A non-parametric effect size measure capturing changes in central tendency and shape of data distributions more flexibly than Cohen's d

Jörn Lötsch (✉ j.loetsch@em.uni-frankfurt.de)
https://orcid.org/0000-0002-5818-6958
Alfred Ultsch
Philipps-Universitat Marburg

Research article

Keywords: A non-parametric effect, size measure capturing changes in central tendency,shape of data distributions more flexibly than Cohen's d

Posted Date: January 17th, 2020

DOI: https://doi.org/10.21203/rs.2.21070/v1

A non-parametric effect size measure capturing changes in central tendency and shape of data distributions more flexibly than Cohen's d

Jörn Lötsch[1,2] and Alfred Ultsch[3]

[1] Institute of Clinical Pharmacology, Goethe - University, Theodor - Stern - Kai 7, 60590 Frankfurt am Main, Germany

[2] 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

[3] DataBionics Research Group, University of Marburg, Hans – Meerwein - Straße, 35032 Marburg, Germany

Correspondence to: Prof. Dr. Dr. Jörn Lötsch, Goethe - University, Theodor - Stern - Kai 7, 60590 Frankfurt am Main, Germany, e-Mail: j.loetsch@em.uni-frankfurt.de, Phone: +49-69-6301-4589, Fax: +49-69-6301-4354

Abstract

Calculating the magnitude of treatment effects or of differences between two groups is a common task in quantitative science. Standard effect size measures based on differences, such as the commonly used Cohen's, fail to capture the treatment-related effects on the data if the effects were not reflected by the central tendency. "Impact" is a novel nonparametric measure of effect size obtained as the sum of two separate components and includes (i) the change in the central tendency of the group-specific data, normalized to the overall variability, and (ii) the difference in the probability density of the group-specific data. Results obtained on artificial data and empirical biomedical data showed that impact outperforms Cohen's d by this additional component. It is shown that in a multivariate setting, while standard statistical analyses and Cohen's d are not able to identify effects that lead to changes in the form of data distribution, "Impact" correctly captures them. The proposed effect size measure shares the ability to observe such an effect with machine learning algorithms. It is numerically stable even for degenerate distributions consisting of singular values. Therefore, the proposed effect size measure is particularly well suited for data science and artificial intelligence-based knowledge discovery from (big) and heterogeneous data.

Introduction

Calculating the magnitude of treatment effects or group differences is a common task in quantitative biomedical science [1]. Effect sizes allow the quantification of the influence of independent variables (features) on dependent variables (e.g. treatment results) [2]. They are also useful to describe associations between features [3]. Several measures have been proposed and their use in biomedical research remains an active research topic [3]. Among the most commonly used measures are difference-based effect measures, among which Cohen's d [4] is frequently reported in the biomedical literature. Since these measures are based on the difference in the central tendency, they do not indicate an effect if this parameter does not change. However, this means that more general treatment-related effects on the data cannot be recorded if the effects are not reflected in the central tendency.

For example, due to an action that changes a known and neutrally evaluated object, the subjects can split into two opposing parties who either welcome or reject the change. Although the mean values of the evaluations before and after the action are similar, clearly visible group differences can be observed in different data distributions (Figure 1). While some of the limitations of Cohen's d's original proposal [5] have been addressed in modified effect size measures such as Hedges'g [6] or Glass's Δ [7], these measures continue to focus on the central tendency and will not capture the described effect. In contrast, a non-parametric comparison of the structure of the data might allow a more adequate quantification of an effect or a group difference.

We propose a novel effect size measure, called "impact", which captures effects a change the central tendency of the data as well as effects that change the shape of the data distribution. This may increase its usefulness as a generic effect size measure for the initial exploration of large and extensive data sets and provide a unifying description of effects on many different and

heterogeneously distributed variables. Since typical two-class problems such as "healthy" or "sick" occur in biomedical research, an effect size measure that compares two groups is largely applicable.

Methods

Impact effect size measure

Design criteria for the proposed measure of effect size were, first, that the measure should not be parametric. Secondly, the measure should be invariant to the scaling ($X' = c \cdot X$) and translation ($X' = X + c$) of the data X. Third, if the changes in the probability distributions are negligible, it should reflect only the change in the central tendency. Fourthly, changes in probability distributions should be recorded as consequences of treatment and, finally, the measure should be numerically stable, especially if the variances of data set X or its subgroups disappear.

Impact(X1, X2) defines an effect size based on the difference in central tendency between two groups or experimental conditions, *X1* and *X2*, as subgroups of a data set $X = \{X1 \cup X2\}$. "Impact" is the sum of two separate measures of the effects comprising (i) the change in the central tendency of the group-specific data, *CTdiff(X1,X2)* normalized to pooled variability and (ii) the difference in probability density of the group-specific data, called morphic difference *MorphDiff(X1,X2)*. Let *deltaM(X1,X2)* denote the difference of the medians in the two subgroups,

$$deltaM(X1, X2) = median(X2) - median(X1) \qquad \text{Equation 1}$$

and

$$pdf(\xi) \qquad \text{Equation 2}$$

denote the probability distribution of data ξ calculated empirically by a suitable estimation such as the Pareto density estimation (PDE) [8]. The central tendency difference is then defined as

$$CTdiff = \frac{deltaM(X1,X2)}{GMD(X1,X2)}$$ Equation 3

with *GMD(X1,X2)* denoting the expected value of absolute inner differences in *X*, which has been derived from Gini's mean difference [9]. This has been shown to be an appropriate measure of the variability of non-normal distributions [10] and is defined as

$$GMD(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|$$ Equation 4

with

$$GMD(X1,X2) = \begin{Bmatrix} sqrt(GMD(X1)^2 + GMD(X2)^2)\ if\ Var(X1) > 0\ and\ Var(X2) > 0 \\ GMD(X)\ if\ (Var(X1) = 0\ or\ Var(X2) = 0)\ and\ Var(X) > 0 \\ \varepsilon\ if\ Var(X) = 0\ with\ 0 < \varepsilon \ll 1 \end{Bmatrix}$$

Equation 5

The morphic difference describes the differences in the pdfs of *X1* and *X2*, including a directional factor related to the change in the centre of gravity c of the two pdfs

$$cgd(X1,X2) = \text{sign}(\text{cg}(\text{X1}) - cg(X2))$$ Equation 6

where cg(X) denotes the center of gravity of X.

$$MorphDiff(X1,X2) = \text{cgd}(\text{X1},\text{X2}) \cdot \int(|pdf(X2) - pdf(X1)|)$$ Equation 7

"Impact" is then the sum of the central tilt difference and the morphic difference, which by its definition fulfills the above mentioned design criteria:

$$Impact(X1,X2) = CTdiff(X1,X2) + MorphDiff(X1,X2)$$ Equation 8

Data sets

To evaluate the properties of the proposed effect size measure, to compare its results with those of Cohen's d and to assess its usefulness for two-class comparison problems, artificially generated and empirically collected biomedical data sets were used, which contained all two groups.

The **first data set** (Figure 1) was created with the property that both groups have the same means. Six subsets were created with n = 2000 points, unless otherwise specified $n_1 = |X_1| = 1000$ and $n_2 = |X_2| = 1000$. The first subset had the property that the means and variances were the same in both groups. The effect of an assumed treatment is that a standard unimodal normal distribution (N(0,1)) is changed to a bimodal distribution containing 50 % of the data in each mode. The second and third subgroups were essentially the same, but contained 80 % of the data in one mode and 20 % in the other mode. The data subset three X_3 was the data set two X_2, mirrored on the y axis: $X_3 = -X_2$. The fourth subset of data consisted of a standard normal distribution for X_1 and a Gaussian distribution with the same mean but with a standard deviation of four (N(0.4)). The fifth (X_5) and sixth (X_6) data subset consisted of a normal distribution in one group and a chi-square distribution in the other group, with the same mean as the Gaussian distribution, and with $X_6 = - X_5$.

The **second data set** (Figure 2) comprised subsets of two groups of n = 100 each, which were generated to compare the impact with Cohen's d in different scenarios. Data sets were created in which (i) both groups contained only a single value, a single value per group but different between groups, different values but identical in both groups, different values but only partially divided between groups, the values from the previous subset multiplied by 10, or a constant value in one group and normally distributed values with a different mean in the other group. In addition, the data set included data subsets with (ii) groups with the same mean but increasing variance in one group but not in the other, and (iii) groups with the same variance but increasing mean in one group but

not in the other. The data set was used for experiments comparing correlations of “Impact” with Cohen’s d (Figure 3).

A third data set (Figure 4) was created to examine the properties of Cohen's d compared to the “impact” measure for their behavior in a machine learning context (see experiments). It contained d = 20 variables (characteristics) with group sizes of n_1, n_2 = 1000. Ten variables were created as standard normal distributions (N(0,1)) using the same random number generator for all data subsets. The differences in these subsets should give values around zero in all effect measures. Five variables consist of a subset drawn from a standard normal distribution, the other subsets were drawn from a Gaussian distribution with mean = 3,...,7 and unit variance. For these variables, the effect measures should be significant and proportional to the difference in mean values. The last five characteristics consist of a subset drawn from a standard normal distribution, the other subsets were drawn from a bimodal distribution, so that the mean value of these subsets is zero, i.e. no change in the mean values between the two subsets, but with significant and increasing changes in their probability distribution. An appropriate sorting of the characteristics in this data set in descending order of absolute effect size should be 15,...,11 (i.e. differences in the central tendency), then the variables numbered 20,...,16 (differences in pdf) and then any order of variables 1 to 10 (no significant differences in the subsets).

A **fourth data set** (Figure 5) consisted of biomedical data obtained in a hematological context. It comprises eight different immunological markers associated with the diagnosis of lymphoma from a flow cytometric panel-based blood analysis. The measurements consist of a subset of n = 1,494 cells from healthy volunteers and a second subset of n = 1,302 cells from lymphoma patients. Cell surface molecules that provide targets for the immunophenotyping of the cells, i.e. CD3, CD4, CD8, CD11, CD19, CD103, CD200 and IgM, were used as measurement parameters.

Experiments

The evaluations of Impact's features and its usefulness for feature selection were primarily performed with the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) on an Intel Core i9® computer running Ubuntu Linux 18.04.3 64-bit).

Data set 1 (Figure 1) was used to show the differences between Cohen's d and Impact in brief. This data set consists mainly of subsets, with no differences in the central tendency, but a significant change in the shape of the distributions of the subsets. **Data set 2** (Figure 2) was used to check the numerical stability of the effect measures.

Data sets 2 and 3 were used to evaluate the effects identified by "Impact" and Cohen's d in a comparison scenario of machine learning and classical statistics, with the aim of ranking variables according to their suitability for mapping group differences reflected in a shift in the central tendency or in a change in the form of the data distribution. A ranking of 20 characteristics of **data set 3** (Figure 4) was made with regard to their differences between the groups. For each variable both “Impact” and Cohen's d were calculated. The variables relevant for group separation were then picked by applying an item categorization technique to the calculated effect sizes of each characteristic. This was implemented as a computed ABC analysis that met the basic requirements of feature selection by filtering techniques [12]. The method easily scales to very high-dimensional data sets, is computationally simple and fast and independent of the classification algorithm. The ABC analysis aims at dividing a data set into three disjoint subsets named "A", "B" and "C". The set "A" should contain the "important few", i.e. those elements that allow a maximum yield to be achieved with minimum effort [13, 14]. The ABC set B includes those elements where the increase in effort is proportional to the increase in yield. The set "C", on the other hand, contains the "trivial many", i.e. those elements with which the yield can only be achieved with a disproportionately large additional effort. In the calculated version of the ABC Analysis, the set limits are determined by mathematical

calculations performed with the R-package "ABC Analysis" (http://cran.r-project.org/package=ABCanalysis [15]).

Subsequently, the variables selected on the basis of a computed ABC analysis of the values of Impact or Cohen's d were used in classification tasks. First, classification and regression trees (CART) [16] were created with variables as vertices, conditions on these variables as edges and classes as leaves. In the present form, the Gini impurity was used to find optimal (local) dichotomous decisions. Additionally, a random forest classifier [17, 18] was trained. This generates sets of different, uncorrelated and often very simple decision trees with conditions on features as vertices and classes as leaves. The distribution of the features is random and the classifier refers to the majority vote for class membership. In the present analysis 500 decision trees were created containing *sqrt(d)* features as a standard of the R-library "caret" (https://cran.r-project.org/package=caret [19]), which was used together with the R-library "doParallel" (https://cran.r-project.org/package=doParallel [20]). The default settings were considered sufficient for the present demonstration purpose, and since elsewhere [21] it was found that there is no penalty for "too many" trees, the risk of over-adaptation was considered low.

The classification tasks were performed in cross-validation runs using 100-fold Monte-Carlo [22] resampling and data splitting into non-overlapping training (2/3 of the data) and test data (1/3 of the data). Classification performance was primarily evaluated as balanced accuracy [23, 24]. Other secondary measures of average classification performance were test sensitivity and specificity, and negative and positive predictive values calculated using standard equations [25, 26].

The variable selection and the subsequent classification experiments with data set 3 were performed twice, once with all 20 variables as candidates for selection based on a calculated ABC analysis and again with omission of variables 11 to 15 (third row of panels in Figure 4), i.e. all variables, since the central tendency varied considerably between the groups. To compare the data science-informatics

based approach with a classical statistical approach, i.e. performing an analysis of variance for repeated measurements (rm-ANOVA) with "measurements", i.e. the 20 variables, as the inter-theme factor and "group" as the inter-theme factor. The focus in this artificial data set was on the ability of the statistical procedure to detect significant group differences on the basis of all variables or the reduced set of variables, with the level of α set at 0.05. These calculations were performed using the SPSS software package (version 26 for Linux, IBM SPSS, IBM Corp, Armonk, NY, USA; https://www.ibm.com/analytics/spss-statistics-software).

Finally, experiments with the biomedical **data set 4** (Figure 5) were carried out analogously, although no predefined subsets as in data set 3 were excluded. In particular, the performance of CART and random forest classifiers when trained with variables selected on the basis of either Cohen's d or Impact was compared.

Implementation

The implementation of the impact effect size measurement in the R library "ImpactEffectsize" (https://cran.r-project.org/package=ImpactEffectsize) uses the PDE of our R package "AdaptGauss" (https://CRAN.R-project.org/package=AdaptGauss [27]). The effect size can be calculated with the *Impact(Data,Cls)* function. The input is expected to be a data vector, *Data*, and a bivalent integer vector of class information, *Cls*. The output consists of all values calculated when the effect size was estimated. The user can display the distributions of the data using either the PDE or a standard density estimation provided as an R-core function. The library uses additional functions provided in the R packages "RcppAlgos" (https://cran.r-project.org/package=RcppAlgos [28]), "caTools" (https://cran.r-project.org/package=caTools [29]), "matrixStats" (https://cran.r-project.org/package=matrixStats [30]) and "parallelDist" ([31] https://cran.r-project.org/package=parallelDist).

Results

Scaling and stability of effect sizes

The first data set served for introductory purposes and was intended to show that Impact recognizes an effect where Cohen's d leads to values close to zero (Figure 1). The results of the second data set (Figure 2) show that Impact (i) provides values where Cohen's d is not defined, such as scenarios where the data in one or both groups have a variance of zero, (ii) Impact scales proportionally to Cohen's d, and (iii) Impact is scale-invariant, i.e., it gives the same value when the values of a data set are multiplied by only one factor.

Data set 3 provides information on the relationship between Cohen's d and Impact. That is, var0011 to var0015 represent an increasing effect on the differences in the group mean values. In this case, Cohen's d and Impact are perfectly correlated (Pearson correlation [32] coefficient r = 0.998; Figure 3 middle panel). The variables var0001 to var0010 of data set 3 were obtained using a random number generator that produces standard normally distributed numbers. Therefore the effects should be insignificant, i.e. around zero. Cohen's d yields absolute values less than 0.1 for these variables, which for a small effect is below the proposed limit of d = 0.2 [33]. It is noteworthy that in this case too, Cohen's d and Impact are proportional (Figure 3 left panel). The variables var0016 to var0020 of data set 3 show no change in the central tendency, while their distribution undergoes significant changes. For all these characteristics, however, Cohen's d does not take Impact to zero, which means that Cohen's d, unlike Impact, does not capture these effects (Figure 3, right panel).

Recognition of group differences with effects on the central tendency or on the distribution form

The use of Cohen's d to point to group effects in the d = 20 artificially generated variables of data set 3 resulted in an ABC set "A" containing only variables in which the groups differed by a shift in the central tendency (middle line of the panels in Figure 4). In contrast, when Impact was used as the

basis for item categorization, the ABC set "A" contained additional variables in which the groups differed in the form of the pdf but had the same mean value (variables var0018, var0019, and var0020 in the bottom line of the panels in Figure 4). The training of both CART and random forests allowed a correct classification at equal performance with a median accuracy of 100 % (Table 1).

The picture changed when all variables in which the groups differed by a shift in the central tendency were omitted. Cohen's then provided a rather random set of characteristics for the ABC set "A" (variables var0001, var0004, var0008, var0009) from the variables that were designed to show no group differences but a random variation between groups. In contrast, in Impact-based selection, all variables with the same means but different forms of the distributions were assigned as members of ABC set "A". As expected, training CART and random forests with the respective sets of variables led to a complete failure of classification for the variables selected on the basis of Cohen's d (median classification accuracy 50 % similar to guessing), whereas correct classification was almost completely possible (median classification accuracy 99.8 %) with the variables having similar means but different forms between the groups selected on the basis of the measure of impact effect size (Table 1).

For comparison, the classical statistical approach could, as expected, detect a group difference when the full set of variables was available (Table 2). In contrast to the artificial intelligence-based approaches (random forests, CART), which had no problems separating the groups due to the form of data distribution, the analysis of variance could not detect the group difference if only variables with the same central tendency were available, i.e. those with a group shift in the mean were omitted (Table 2). The same was achieved by using several t-tests [34], which only found significant group differences for the variables with median shift but nut with the same mean but shape differences (details not shown).

The use of Cohen's d to identify most of the group discriminating variables among the d = 10 immunological markers of data set 4, related to lymphomas (Figure 5), resulted in an ABC set "A"

containing CD8 and CD103. The classification accuracies obtained with CART or random forest based classifiers were between 69 and 66 % (Table 3). When using Impact to identify the group discriminating variables, a third marker CD4 was selected in addition to the two markers, also based on the value of Cohen's d. This third marker increased the balanced classification accuracy by up to 10% from the marker set selected on the basis of Cohen's d to 71-75% (Table 3).

Discussion

The processing of large amounts of data in the life sciences often implies a large number, e.g. thousands of variables. In such an environment, visual inspection or manual analysis of all variables for their suitability to quantify treatment effects or group differences is not feasible. This increases the need for robust calculations that are defined even in extreme cases, e.g. when the variances of subsets deteriorate to zero. However, the popular Cohen's d measure for effect size is undefined in this case and an algorithmic implementation of Cohen's d would yield unpredictable values. This is unacceptable if a measure of effect size is to be used for feature selection, that is, the selection of a few relevant features from a large corpus of mostly irrelevant candidate features. Furthermore, in order not to miss any important effects, as shown in this report, more than a comparison of the mean values of the untreated and treated subgroups to which the Cohen'd measure is limited is required.

An effect size measure is presented that captures changes in the central tendencies as well as changes in the forms of data distribution. This is missed with the classical effect size measure Cohen's d. Demonstrations on artificial data and empirical biomedical data from real measurements have shown that this additional property allows Cohen's d, as a typical classical effect size measure, to outperform Cohen's d in the assessment of group differences when the global form of data distribution is more relevant than the central tendency.

Impact regards the difference in central tendency as one of its components (*CTdiff*). The morphic difference (*MorphDiff*) as the second component takes into account changes in the shape of the distribution. The scaling of these two components was chosen so that the absolute values of the *CTdiff* are unbound, while the *MorphDiff* is between -1 and 1, which means that for large effect sizes the Cohen's d style of effect size measure (*CTdiff*) dominates in the calculation of the Impact. If the effects are small, i.e. the (normalized) central tendency is in a range from -1 to 1, the morphic difference in the effect size measure becomes more important.

Within these limits, it has been shown that Cohen's d is not able to detect effects (Figure 3 right), while Impact is proportional to the amount by which the distribution has changed. Typical implementations of Cohen's d use the pooled variance for two subgroups [35].. However, when this pooled variance is used, it makes the calculation of Cohen's d dependent on the relative sizes of the two subgroups. The variance of the larger subgroup will dominate the pooled variance and is therefore crucial for unifying the central tendency. Impact, however, is completely independent of the sizes of the two subgroups (treated versus untreated). This is particularly advantageous if the treated group of patients is small.

Effect size measures, if not used for feature selection, are a basis for meta-analyses. If not reported in the original publication, they are usually estimated from the reported measures of central tendency and variance. This can be achieved in a similar way for the proposed measure, i.e. the impact can be estimated from the parametric information on the variance. However, this may miss the difference in the shape of the distribution. Ideally, the original data are available that allow the form of the distribution to be estimated, including possible multimodality that is not covered by the standard statistical measures usually reported in scientific papers. A very clear example (Table 2) showed that in some cases typical statistical analyses such as ANOVA with repeated measurements are not able to detect differences in groups where a machine-learned classifier has no difficulty in

doing so. Therefore, the proposed effect size measure is directed more towards a data science and machine learning context than statistical data analysis.

Conclusions

An effect size measure is proposed that, first, is robust to any type of data distribution and, second, believes that a treatment can have complex effects on the measured characteristics, either on the central tendency or on the shape of the distribution, or on both. However, the established characteristics of the Cohen's d remain intact when using the newly defined effect size measure "impact". Based on artificial and real empirical data, it was shown that a purely algorithmic procedure for feature selection can be used to find the most relevant features of data sets with this new effect measure. Furthermore, the present experiments clearly show an advantage of the machine-learned algorithms and the prosed effect size measure over classical statistical analyses and the standard Cohen's d-effect size measure for capturing complex treatment effects or group differences. Impact has been shown to outperform Cohen's d and other statistical tools for data analysis.

Declarations

Ethics approval and consent to participate: n/a

Consent for publication: n/a

Availability of data and materials: The “ImpactEffectsize” R package is freely available at https://cran.r-project.org/package=ImpactEffectsize.

Competing interests: The authors have declared that no further conflicts of interest exist.

Funding: This work has been funded by the Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL). The funders had no role in the decision to publish or in the preparation of the manuscript.

Authors' contributions: AU – Conceptualization of the project, mathematical implementation, writing of the manuscript, JL – Conceptualization of the project, programming, writing of the manuscript, data analyses and creation of the figures

Acknowledgements: A preliminary version of the proposed effect size measure has been communicated in an oral presentation at the Fifth European Conference on Data Analysis (ECDA2019), March 18 – 20, 2019, Bayreuth, Germany.

References

1. Cohen J: **Statistical Power Analysis for the Behavioral Sciences**. New York: Routledge; 1988.

2. Monsarrat P, Vergnes J-N: **The intriguing evolution of effect sizes in biomedical research over time: smaller but more often statistically significant**. *GigaScience* 2017, **7**(1):1-10.

3. Rosenthal JA: **Qualitative Descriptors of Strength of Association and Effect Size**. *Journal of Social Service Research* 1996, **21**(4):37-59.

4. Cohen J: **A coefficient of agreement for nominal scales**. *Educ Psychol Meas* 1960, **20**.

5. Grice JW, Barrett PT: **A note on Cohen's overlapping proportions of normal distributions**. *Psychol Rep* 2014, **115**(3):741-747.

6. Hedges LV: **Distribution Theory for Glass's Estimator of Effect size and Related Estimators**. *Journal of Educational Statistics* 1981, **6**(2):107-128.

7. Glass G, McGaw B, Smith M: **Meta-analysis in social research 1981 Beverly Hills**. In.: CA Sage.

8. Ultsch A: **Pareto Density Estimation: A Density Estimation for Knowledge Discovery.** In: *Innovations in Classification, Data Science, and Information Systems - Proceedings 27th Annual Conference of the German Classification Society (GfKL): 2003; Berlin*. Springer.

9. Gini C: **Measurement of Inequality of Incomes**. *The Economic Journal* 1921, **31**(121):124-126.

10. Shlomo Y: **Gini's Mean difference: a superior measure of variability for non-normal distributions**. *Metron - International Journal of Statistics* 2003, **0**(2):285-316.

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

12. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.

13. Pareto V: **Manuale di economia politica, Milan: Società editrice libraria, revised and translated into French as Manuel d'économie politique**. 1909.

14. Juran JM: **The non-Pareto principle; Mea culpa.** *Quality Progress* 1975, **8**(5):8-9.

15. Ultsch A, Lötsch J: **Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data**. *PLoS One* 2015, **10**(6):e0129767.

16. Breimann L, Friedman JH, Olshen RA, Stone CJ: **Classification and Regression Trees**. Boca Raton: Chapman and Hall; 1993.

17. Ho TK: **Random decision forests**. In: *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*. 844681: IEEE Computer Society 1995: 278.

18. Breiman L: **Random Forests**. *Mach Learn* 2001, **45**(1):5-32.

19. Kuhn M: **caret: Classification and Regression Training**. In.; 2018.

20. Weston S: **doParallel: Foreach Parallel Adaptor for the 'parallel' Package**. In.: Corporation, Microsoft; 2019.

21. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP: **Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling**. *J Chem Inf Comput Sci* 2003, **43**(6):1947-1958.

22. Good PI: **Resampling methods : a practical guide to data analysis**. Boston: Birkhäuser; 2006.

23. Brodersen KH, Ong CS, Stephan KE, Buhmann JM: **The Balanced Accuracy and Its Posterior Distribution**. In: *Pattern Recognition (ICPR), 2010 20th International Conference on: 23-26 Aug. 2010 2010*. 3121-3124.

24. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH: **A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction**. *Genet Epidemiol* 2007, **31**(4):306-315.

25. Altman DG, Bland JM: **Diagnostic tests. 1: Sensitivity and specificity**. *BMJ* 1994, **308**(6943):1552.

26. Altman DG, Bland JM: **Diagnostic tests 2: Predictive values**. *BMJ* 1994, **309**(6947):102.

27. Ultsch A, Thrun MC, Hansen-Goos O, Lötsch J: **Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss)**. *Int J Mol Sci* 2015, **16**(10):25897-25911.

28. Wood J: **RcppAlgos: High Performance Tools for Combinatorics and Computational Mathematics**. In.; 2019.

29. Tuszynski J: **caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.** In.; 2019.

30. Bengtsson H: **matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)**. In.; 2019.

31. Eckert A: **parallelDist: Parallel Distance Matrix Computation using Multiple Threads**. In.; 2018.

32. Pearson K: **LIII. On lines and planes of closest fit to systems of points in space**. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901, **2**(11):559-572.

33. Cohen J: **Statistical power analysis for the behavioral sciences**. New York: Academic press; 1998.

34. Student: **The Probable Error of a Mean**. *Biometrika* 1908, **6**(1).

35. Hartung J, Knapp G, Sinha BK: **Statistical meta-analysis with applications**: Wiley, New York; 2008.

Table 1: Performance measures of classifiers applied to sets of variables selected based on the magnitude of either Cohen's or the Impact effect size measure. Two different machine-learned methods (classification and regression trees (CART) and random forests (RF) were applied on artificially created data comprising two groups with sizes of n = 1000 and d = 20 variables (var0001 - var0020, Figure 2). Of these variables, in var0001 – var0010 the means and variances were randomly jittered between the two groups (upper two lines of panels in Figure 2), in variables va0011 – var0015 the means differed substantially between groups (third line of panels in Figure 2), and in var0016 – var0020 the groups had the same mean but one group the data was spilt into two distinct modes whereas in the other group the data varied around the mean (bottom line of panels in Figure 2). 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).

	All parameters [1,…,20] (see Figure 4)				**Reduced set of parameters [1,…,10,16,…,20] (see Figure 4)**			
	Cohen's d		Impact		Cohen's d		Impact	
Selected features (Var00...)	11, 12, 13, 14, 15		11, 12, 13, 14, 15, 18, 19, 20		1, 4, 8, 9		16, 17, 18, 19 20	
	CART	RF	CART	RF	CART	RF	CART	RF
Sensitivity, recall	100 (99.1 - 100)	100 (100 - 100)	100 (99.1 - 100)	100 (100 - 100)	58.6 (8.3 - 100)	50.3 (44.7 - 55.6)	100 (99.7 - 100)	100 (100 - 100)
Specificity	100 (100 - 100)	100 (100 - 100)	100 (100 - 100)	100 (100 - 100)	41.7 (0 - 89.8)	50.3 (45.1 - 54.7)	99.7 (99.1 - 100)	100 (100 - 100)
Pos. pred. value, precision	100 (100 - 100)	100 (100 - 100)	100 (100 - 100)	100 (100 - 100)	50 (44.6 - 53)	50.2 (47 - 53.4)	99.7 (99.1 - 100)	100 (100 - 100)
Negative predictive value	100 (99.1 - 100)	100 (100 - 100)	100 (99.1 - 100)	100 (100 - 100)	50.1 (47.3 - 55.8)	50.2 (47.3 - 53.2)	100 (99.7 - 100)	100 (100 - 100)
F1	100 (99.5 - 100)	100 (100 - 100)	100 (99.5 - 100)	100 (100 - 100)	53.9 (13.9 - 66.7)	50.2 (45.8 - 53.9)	99.8 (99.5 - 100)	100 (100 - 100)
Balanced Accuracy	100 (99.5 - 100)	100 (100 - 100)	100 (99.5 - 100)	100 (100 - 100)	50 (47.3 - 53.1)	50.2 (47.2 - 53.3)	99.8 (99.5 - 100)	100 (100 - 100)
AUC ROC	100 (99.5 - 100)	100 (100 - 100)	100 (99.5 - 100)	100 (100 - 100)	50 (47.2 - 53.4)	50 (47.2 - 53.4)	99.8 (99.5 - 100)	100 (100 - 100)

Table 2: Results of an analysis of variance for repeated measures (rm-ANOVA) applied onto the artificially created data set comprising two groups with sizes of n = 1000 and d = 20 variables (var0001 - var0020, Figure 2). Of these variables, in var0001 – var0010 the means and variances were randomly jittered between the two groups (upper two lines of panels in Figure 2), in variables va0011 – var0015 the means differed substantially between groups (third line of panels in Figure 2), and in var0016 – var0020 the groups had the same mean but one group the data was spilt into two distinct modes whereas in the other group the data varied around the mean (bottom line of panels in Figure 2).

rm-ANOVA effects	**All parameters [1,…,20] (see Figure 4)**			**Reduced set of parameters [1,…,10,16,…,20] (see Figure 4)**		
	df	F	p	df	F	p
Measure	19,3796	735.878	$< 6.65 \cdot 10^{-244}$	14,2792	0.139	1
Measure * Class	19,3796	736.578	$< 6.65 \cdot 10^{-244}$	14,2792	0.228	0.999
Class	1,1998	917.764	$3.23 \cdot 10^{-166}$	1,1998	0.001	0.978

Table 3: Performance measures of classifiers applied to variables selected based on the magnitude of either Cohen's or the Impact effect size measure. Two different machine-learned methods (classification and regression trees (CART) and random forests (RF) were applied on biomedical data of a hematological context comprising a flow cytometry-based lymphoma makers CD8, CD4, CD3, CD200, CD11 CD20, IgM , CD19, and CD103 (marker names truncated for non-disclosure reasons) from healthy subjects and patients (Figure 5). 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).

	Based on Cohen's d		**Based on Impact**	
Selected features	CD8, CD103		CD8, CD4, CD103	
	CART	RF	CART	RF
Sensitivity, recall	69.1 (51.4 - 82.6)	67.3 (57.6 - 73.1)	71.4 (61 - 80.1)	75.7 (67.8 - 81.9)
Specificity	70.5 (51.5 - 81.3)	64 (56.7 - 74)	69.7 (61.6 - 79.7)	74.2 (67.7 - 81.3)
Pos. pred. value, precision	69.9 (64.1 - 75)	66.1 (61.4 - 68.9)	71.6 (68.3 - 75.4)	75.5 (72.4 - 78)
Negative predictive value	69.4 (64.5 - 74.3)	65.6 (61.1 - 69.7)	71.4 (67 - 74.3)	75.3 (72.3 - 77.3)
F1	69.3 (58.9 - 74)	66.6 (60.1 - 70.5)	72 (64.9 - 76.3)	75.5 (70.7 - 79)
Balanced Accuracy	69.1 (65.6 - 71.3)	65.6 (62.6 - 67.9)	71.2 (68.5 - 73.2)	75.1 (72.8 - 76.8)
AUC ROC	70.7 (65.6 - 73.8)	70.8 (67.8 - 73.5)	71.1 (68.8 - 73.3)	82.7 (80.4 - 84.2)

Figure 1: Example cases where the two samples (red and blue) display no differences in mean but possess clearly different data distributions (data set 1). The plots show the probability distribution function (pdf) of the data (ordinate) along the data's range (abscissa). The perpendicular lines indicate the means for both data subsets. The colors correspond to the sample pdf colors. Please note that the means are numerically identical but have been optically separated by one pixel. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

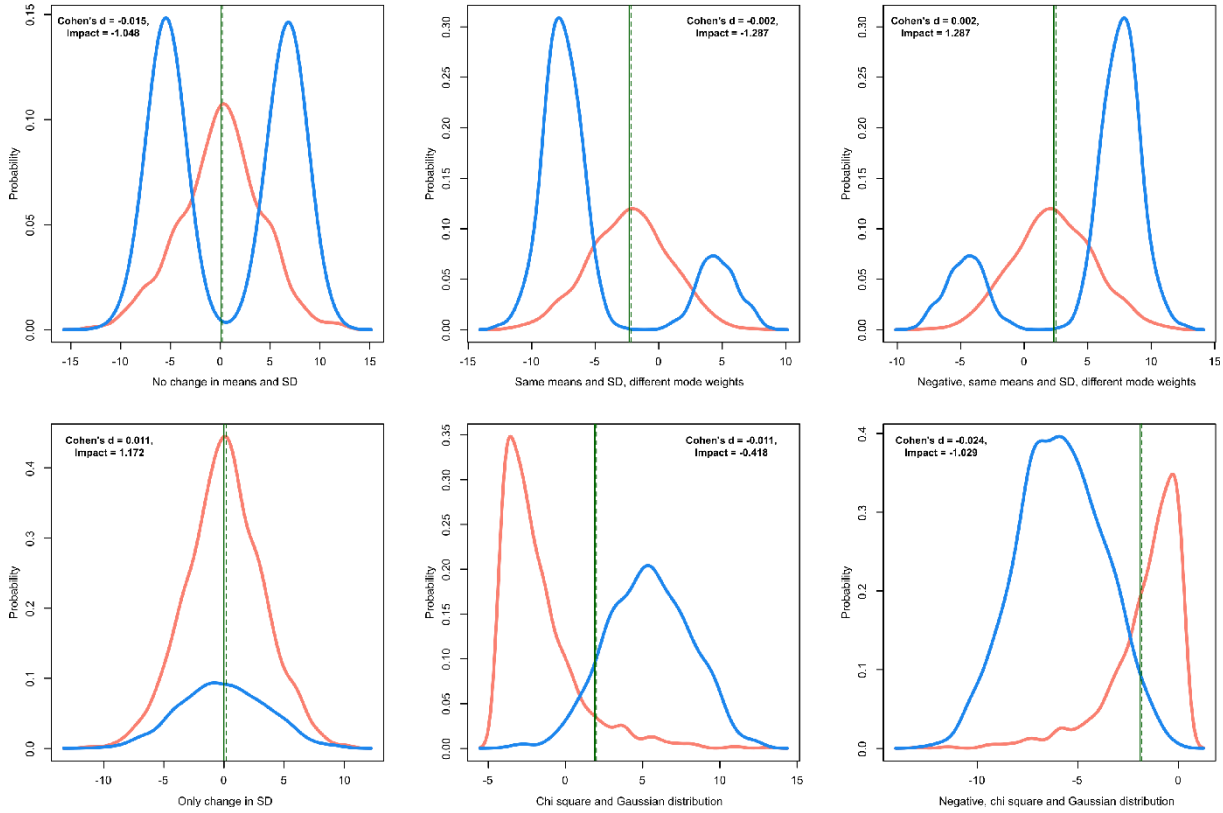


Figure 2: Effect sizes expressed as Cohen's d respectively Impact calculated for artificially created data sets (data set 2) comprising subsets of each two groups (red and blue) with identical sizes of n = 100. **Left panel:** Data subsets in which (i) both groups contained only one single value, one single value per group but different between groups, various different values but identical in both groups, various different values but only partly shared between groups, the values from the previous subset multiplied with 10, or a constant value in one group and normally distributed values with a different mean in the other group. **Middle panel:** Several data subsets with groups with the same mean but increasing variance in one but not the other group. **Right panel:** Data subsets with groups with the same variance but increasing mean in one but not the other group. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

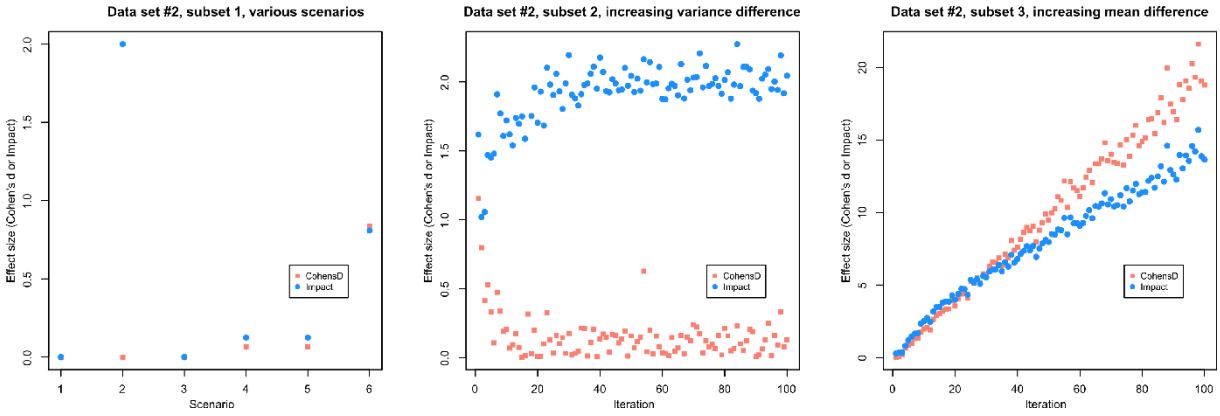


Figure 3: Relations between the “Impact” effect size measure and Cohen’s d in the different data scenarios of data set 3 (Figure 4). Left panel: When data of both groups are randomly generated normally distributed numbers with small between-group differences in mean and variance (var0001 to var0010 of data set 3), the effects are small although still correlated. Middle panel: When differences in the group means increase (var0011 to var0015), Cohen’s d and Impact are perfectly correlated. Right panel: With no change in the central tendency difference but group difference merely in the distribution of data. Cohen’s d takes a value of zero in these cases, Impact measures the effect. The dotted lines indicate linear regressions. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]).

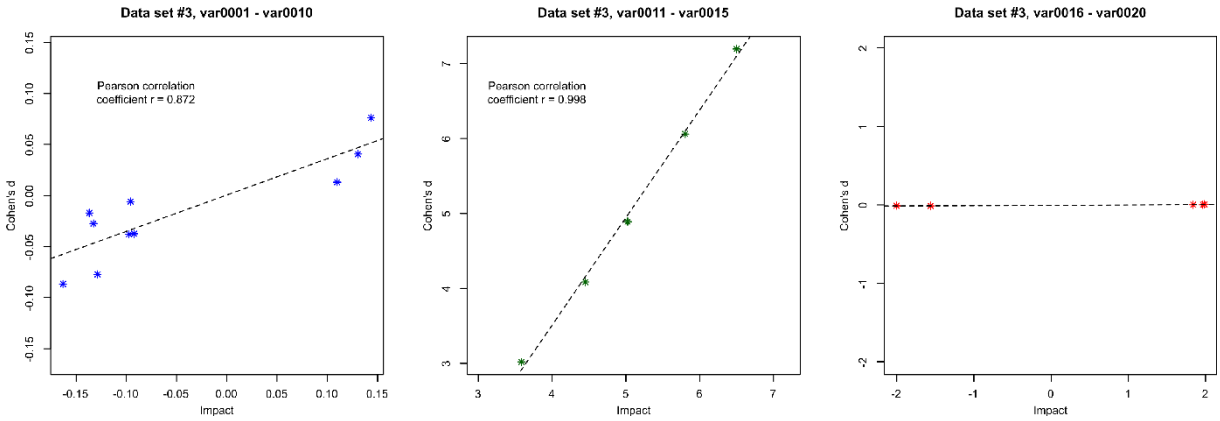


Figure 4: Feature selection using either Cohen's d or Impact: **A:** Artificially created data comprising two groups (red and blue) with sizes of n = 1000 and d = 20 variables (data set 3). Of these variables, in var0001 – var0010 the means and variances were randomly jittered between the two groups (upper two lines of panels), in variables va0011 – var0015 the means differed substantially between groups (third line of panels), and in var0016 – var0020 the groups had the same mean but one group the data was spilt into two distinct modes whereas in the other group the data varied around the mean (bottom line of panels). **B:** Results of feature selection based on calculation of the effect size followed by computed ABC analysis. The bar graphs show the effect size in descending order. The relevant features, i.e., those in ABC sets A and B, are shown in blue color. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

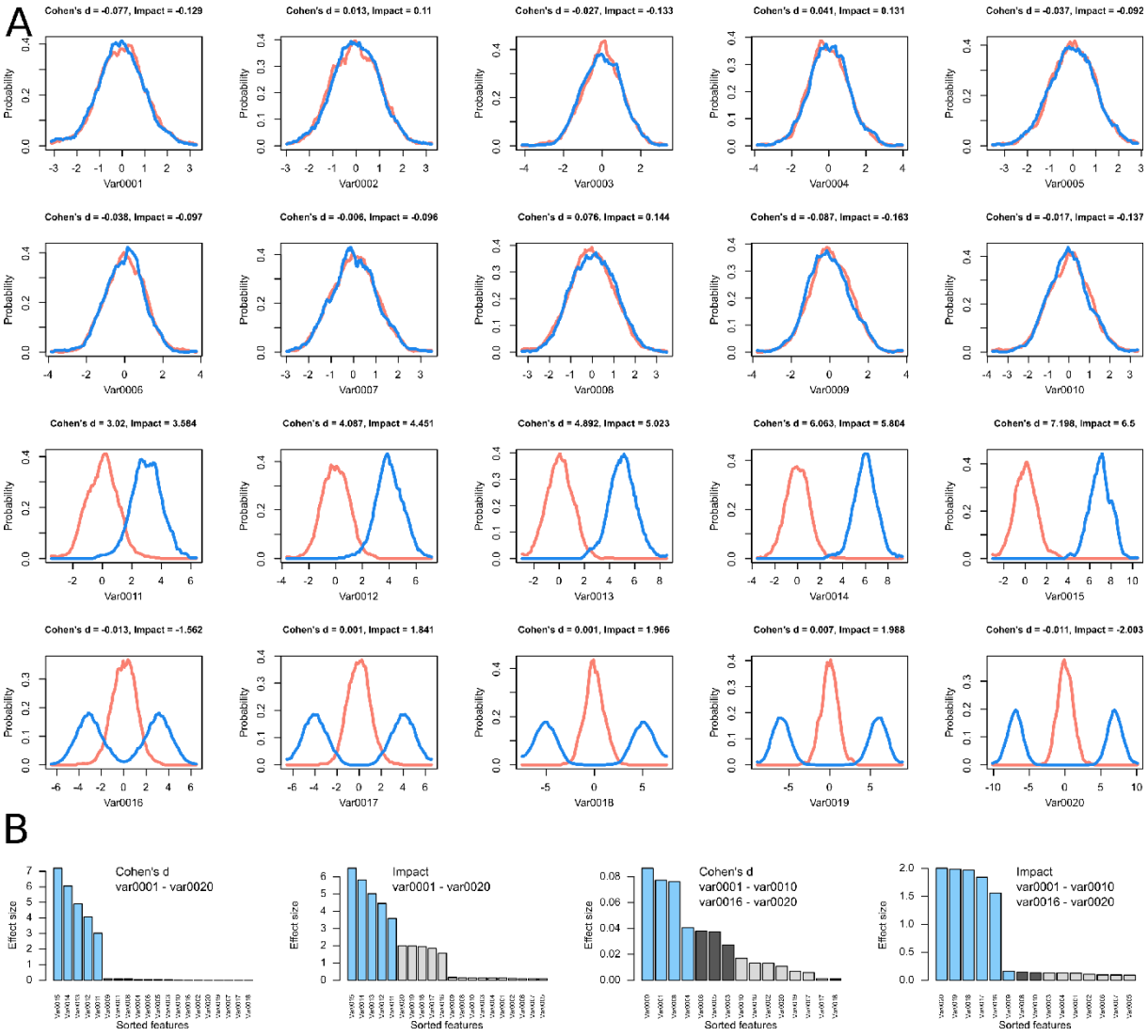
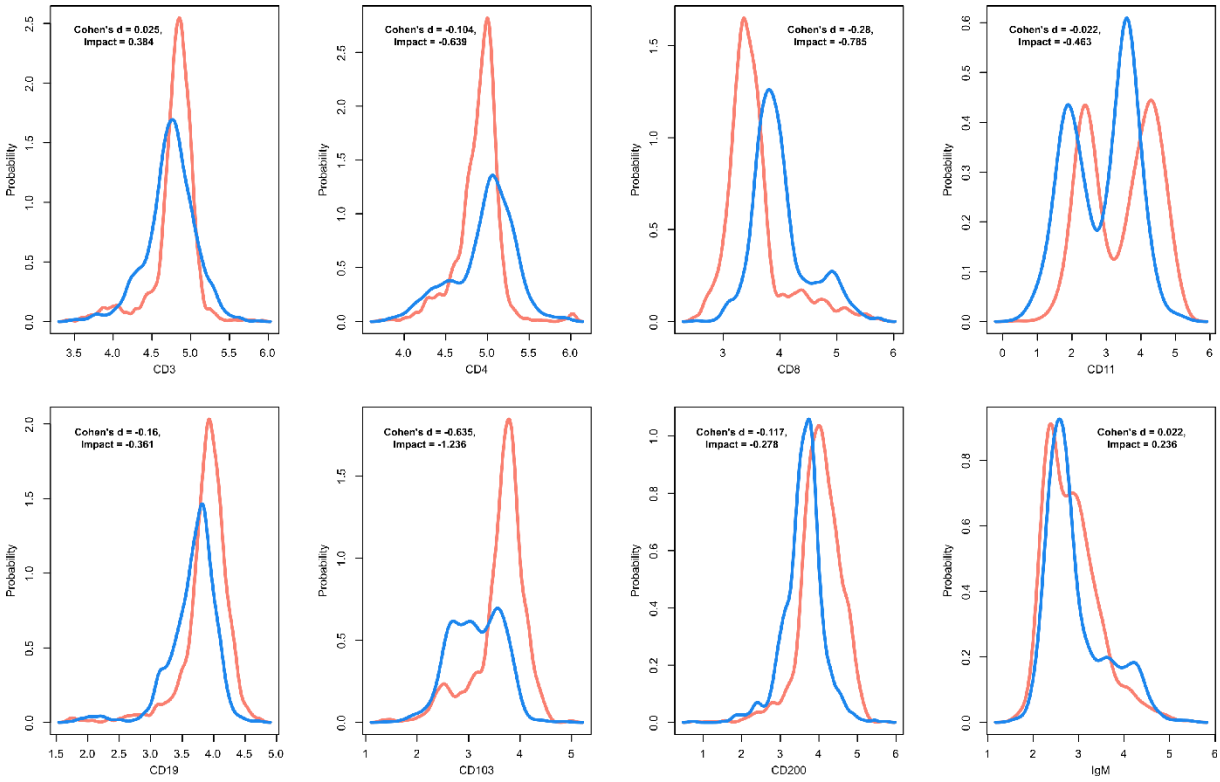


Figure 5: Biomedical data of a hematological context comprising a flow cytometry-based lymphoma makers CD3, CD4, CD8, CD11, CD19, CD103, CD200 and IgM comprising of one subset of n = 1,494 cells from healthy subjects (red) and a second set of n = 1,302 cells from lymphoma patients (blue) (data set 4). The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).



Figures

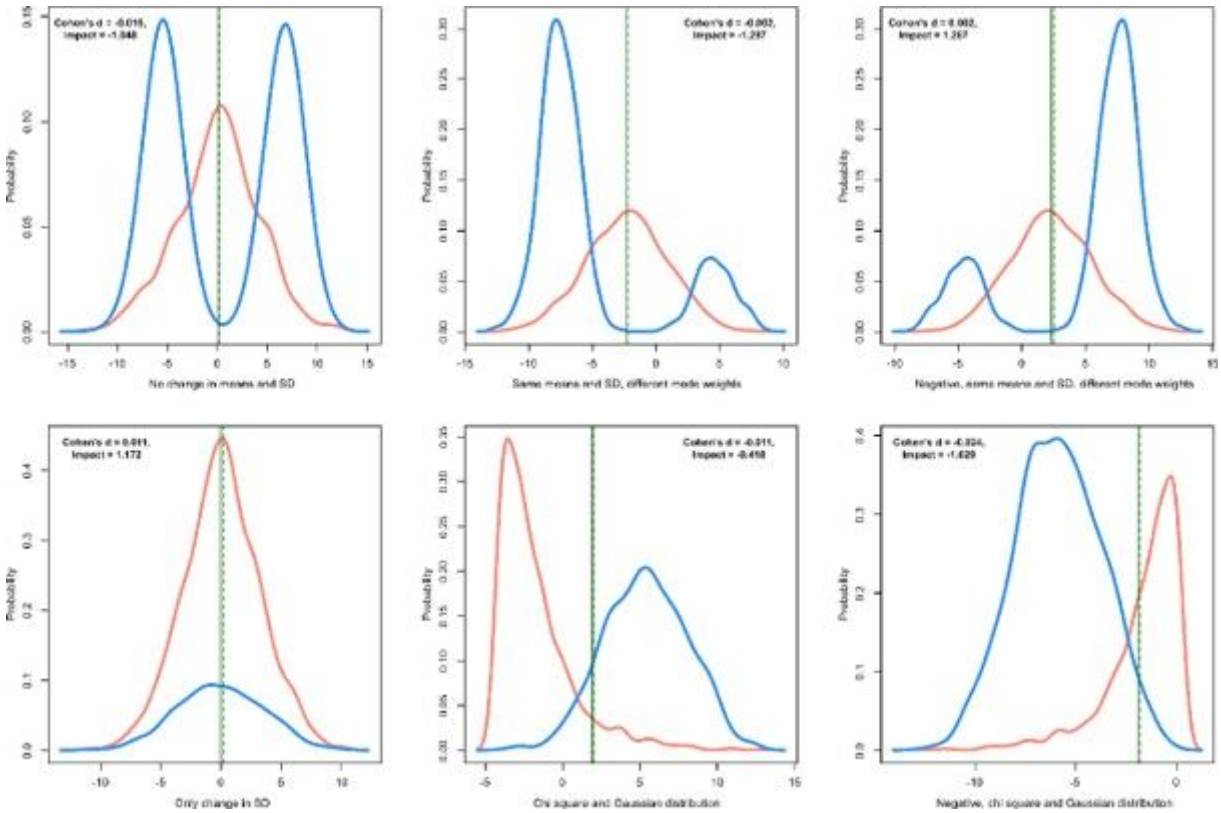


Figure 1

Example cases where the two samples (red and blue) display no differences in mean but possess clearly different data distributions (data set 1). The plots show the probability distribution function (pdf) of the data (ordinate) along the data's range (abscissa). The perpendicular lines indicate the means for both data subsets. The colors correspond to the sample pdf colors. Please note that the means are numerically identical but have been optically separated by one pixel. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

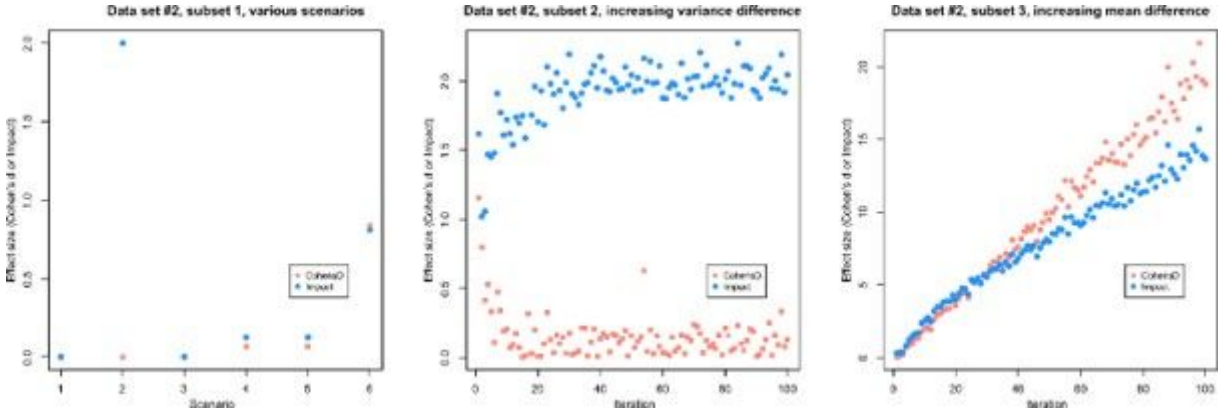


Figure 2

Effect sizes expressed as Cohen's d respectively Impact calculated for artificially created data sets (data set 2) comprising subsets of each two groups (red and blue) with identical sizes of n = 100. Left panel: Data subsets in which (i) both groups contained only one single value, one single value per group but different between groups, various different values but identical in both groups, various different values but only partly shared between groups, the values from the previous subset multiplied with 10, or a constant value in one group and normally distributed values with a different mean in the other group. Middle panel: Several data subsets with groups with the same mean but increasing variance in one but not the other group. Right panel: Data subsets with groups with the same variance but increasing mean in one but not the other group. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

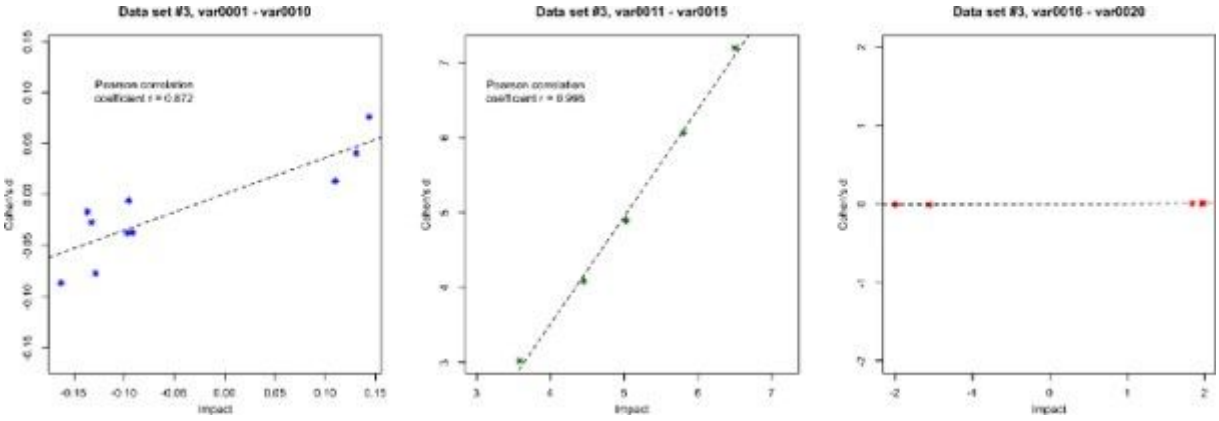


Figure 3

Relations between the "Impact" effect size measure and Cohen's d in the different data scenarios of data set 3 (Figure 4). Left panel: When data of both groups are randomly generated normally distributed numbers with small between-group differences in mean and variance (var0001 to var0010 of data set 3), the effects are small although still correlated. Middle panel: When differences in the group means increase (var0011 to var0015), Cohen's d and Impact are perfectly correlated. Right panel: With no change in the central tendency difference but group difference merely in the distribution of data. Cohen's d takes a value of zero in these cases, Impact measures the effect. The dotted lines indicate linear regressions. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]).

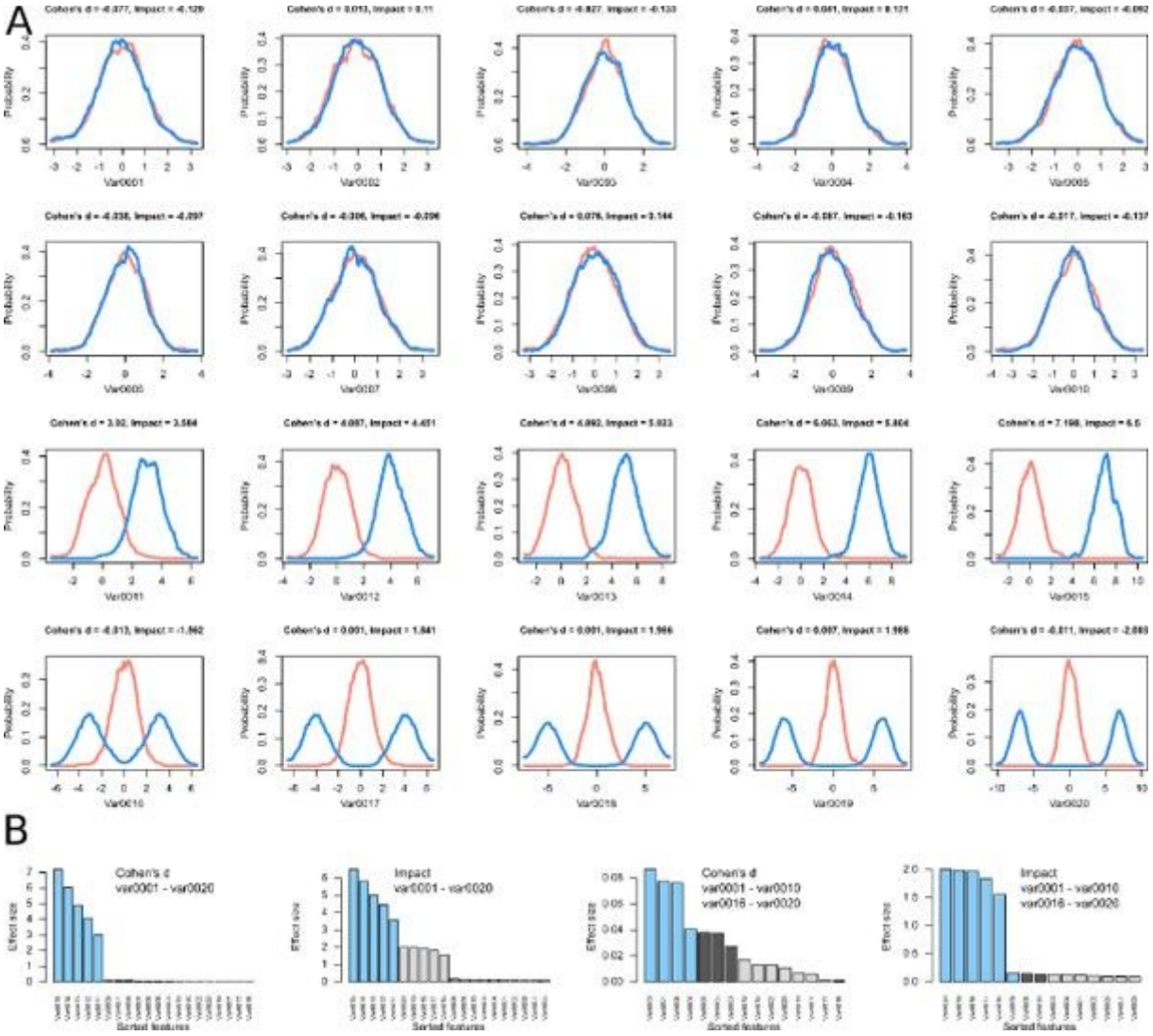


Figure 4

Feature selection using either Cohen's d or Impact: A: Artificially created data comprising two groups (red and blue) with sizes of n = 1000 and d = 20 variables (data set 3). Of these variables, in var0001 – var0010 the means and variances were randomly jittered between the two groups (upper two lines of panels), in variables va0011 – var0015 the means differed substantially between groups (third line of panels), and in var0016 – var0020 the groups had the same mean but one group the data was spilt into two distinct modes whereas in the other group the data varied around the mean (bottom line of panels). B: Results of feature selection based on calculation of the effect size followed by computed ABC analysis. The bar graphs show the effect size in descending order. The relevant features, i.e., those in ABC sets A and B, are shown in blue color. The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package "ImpactEffectsize" (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).

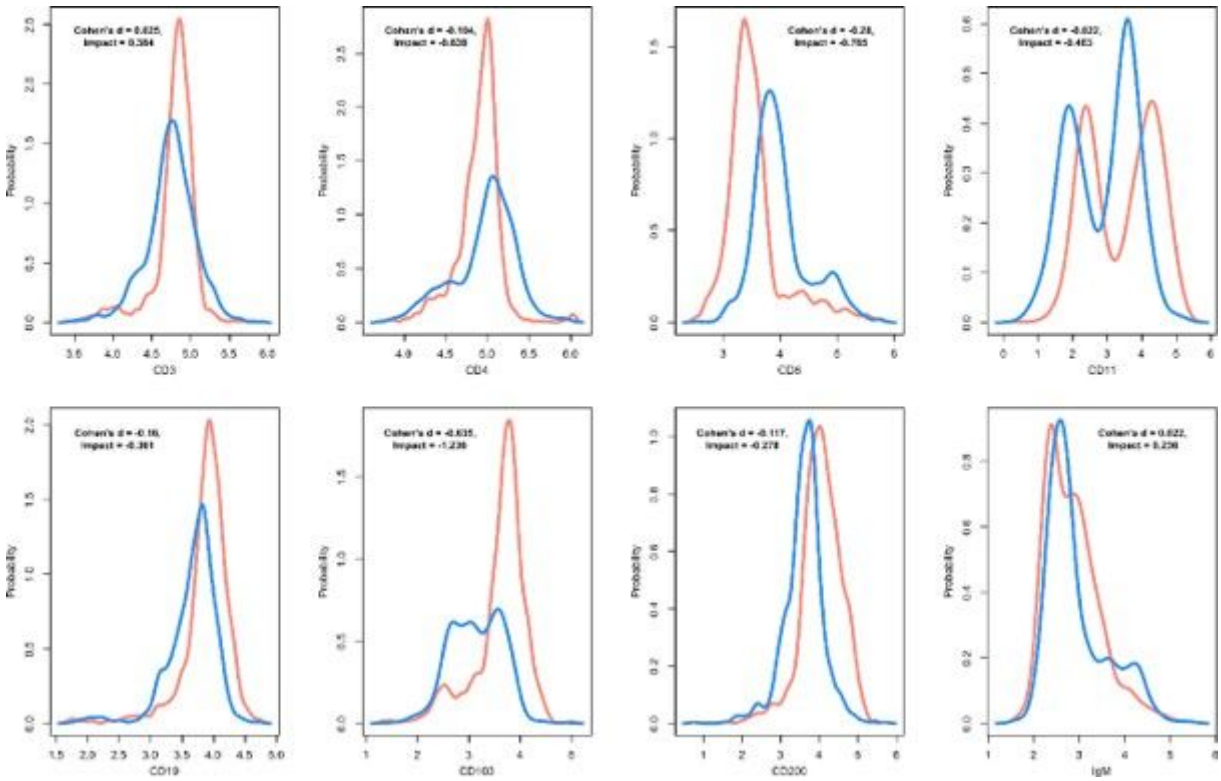


Figure 5

Biomedical data of a hematological context comprising a flow cytometry-based lymphoma makers CD3, CD4, CD8, CD11, CD19, CD103, CD200 and IgM comprising of one subset of n = 1,494 cells from healthy subjects (red) and a second set of n = 1,302 cells from lymphoma patients (blue) (data set 4). The figure has been created using the R software package (version 3.6.1 for Linux; http://CRAN.R-project.org/ [11]) using the R package “ImpactEffectsize” (https://www.kgu.de/zpharm/klin/research/ImpactEffectsize_0.1.0.tar.gz; CRAN upload pending).