarized areas, i.e. 1 if the compound is present, or 0 if it is absent from the sample, was also investigated. Also, this approach might be of interest for forensic identification problems dealing with intrinsically binary features, such as gradient, structural and concavity (GSC) binary features in handwriting identification [Srihari et al. 2008]. In the following, we refer to these traces of 741 features, continuous or binarized, as "odor traces".

As stated above, the subjects were sampled in quadruplicate, but due to unavoidable mishaps with some samples (like accidently dropping a polymer on the floor during sampling) and to chromatographic problems (such as failures of the cryogenic modulator), 1690 odor traces were obtained for the 534 subjects (44 were sampled once, 77 twice, 160 three times, and the remaining 253 subjects four times, leading to an average of 3.2 odor traces per subject). This data set was split into a calibration set for training and validation, and an independent test set for performance estimation. The split was made so as to respect the gender proportions, with subject of all ages in the two sets, and odor traces of the same subject being put in the same set. As a result, the calibration set comprises 412 subjects and their 1 299 odor traces (corresponding to 1 594 H_{ss} and 841 457 H_{ds} pairs), and the test set comprises the remaining 122 subjects and their 391 odor traces (leading to 481 H_{ss} and 75 764 H_{ds} pairs). The way the odor traces distribute between calibration and test set can be grasped through the Principal Component Analysis (PCA) of Figure 1.

2.5. Implementation

Three methods are implemented:

1) the direct method using a scalar distance between odor traces,

2) the indirect method using a scalar distance between odor traces,

3) the indirect method using a vectorial distance, i.e. a distance on each odor compound.

Note that, given the large dimension of the problem (n=741) and the known correlations between features, we did not attempt to implement the direct method using a vectorial distance (i.e. a scalar distance on each feature and the naïve Bayes classifier).

2.5.1. Distances between two odor traces

Concerning the choice of the scalar distance, correlation-based distances are robust with respect to shifts and linear transformations of the features, and since Spearman's correlation coefficient is able to capture a monotonic nonlinear association, as opposed to Pearson's linear correlation coefficient [Daniels 1944], the Spearman correlation based distance is also expected to be more robust with respect to nonlinear variations of peak areas. Since this proved to be true in a previous study where Euclidian distance, Pearson correlation and Spearman correlation based distances were compared [Cuzuel et al. 2018bis], we restrict here to Spearman's correlation based distance for the direct and indirect methods using a scalar distance. The chosen vectorial distance for the third method is simply the vector of the absolute differences between feature values.

2.5.2. Estimation of the likelihoods for the direct method

The calibration set was used to build pairs of odor traces of same and different sources, and to compute their distances. The empirical densities were fitted with a two-Gaussian mixture distribution, using Matlab's function "fitgmdist", leading to estimates of the likelihoods $f(d|H_{ss})$ and $f(d|H_{ds})$.

2.5.3. Estimation of the logistic model for the indirect methods

The logistic regression model of Equation (3) with parameters $\theta = [a^T\ b]^T$ was fitted to minimize the cross-entropy cost function using Matlab's function "glmfit".

2.5.4. Likelihood ratio and performance estimation

For the direct method, the LR was evaluated using the estimates of the likelihoods $f(d|H_{ss})$ and $f(d|H_{ds})$ and Equation (2), as a function of the distance d. For the indirect methods, the LR was obtained from the fitted logistic regression and Equation (6), and plotted as a function of the distance d or of the score $a^T d + b$, depending on d being scalar of vectorial.

Since there is no true reference for the LR, the performance of the different methods was evaluated by estimating the posterior probability $P(H_{ss}|d)$ according to Equations (1) and (5) for the direct and indirect methods respectively, and by performing a binary classification using equal prior probabilities ($P(H_{ss}) = P(H_{ds}) = 0.5$). Varying the decision threshold on $P(H_{ss}|d)$, the sensitivity and the specificity were estimated on the calibration and test sets, and used to compute the corresponding areas under the receiver operating characteristic

"ROC" curve (AUC) [Hanley and McNeil 1982]. The performance was further characterized by the sensitivity and specificity maximizing Youden's index [Youden 1950], i.e. their sum.

2.5.5. Feature selection

In a previous study [Cuzuel et al. 2018bis], improved results were obtained using feature selection. Given the large number of features (odor compounds) and the large size of the data set, an economic and robust filter approach to this selection was chosen. The idea is to retain the features that contribute the most to the difference between densities under H_{ss} and H_{ds}. For each feature, using the absolute values of the difference for the H_{ss} and H_{ds} pairs, we computed Wilcoxon's non-parametric test statistic in the case of continuous features, and Fisher's exact test statistic in the case of binarized features. Then, the features were ranked in decreasing order of the one-sided p-value of the test (it is a one-sided test since smaller differences between features under H_{ss} than under H_{ds} are sought for). The number of features maximizing the AUC was estimated on the calibration set using 3-fold cross-validation. The cross-validation partitions were randomly chosen with the constraint that the odor traces of the same subject were put in the same partition. Note that cross-validation also enabled us to estimate the uncertainty on the AUCs through the mean standard deviation on the three partitions.

3. RESULTS AND DISCUSSION

The three methods are first evaluated using all the features of the odor traces (baseline comparison) and then, the possibility to further improve their performance using feature selection is investigated.

3.1. Baseline comparison of the three methods (without feature selection)

The results obtained with the three methods on the calibration and test sets are summarized in Tables 3 and 4 for binarized and continuous features respectively. As a first remark, the performance of the direct and indirect methods using a scalar distance in terms of AUC and of sensitivity and specificity are almost identical, for both binarized and continuous features. Thus, the higher flexibility of the direct method does not increase the performance. On the contrary, its lack of robustness can be visualized on Figure 2 depicting the posterior probability and the LR obtained with the binarized features: due to the larger variance of the likelihood under H_{ss}, the posterior probability and the LR, instead of being monotonous, start to increase with the distance at some point ($d \approx 0.7$). Whereas whatever the situation with the indirect method, posterior probability and LR always decrease with d, see Figure 3 depicting the posterior probability and the LR obtained with indirect method, this time on the continuous features.

Also noteworthy, the performance obtained with the indirect method using a vectorial distance is significantly better than those of the methods working with a scalar distance: the AUC on the calibration and test sets jumps from 91-92% to 97-98%, the standard deviation of the AUC being estimated at 0.7% using 3-fold cross-validation on the calibration set. The distributions of the score ($a^T d + b$) resulting from the logistic regression, the regression itself, the posterior probability and the LR are shown in Figure 4. The only drawback lies in the increased, but perfectly tractable computational cost (10 minutes instead of a few seconds, on a 4,2 GHz Intel Core i7).

Finally, the binarization of the features decreases the performance, but only marginally (the AUC is decreased by ≈ 1%). Note that, in this precise case where the features quantify the amount of odor compounds, this could be due to the fact that the normalization of the compound proportion uses all these compounds whereas it is not known whether they are all relevant. Note also that the normalization was improved by performing it in the logarithmic scale rather than in the linear scale (the former improving the reproducibility, see Table 2), with which continuous features did not outperform binarized features, as shown in a previous study [Cuzuel et al. 2018bis]. Finally, other normalization methods specific to GCxCG-MS data might advantageously be investigated [Chen et al. 2017], but are outside the scope of this paper.

3.2. Comparison of three methods with feature selection

The number of selected features using the filter approach is reported in Tables 5 and 6, together with the corresponding results on the calibration and test sets, for binarized and continuous feature respectively.

Again, there is almost no difference in performance between the direct and indirect methods with a scalar feature, be it on binarized or continuous features. In terms of AUC, the selection is more efficient on binary features than on continuous ones (94.4% with selection instead of 91.5% without for binarized features, 93.5% instead of 93.0% for continuous features, on the test set), with an important reduction of the number of the binarized features (267 instead of 741), and a moderate one for continuous features (535 instead of 741). Note also that in both cases, this increased performance benefits the specificity, which is highly desirable in a forensic application (it is crucial in this context not to reject the different-source hypothesis, i.e. the defense hypothesis, when in fact it is true).

For both binarized and continuous features, the indirect method using a vectorial distance is again significantly better than the previous ones (AUCs around 97-98% instead of 94-95% on both calibration and test sets), with a similar number of selected features (440 for binarized features, 500 for continuous ones). However, in both cases, the parsimony due to feature selection does not increase the performance as compared to the baseline method, it is quasi-identical with and without selection. In return, this testifies to a robustness of the indirect method with respect to possibly irrelevant features. And of course, not to have to perform the selection spares computation time.

3.3. Discussion of the choice of equal priors

In this manuscript, the methods are compared in terms of AUC, sensitivity and specificity. In the case of the indirect methods, the regression being obtained by fitting a logistic function to the data, whatever the prior probabilities $P(H_{ss})$ and $P(H_{ds})$, the posterior probability $P(H_{ss}|d)$ given by Equation (5) is also a logistic function. Thus, when the threshold on $P(H_{ss}|d)$ is varied from 1 to 0, the same ROC curve is described, whose AUC only depends on the distance distributions under H_{ss} and H_{ds}: only the threshold maximizing Youden's index changes.

With the direct method, the choice of the prior has an influence on the shape of the posterior probability, so that the threshold on $P(H_{ss}|d)$ can possibly be varied in a different interval (see Figure 3 where $P(H_{ss}|d)$ never reaches 0 for example). However, in practice, there is no influence on AUC, sensitivity and specificity because, again, the AUC depends essentially on the distance distributions under H_{ss} and H_{ds}. The only palpable change is on the threshold yielding the best compromise between sensitivity and specificity, threshold which adjusts to $P(H_{ss})$ by roughly following it.

Thus, the assumption of equal prior probabilities has practically no impact on the LR estimate.

3.4. Limitations

From a practical point of view, our work suffers several limitations for a real-world forensic application. First, for practical reasons, the subjects were sampled at a single time point, so that the variability of the data is essentially due to the analytical variability. Second, the chromatograms being compared are of the same nature, i.e. obtained on samples provided by directly sampling the subjects (with contact with the subjects' hands) whereas in real life, the unknown source sample will be obtained indirectly from an object on the crime scene (without contact with the subject). Third, the odor collected on the crime scene might be contaminated by other odors, from the environment or from other people. A study focused on mixtures of odors, contaminations, and weathered traces has not been carried out yet but is considered. However, despite these controlled conditions, the PCA of Figure 1 and the statistics of Table 2 show that the data is already of limited reproducibility, so that the good results we have obtained are encouraging concerning the robustness of the best method to more realistic sampling conditions.

From a methodological point of view, the proposed methods are based on a common source scenario, where it is asked whether the two traces originate from the same source or from different sources without specifying which sources are considered, and not on a specific source scenario, where the question is whether the two traces stem specifically from the known source [Neuman and Ausdemore 2020]. The problem of the common source scenario is that it does not take account of the typicality of the source, contrary to recommendations for a better estimation of the strength of evidence through the LR [Morrison 2013, Tang and Srihari 2014, Morrison and Enzinger 2018]. But to implement a specific source scenario, a number of traces from the known source are needed in order to be able to estimate the distribution under H_{ss} (for the direct method) or to discriminate between the H_{ss} and H_{ds} populations (for the indirect methods), which is quite unpractical when dealing with human hand odor, and not feasible at this stage of the study (at most four usable odor traces were obtained, for only 253 subjects among the 534).

4. CONCLUSIONS

To summarize, the advantages expected from an indirect method are fully obtained, in particular the dispensation to parameterize the likelihoods, and the robustness with respect to differences in their variance and/or to possible outliers. Moreover, an increase in performance of the indirect method as compared with the direct one is not obtained with a scalar distance between odor traces, but when using the vector of the distances between each feature of the odor traces. This improvement was not really expected, because, especially in the forensic context, it is often advocated to convert multivariate data to a univariate datum summarizing the relationship between features. Finally, the indirect method with a vectorial distance proves also robust with respect to potentially irrelevant features since removing them does not modify the performance, an appealing quality for dealing with traces which are not yet solidly characterized, such as odor traces.

REFERENCES

T. Ali, L. Spreeuwers, R. Veldhuis, D. Meuwly (2015), Sampling variability in forensic likelihood-ratio computation: A simulation study, Science & Justice 55 (6), 499-508, https://doi.org/10.1016/j.scijus.2015.05.003

C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

J. Chen, P. Zhang, M. Lv, H. Guo, Y. Huang, Z. Zhang, and F. Xu (2017), Influences of Normalization Method on Biomarker Discovery in Gas Chromatography–Mass Spectrometry-Based Untargeted Metabolomics: What Should Be Considered? Analytical Chemistry 89 (10), 5342-5348 , https://doi.org/10.1021/acs.analchem.6b05152

C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Second Edition (2004).

A. M. Curran, P. A. Prada, K. G. Furton (2010), The differentiation of the volatile organic signatures of individuals through SPME-GC/MS of characteristic human scent compounds. Journal of forensic sciences 55(1), 50–57, doi:10.1111/j.1556-4029.2009.01236.x

V. Cuzuel (2017), Développement d'une stratégie de caractérisation chimique de la signature odorante d'individus par l'analyse chimiométrique de données issues de méthodes séparatives bidimensionnelles, Thèse de doctorat de l'Université Pierre et Marie Curie, ED388. https://hal.archives-ouvertes.fr/tel-01667972.

V. Cuzuel, G. Cognon, I. Rivals, C. Sauleau, F. Heulard, D. Thiébaut, J. Vial (2017), Origin, analytical characterization and use of human odor in forensics, Journal of forensic sciences 62, 330–350, https://doi.org/10.1111/1556-4029.13394

V. Cuzuel, E. Portas, G. Cognon, I. Rivals, F. Heulard, D. Thiébaut, J. Vial (2017bis), Sampling method development and optimization in view of human hand odor analysis by thermal desorption coupled with gas chromatography and mass spectrometry., Analytical and Bioanalytical Chemistry 409, 5113–5124. doi:10.1007/s00216-017-0458-8.

V. Cuzuel, G. Cognon, I. Rivals, F. Heulard, D. Thiébaut, J. Vial (2018), Human odor and forensics. Optimization of a comprehensive gas chromatography method based on orthogonality: how not to choose between criteria., J. Chromatogr. A. 1536 58-66, doi:10.1016/j.chroma.2017.08.060.

V. Cuzuel, R. Leconte , G. Cognon, D. Thiébaut, J. Vial , C. Sauleau , I. Rivals (2018bis), Human odor and forensics: towards Bayesian suspect identification using GCxGC-MS characterization of hand odor, *Journal of Chromatography B* 1092, 379-385, https://www.sciencedirect.com/science/article/abs/pii/S1570023218306068?via%3Dihub

H. E. Daniels (1944), The relation between measures of correlation in the universe of sample per-mutation, Biometrika, Volume 33, Issue 2, 129-135, https://doi.org/10.1093/biomet/33.2.129.

E. Enzinger, G. S. Morrison, F. Ochoa (2016) A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Science & Justice 56(1), 42-57, https://doi.org/10.1016/j.scijus.2015.06.005.

H. Jeffreys, Theory of probability, Oxford University Press, Oxford, 1939.

T. Hastie, R. Tibshirani, Friedman J. The elements of statistical learning: data mining, inference, and prediction, 2nd ed., Springer, New York, 2009.

A. J. Hanley, J.B. McNeil (1982), The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, Radiology. 143, 29–36. doi:10.1148/radiology.143.1.7063747

Erwin J A T Mattijssen , Cilia L M Witteman, Charles E H Berger, Nicolaas W Brand, Reinoud D Stoel (2020), Validity and Reliability of Forensic Firearm Examiners, Forensic Science International 307 110112, https://doi.org/10.1016/j.forsciint.2019.110112

G. S. Morrison (2011), A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM), Speech Communication 53 (2), 242-256, https://doi.org/10.1016/j.specom.2010.09.005

G. S. Morrison (2013), Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. Australian Journal of Forensic Sciences, 45(2), 173-197. https://doi.org/10.1080/00450618.2012.733025

G. S. Morrison, E. Enzinger (2018) Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality, Science & Justice 58 (1), 47-58, https://doi.org/10.1016/j.scijus.2017.06.005.

C. Muehlethaler, G. Massonnet, T. Hicks (2016), Evaluation of infrared spectra analyses using a likelihood ratio approach: A practical example of spray paint examination, Science & Justice, 56 (2), 61-72, https://doi.org/10.1016/j.scijus.2015.12.001.

C. Neuman, C. Champod, R. Puch‐Solis, N. Egli, A. Anthonioz, A. B.‐Griffiths (2007), Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, Journal of Forensic Sciences, 52(1), 54-64, https://doi.org/10.1111/j.1556-4029.2006.00327.x

C. Neumann, Ian W Evett, J. Skerrett (2012), Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm, Journal of the Royal Statistical Society Series A (Statistics in Society) 175(2), 1-26 https://doi.org/10.1111/j.1467-985X.2011.01027.x

C. Neuman, M.A. Ausdemore (2020), Defence Against the Modern Arts: the Curse of Statistics -Part II: "Score-based likelihood ratios", *Law, Probability and Risk* 19(1), 21–42, https://doi.org/10.1093/lpr/mgaa006

F. Riva, C. Champod (2014), Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, Journal of forensic sciences 59(3), 637-47, https://doi.org/10.1111/1556-4029.12382

F. Riva, E. J.A.T. Mattijssen, R. Hermsen, P. Pieper, W. Kerkhoff, C. Champod (2020), Comparison and interpretation of impressed marks left by a firearm on cartridge cases – Towards an operational implementation of a likelihood ratio based technique, Forensic Science International 313, 110363, https://doi.org/10.1016/j.forsciint.2020.110363.

C. Robert, The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation, Springer, New York, 2001.

S.N. Srihari, C. Huang, H. Srinivasan (2008) On the Discriminability of the Handwriting of Twins, Journal of Forensic Sciences 53(2), 430–446, https://doi.org/10.1111/j.1556-4029.2008.00682.x

Y. Tang, S. N. Srihari (2014) Likelihood ratio estimation in forensic identification using similarity and rarity, Pattern Recognition 47, 945-958, https://doi.org/10.1016/j.patcog.2013.07.014

Y. Tang , S. N. Shrihari (2014bis) Computational methods for the analysis of footwear impression evidence, In "Computational Intelligence in Digital Forensics: Forensic Investigation and Application", A. Muda, Y-H Choo, A. Abraham and S. N. Srihari (eds.), Springer 2014, pages 333-383.

P. Vergeer, J. N. Hendrikse, M. M. P. Grutters, L. J. C. Peschier (2020). A method for forensic gasoline comparison in fire debris samples: A numerical likelihood ratio system. Sci Justice. 2020 Sep;60(5):438-450, https://doi.org/10.1016/j.scijus.2020.06.002

W. J. Youden (1950). Index for rating diagnostic tests. Cancer. 3, 32–35, https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

FIGURE CAPTIONS

Figure 1. PCA of the data set showing the distribution of the odor traces between calibration and test sets. The PCA was performed on the covariance matrix of the continuous features, i.e. normalized in logarithmic scale.

Figure 2. Results of the direct method with the scalar distance d on binarized feature, as functions of d, on the calibration set. a) Empirical distribution of d under H_{ss} (1 594 pairs), and estimated density (mixture of 2 Gaussians). b) Empirical distribution of d under H_{ds} (841 457pairs), and estimated density (mixture of 2 Gaussians). c) Posterior probability of H_{ss} obtained using Bayes' formula with equal priors. d) Likelihood ratio.

Figure 3. Results of the indirect method with the scalar distance d on continuous features, as functions of d, on the calibration set. a) Empirical distribution of d under H_{ss} (1 594 pairs). b) Empirical distribution of d under H_{ds} (841 457pairs). c) Logistic regression (dotted line), and deduced posterior probability of H_{ss} with equal priors (continuous line). d) "Likelihood ratio" $\exp(a^T d + b)$ corresponding to the logistic regression (dotted line), and likelihood ratio (continuous line).

Figure 4. Results of the indirect method with the vectorial distance d on continuous features, as functions of the score $a^T d + b$, on the calibration set. a) Empirical distribution of the score $a^T d + b$ under H_{ss} (1 594 pairs). b) Empirical distribution of the score $a^T d + b$ under H_{ds} (841 457pairs). c) Logistic regression (dotted line), and deduced posterior probability of H_{ss} with equal priors (continuous line). d) "Likelihood ratio" $\exp(a^T d + b)$ corresponding to the logistic regression (dotted line), and likelihood ratio (continuous line).

TABLE CAPTIONS

Table 1. Panel composition in terms of sex and age.

Table 2. Reproducibility of the odor trace features (peak areas), without and with two normalizations, estimated on the 490 subjects sampled at least twice (IQI stands for interquartile interval).

Table 3. Baseline comparison between the three methods on the calibration and test sets, using binarized features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, all in %.

Table 4. Baseline comparison between the three methods on the calibration and test sets, using continuous features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, all in %.

Table 5. Comparison between the three methods on the calibration and test sets, using binarized features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, with selection of the number of features among the 741 by cross-validation on the calibration set (AUC-CV3 denotes the mean 3-fold cross-validation AUC on the calibration set, and #feat. the number of features selected by cross-validation).

Table 6. Comparison between the three methods on the calibration and test sets, using continuous features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, with selection of the number of features among the 741 by cross-validation on the calibration set (AUC-CV3 denotes the mean 3-fold cross-validation AUC on the calibration set, and #feat. the number of features selected by cross-validation).

TABLES

Table 1. Panel composition in terms of sex and age.

age / sex	7-17	18-64	65-94	total
man	18	182	18	218
woman	28	268	20	316
total	46	450	38	534

Table 2. Repeatability of the odor trace features (peak areas), without and with two normalizations, estimated on the 490 subjects sampled at least twice (IQI stands for interquartile interval).

Normalization	**Median relative standard deviation [IQI] in %**
None	61.0 [32.1 ; 78.0]
In scalar scale	56.1 [31.9 ; 74.2]
In logarithmic scale	33.2 [17.9 ; 48.8]

Table 3. Baseline comparison between the three methods on the calibration and test sets, using binarized features, in terms of AUC, sensitivity (Sn) and specificity (Sp) for the threshold maximizing Youden's index, all in %.

	Calibration				Test			
Method	AUC	threshold	Sn	Sp	AUC	threshold	Sn	Sp
Direct	91.2	0.43	80.6	91.3	91.4	0.54	78.4	94.9
Indirect scal. d	91.2	0.62	80.6	91.3	91.5	0.75	78.4	94.9
Indirect vect. d	98.5	0.54	93.4	96.3	97.1	0.56	91.1	94.2

Table 4. Baseline comparison between the three methods on the calibration and test sets, using continuous features, in terms of AUC, sensitivity (Sn) and specificity (Sp) for the threshold maximizing Youden's index, all in %.

	Calibration				Test			
Method	AUC	threshold	Sn	Sp	AUC	threshold	Sn	Sp
Direct	92.1	0.44	81.1	92.5	93.0	0.50	81.3	94.6
Indirect scal. d	92.1	0.64	81.1	92.5	93.0	0.73	81.3	94.6
Indirect vect. d	98.9	0.58	94.4	97.0	97.8	0.57	91.3	95.1

Table 5. Comparison between the three methods on the calibration and test sets, using binarized features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, with selection of the number of features among the 741 by cross-validation of the calibration set (AUC-CV3 the mean 3-fold cross-validation AUC on the calibration set, and #feat. denotes the number of features selected by cross-validation).

			Calibration			Test		
Method	AUC-CV3	#feat.	AUC	Sn	Sp	AUC	Sn	Sp
Direct	94.4	267	94.5	80.9	94.7	94.4	84.6	92.8
Indirect scal. d	94.4	267	94.5	80.9	94.7	94.4	84.6	92.8
Indirect vect. d	96.4	440	97.6	89.5	97.5	97.0	91.3	94.3

Table 6. Comparison between the three methods on the calibration and test sets, using continuous features, in terms of AUC, sensitivity (Sn) and specificity (Sp) maximizing Youden's index, with selection of the number of features among the 741 by cross-validation of the calibration set (AUC-CV3 the mean 3-fold cross-validation AUC on the calibration set, and #feat. denotes the number of features selected by cross-validation).

			Calibration			Test		
Method	AUC-CV3	#feat.	AUC	Sn	Sp	AUC	Sn	Sp
Direct	93.1	535	93.1	81.3	93.3	93.5	82.3	95.5
Indirect scal. d	93.1	535	93.1	81.3	93.3	93.5	82.3	95.5
Indirect vect. d	97.0	500	98.3	91.6	97.3	97.7	91.7	94.9

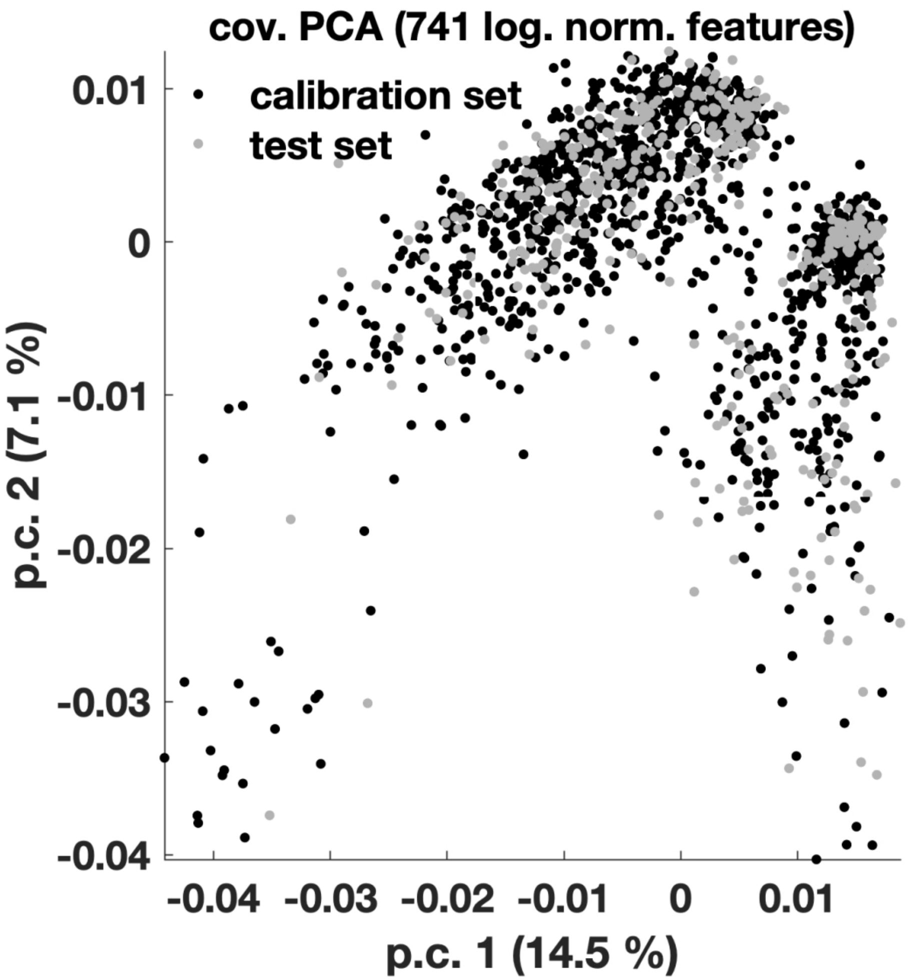

Figure 1: PCA of the data set showing the distribution of the odor traces between calibration and test sets. The PCA was performed on the covariance matrix of the continuous features, i.e. normalized in logarithmic scale.

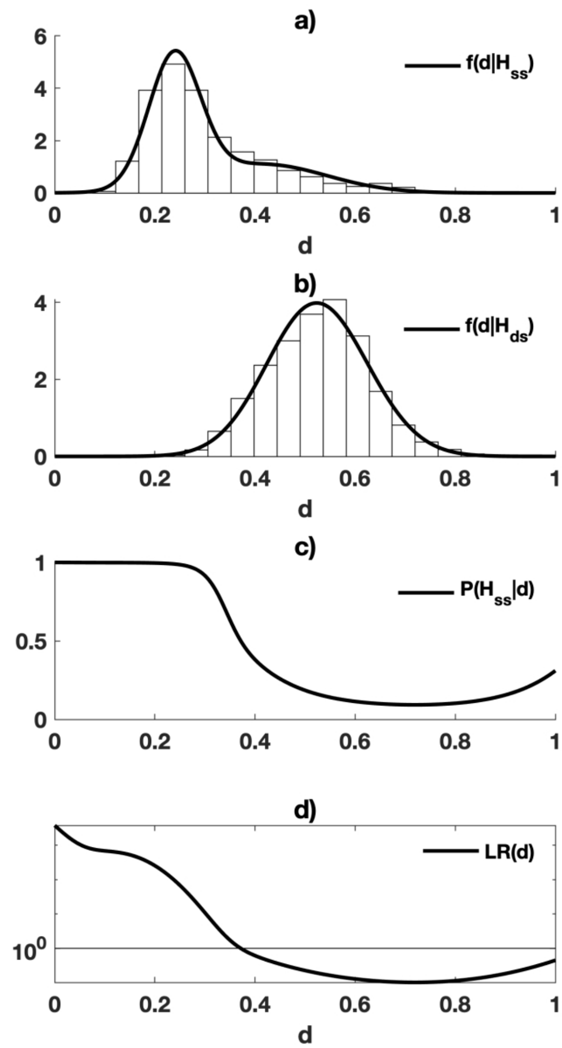

Figure 2: Results of the direct method with the scalar distance d on binarized feature, as functions of d, on the calibration set. a) Empirical distribution of d under Hss (1 594 pairs), and estimated density (mixture of 2 Gaussians). b) Empirical distribution of d under Hds (841 457pairs), and estimated density (mixture of 2 Gaussians). c) Posterior probability of Hss obtained using Bayes' formula with equal priors. d) Likelihood ratio.

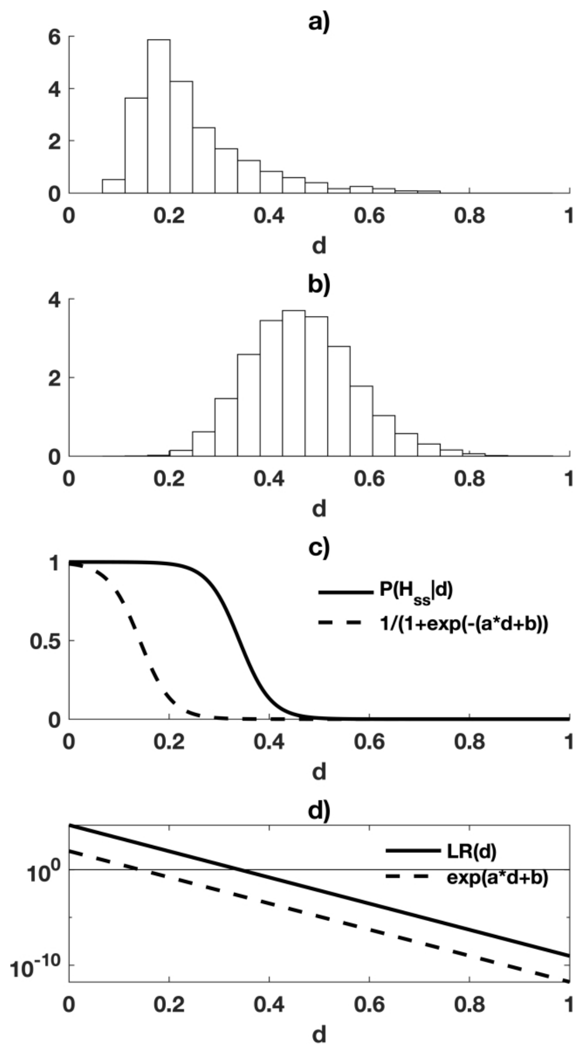

Figure 3: Results of the indirect method with the scalar distance d on continuous features, as functions of d, on the calibration set. a) Empirical distribution of d under Hss (1 594 pairs). b) Empirical distribution of d under Hds (841 457pairs). c) Logistic regression (dotted line), and deduced posterior probability of Hss with equal priors (continuous line). d) "Likelihood ratio" exp(aT d +b) corresponding to the logistic regression (dotted line), and likelihood ratio (continuous line).

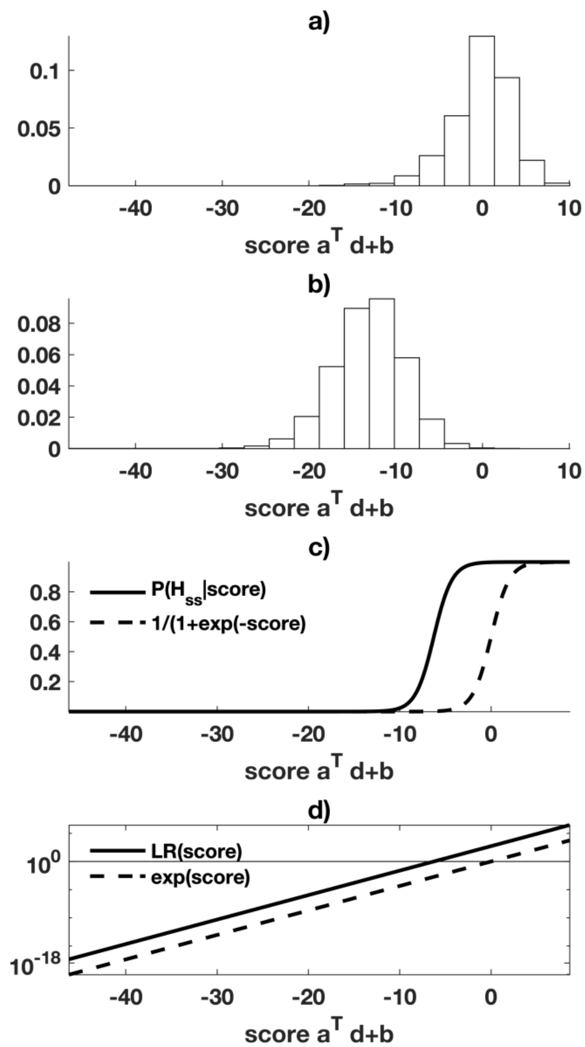

Figure 4: Results of the indirect method with the vectorial distance d on continuous features, as functions of the score aT d +b, on the calibration set. a) Empirical distribution of the score aT d + b under Hss (1 594 pairs). b) Empirical distribution of the score aT d + b under Hds (841 457pairs). c) Logistic regression (dotted line), and deduced posterior probability of Hss with equal priors (continuous line). d) "Likelihood ratio" exp(aT d +b) corresponding to the logistic regression (dotted line), and likelihood ratio (continuous line).

age / sex	7-17	18-64	65-94	total
man	18	182	18	218
woman	28	268	20	316
total	46	450	38	534

Normalization	Median relative standard deviation [IQI] in %
None	61.0 [32.1 ; 78.0]
In scalar scale	56.1 [31.9 ; 74.2]
In logarithmic scale	33.2 [17.9 ; 48.8]

	Calibration				Test			
Method	AUC	threshold	Sn	Sp	AUC	threshold	Sn	Sp
Direct	91.2	0.43	80.6	91.3	91.4	0.54	78.4	94.9
Indirect scal. d	91.2	0.62	80.6	91.3	91.5	0.75	78.4	94.9
Indirect vect. d	98.5	0.54	93.4	96.3	97.1	0.56	91.1	94.2

	Calibration				Test			
Method	AUC	threshold	Sn	Sp	AUC	threshold	Sn	Sp
Direct	92.1	0.44	81.1	92.5	93.0	0.50	81.3	94.6
Indirect scal. d	92.1	0.64	81.1	92.5	93.0	0.73	81.3	94.6
Indirect vect. d	98.9	0.58	94.4	97.0	97.8	0.57	91.3	95.1

			Calibration			Test		
Method	AUC-CV3	#feat.	AUC	Sn	Sp	AUC	Sn	Sp
Direct	94.4	267	94.5	80.9	94.7	94.4	84.6	92.8
Indirect scal. d	94.4	267	94.5	80.9	94.7	94.4	84.6	92.8
Indirect vect. d	96.4	440	97.6	89.5	97.5	97.0	91.3	94.3

			Calibration			Test		
Method	AUC-CV3	#feat.	AUC	Sn	Sp	AUC	Sn	Sp
Direct	93.1	535	93.1	81.3	93.3	93.5	82.3	95.5
Indirect scal. d	93.1	535	93.1	81.3	93.3	93.5	82.3	95.5
Indirect vect. d	97.0	500	98.3	91.6	97.3	97.7	91.7	94.9