modelBuildR: an R package for model building and feature selection with erroneous classifications

Maximilian Knoll, Jennifer Furkel, Juergen Debus and Amir Abdollahi

Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Deutschland
National Center for Tumor Disease (NCT), UKHD and German Cancer Research Center (DKFZ), Heidelberg, Germany
German Cancer Consortium (DKTK), Core Center Heidelberg, DKFZ, Heidelberg, Germany

Corresponding author
Maximilian Knoll, m.knoll@dkfz.de

ABSTRACT

Background. Model building is a crucial part of omics based biomedical research to transfer classifications and obtain insights into underlying mechanisms. Feature selection is often based on minimizing error between model predictions and given classification (maximizing accuracy). Human ratings/classifications, however, might be error prone, with discordance rates between experts of 5–15%. We therefore evaluate if a feature pre-filtering step might improve identification of features associated with true underlying groups.

Methods. Data was simulated for up to 100 samples and up to 10,000 features, 10% of which were associated with the ground truth comprising 2–10 normally distributed populations. Binary and semi-quantitative ratings with varying error probabilities were used as classification. For feature preselection standard cross-validation (V2) was compared to a novel heuristic (V1) applying univariate testing, multiplicity adjustment and cross-validation on switched dependent (classification) and independent (features) variables. Preselected features were used to train logistic regression/linear models (backward selection, AIC). Predictions were compared against the ground truth (ROC, multiclass-ROC). As use case, multiple feature selection/classification methods were benchmarked against the novel heuristic to identify prognostically different G-CIMP negative glioblastoma tumors from the TCGA-GBM 450 k methylation array data cohort, starting from a fuzzy umap based rough and erroneous separation.

Results. V1 yielded higher median AUC ranks for two true groups (ground truth), with smaller differences for true graduated differences (3–10 groups). Lower fractions of models were successfully fit with V1. Median AUCs for binary classification and two true groups were 0.91 (range: 0.54–1.00) for V1 (Benjamini-Hochberg) and 0.70 (0.28–1.00) for V2, 13% ($n = 616$) of V2 models showed AUCs $<= 50\%$ for 25 samples and 100 features. For larger numbers of features and samples, median AUCs were 0.75 (range 0.59–1.00) for V1 and 0.54 (range 0.32–0.75) for V2. In the TCGA-GBM data, modelBuildR allowed best prognostic separation of patients with highest median overall survival difference (7.51 months) followed a difference of 6.04 months for a random forest based method.

Conclusions. The proposed heuristic is beneficial for the retrieval of features associated with two true groups classified with errors. We provide the R package modelBuildR to

simplify (comparative) evaluation/application of the proposed heuristic (http://github.com/mknoll/modelBuildR).

INTRODUCTION

Model training is an important task in biomedical research for the evaluation of omics data, e.g., for classification tasks. The features included in the model and used for classification might hint towards underlying (biological) processes or mechanisms.

Such classifications in biomedical research are often encoded by a human rater as binary, e.g., a given immune-histochemistry staining can be classified as positive/negative (1/0), or as semi-quantitative score (e.g., 0–5) for graduated evaluation (*Balermpas et al., 2017*; *Knoll et al., 2016*). Often, associated changes on molecular level are of interest, measured e.g., by analysis of expression or methylation data with arrays/sequencing yielding a high number of features.

Binary outcome data can be modeled using a logistic regression, a generalized linear model (GLM) with logit link function (*Hastie & Tibshirani, 1986*; *McCullagh & Nelder, 1989*). Semi-quantitative data might be evaluated using linear models.

For model training, a full evaluation of all feature combinations is usually not feasible (high number of features), and standard GLMs cannot be trained for numbers of features >numbers of observations, requiring the usage of heuristics for pre-filtering of features. A set of remaining features can then be used to train a model, e.g., using backward selection in combination with an information criterion (*Akaike, 1973*).

Model fits are usually evaluated for their ability to predict the observed data ("goodness-of-fit"). The latter might, however, contain erroneous assignments, arising e.g., from multiple sources (technical difficulties, sampling or human error). Thus, forcing the model to fit the observed rather than the true underlying groups might lead to the selection of inappropriate features.

We therefore propose to use a heuristic for feature pre-filtering prior to model building which reverses the role of dependent (classification)/independent (features) variables, perform tests for difference and cross validate data with reverted roles of dependent/independent variables, and use only retained features for subsequent model building.

Its performance is compared to a standard cross validation approach (non-reverted roles of variables) in simulated data. Binary and semi-quantitative encodings (with added errors) in features sampled from two or more populations are evaluated, and the ability of both approaches to select meaningful features (high overlap with known ground truth).

We provide an R package to simplify (comparative) analyses with the proposed heuristic, available on github (http://github.com/mknoll/modelBuildR).

METHODS

Feature selection methods

The two evaluated feature selection methods are outlined in Figs. 1B and 2. Variant 2 (V2) uses cross validation to obtain an order on single features (univariate test), using the (erroneous) classification as dependent variable. For binary outcomes, cv.binary() from the DAAG package (*Maindonald & Braun, 2020*) and for semi-quantitative outcomes, cv.lm() was used (default parameters, Fig. 3). The first n features with lowest cross-validation errors or highest accuracy were selected for model selection, with n being the number of evaluated samples (here: 50). Further processing was similar between both evaluated methods. Variant 1 inverts the role of dependent (classification)/independent (features) variables for the initial feature filtering step. First, a significant influence of the observed classification on each measured feature is tested using a linear model and calculating model p-values (null- vs full models, likelihood ratio test, LRT). P-values are then adjusted for multiplicity, Benjamini–Hochberg and Bonferroni adjustment was evaluated, all features with adjusted p-values below 0.05 ($p* = 0.05$, Fig. 1B) were retained. Next, a cross validation step was performed, keeping the inverted roles of independent/dependent variables. Finally, the first n features with lowest cross-validation errors were used for further analysis, with n being the minimum of the number of remaining features and numbers of samples. For the next step, which is similar to variant 2, the original roles of the dependent and independent variables were assumed (classification: dependent variable). Model building was performed by backward model selection using AIC, with a logistic regression for binary outcomes and a linear model for semi-quantitative classification. Predictions were then compared to known underlying group truth by calculation AUCs with pROC::roc() (*Robin et al., 2011*) for binary and pROC::multiclass.roc() for semi-quantitative classifications (Fig. 2). AUCs and AUC ranks (tie methods: average, random) were evaluated.

Evaluated data

An overview of simulated data gives Fig. 2. Two common cases were tested, a binary classification and a semi-quantitative graduated classification together with high-dimensional data. To keep calculation time reasonable, a total number of 100 features was evaluated in 50 samples. Only a fraction of features ($\sim$10%, sampled with runif()) were assumed to show differences between groups. For two group analyses, they were sampled from two normal distributions with varying differences in means and standard deviations. For more than two classes, differences between means of subsequent classes/distributions were constant, as were their standard deviations. Group sizes were balanced, if this was not possible, the sample number of the highest ranked group was expanded. To assure reproducibility, a fixed seed was used. Errors on the classifications were introduced as follows: for a binary classification, the respective group assignment was retrieved from a binomial distribution yielding 0,1 with probabilities prob1 and prob2. For larger numbers of classes, a vector of 0,1 values was obtained similarly for a probability prob1, and
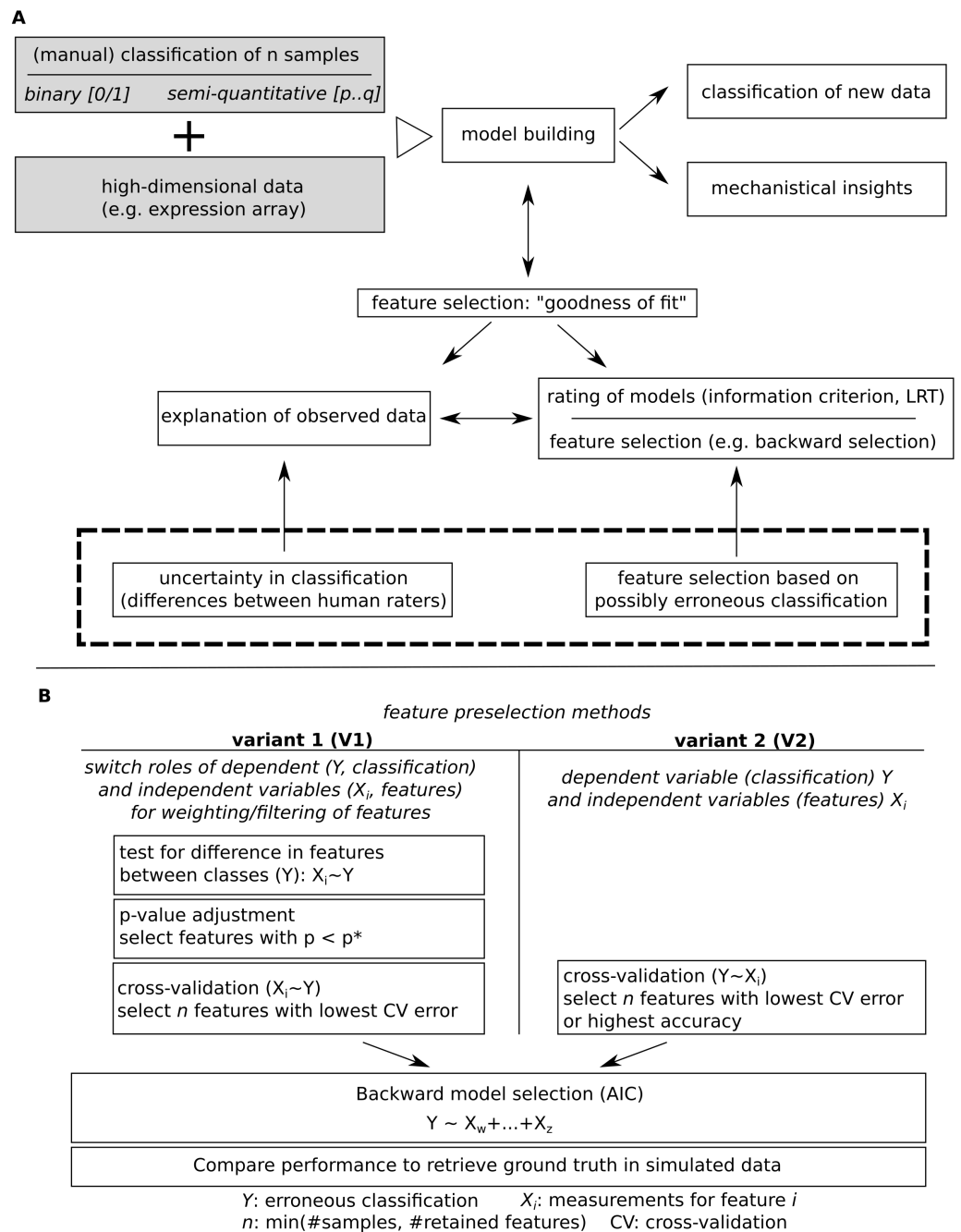
Figure 1 Model building for a given classification using high dimensional data. (A) Major steps in model building and outline of the subsequently addressed issue of potential erroneous classification. (B) Comparatively evaluated strategies for feature pre-filtering/ordering of features prior to model building and outline of how model performance is evaluated.

A | **binary classification**

ground truth

$n_1=25$ $n_2=25$

000...000 111...111

sample classification: binomial distribution {0,1}
  probability: 0.1,0.3,0.6,0.9
    per group: prob1, prob2

010...000 110..001

simulate 100 features

  ~10% features are different between groups

  varying mean difference: 0.1,0.5,1,1.5
  varying sd: 0.1,0.5,1,1.5,2,3

feature weighting / preselection

cross-validation (CV)   LRT + CV
$(Y \sim X_i)$ $(X_i \sim Y)$

train model against erroneous classification

model1   model2
$Y \sim X_1 + ... + X_j$

ROC analysis of model prediction against
ground truth

AUC1   AUC2

B | **semiquantitative classification**

	I	**II**

I: ground truth: 3-10 classes
II: ground truth: two classes (see above)

1..1 2..2 3..3   000...000 111...111

add error: binomial distribution +/- {0,1}
  probability: 0.1,0.3,0.6,0.9

1..2 2..1 2..3   1..2 2..1 2..3

simulate 100 features

  ~10% features are different between groups

  varying mean difference: 0.1,0.5,1,1.5
  varying sd: 0.1,0.5,1,1.5,2,3

feature weighting / preselection

CV   LRT + CV | CV   LRT + CV
$(Y \sim X_i)$ $(X_i \sim Y)$ | $(Y \sim X_i)$ $(X_i \sim Y)$

train model against erroneous classification

model1   model2 | model1   model2
$Y \sim X_1 + ... + X_j$ | $Y \sim X_1 + ... + X_j$
$Y \in \mathbb{N}$ | $Y \in \{0,1\}$

ROC analysis of model prediction against
ground truth

pROC::multiclass.roc() | pROC::roc()

Figure 2 **Simulated data for binary (A) and semi-quantitative classification (B) and the introduction of errors used for evaluation of feature selection/model building approaches.** Observed classification was sampled from a binomial distribution for varying probabilities per group, for semi-quantitative data, equidistance between classes was assumed, and errors were added based on data sampled from binomial distributions.
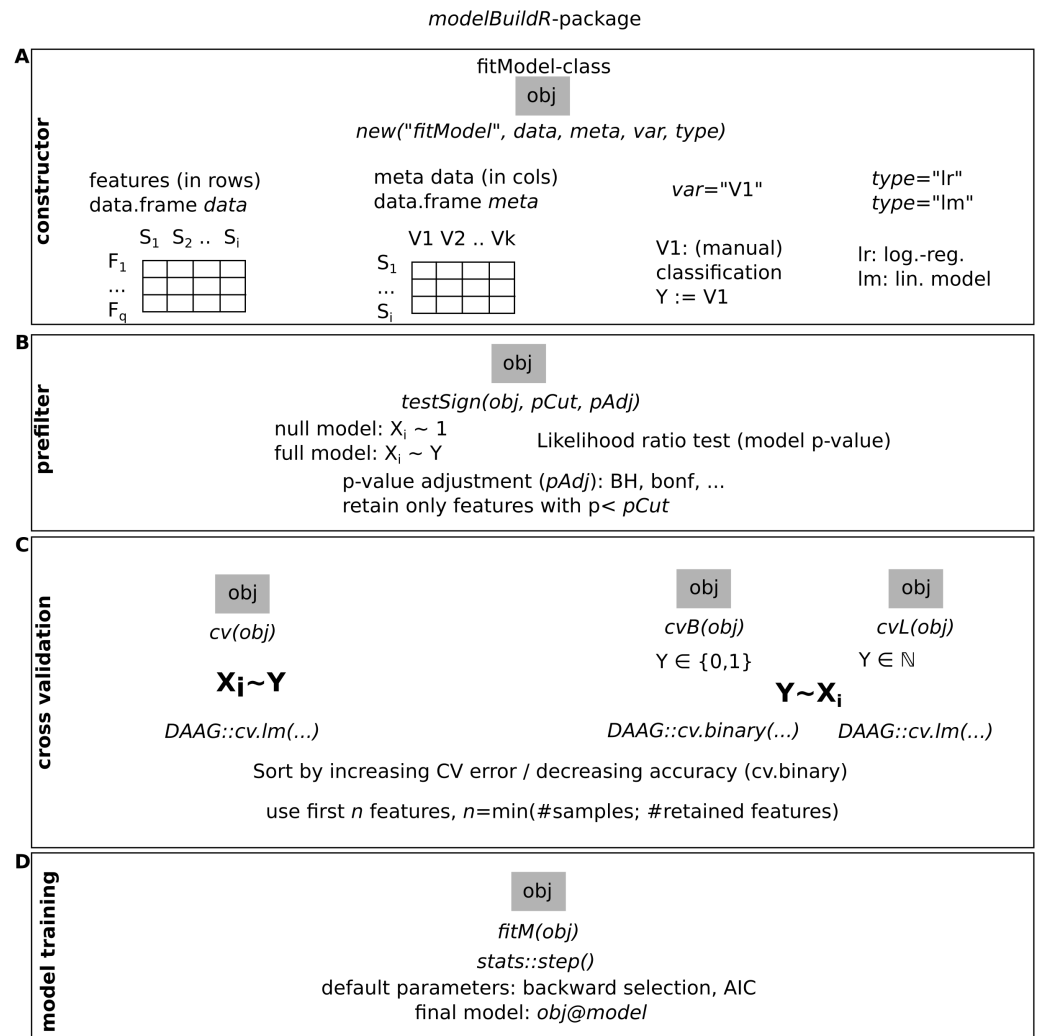
Figure 3 **Overview of the functions implemented in the modelBuildR package and required parameters.** (A) Instantiation of a fitModel object for analysis. (B) Prefiltering of features. (C) Cross-validation approaches. (D) Final model fitting.

was subtracted from the true classification. Absolute values were used as erroneous classification. Reproducible code and analyses are available as CodeOcean capsule: https://codeocean.com/capsule/3333162/tree/v1.

modelBuildR package

The presented analyses were performed with the modelBuildR package. An overview of its functionality is shown in Fig. 3, additional functionality is outlined in the package vignette.

The constructor for a new fitModel instance requires a feature data.frame data with features in rows and samples in colums, a metadata data.frame meta with samples in rows and covariates in columns, specification of the classification (dependent) variable var and the type of model to train (type, lr for logistic regression and lm for linear model). The evaluation of an association of the classification variable on feature measurements

is performed with testSign(), expecting a multiplicity adjustment parameter (pAdj, all allowed methods from stats::p.adjust()) and a *p*-value cutoff (pCut). Different cross-validation methods are implemented, cv() performs the cross validation on inverted roles of dependent/independent variables as described above (Figs. 1B and 2). cvB() and cvL() perform cross-validation on non-inverted roles of variables. fitM() finally performs the model training using R's stats::step() function with default parameters (using AIC or BIC) or using a cross-validation approach (see Suppl. Methods for details).

In addition to previously outlined analyses, the feature preselection step can also be performed while including additional covariates (both for the model evaluation and cross validation step, refer to the package vignette for details: vignette("modelBuildR")).

Omics data and alternative feature selection methods

450k Illumina human methylation array data and clinical information of the TCGA-GBM cohort was retrieved through the GDC data portal on 2019-11-07. Logit transformed methylation data was used for analysis if not stated otherwise (*M* values). G-CIMP classification was performed as follows: $L = 282.7 + 114.2*cg06903384$, $p = exp(L)/(1 + exp(L))$. Samples with $p < 0.5$ were classified as CIMP- and CIMP+ otherwise. The glmnet package (*Friedman, Hastie & Tibshirani, 2010*) was used for lasso regression, utilizing cross validation to select an appropriate lambda value, and randomForest (*Liaw & Wiener, 2002*) for random forest analysis. Student's *t*-tests were used. Optimal cutoffs of prognostic separation (minimal *p*-value) were calculated with dataAnalysisMisc::findOptCutoff() (*dataAnalysisMisc, 2020*). The pvclust package (*pvclust, 2019*) was used for consensus clustering, the umap R package in combination with umap-learn for dimensionality reduction (*Konopka, 2020*; *McInnes & Healy, 2018*). Significance level alpha was fixed at 0.05 (two-sided).

RESULTS

Semi-quantitative classification and true graduated differences in underlying data

Model fitting (>0 as significantly different identified features, *pAdj* < 0.05, Benjamini–Hochberg adjustment) was successful more frequently when using V2. V1 showed lower fractions for fewer categories (successful model fits, reference V2: median: 88%, range: 70–97%, Fig. 4A). Observed median AUCs were similar between V1 and V2 (Fig. 4B) and did not differ between *p*-value adjustment methods (V1, Fig. 4B). AUCs stratified by numbers of true underlying groups and mean difference for BH and Bonferroni adjustments are shown in Figs. S2 + S3. Differences between V1/V2 of observed AUCs decreased for increasing numbers of categories, single outliers were observed mostly for V1. Higher uncertainty in classification (prob1 0.3/0.6) showed lower AUCs especially for fewer groups (Figs. S2 + S3). Larger standard deviations with smaller mean differences decreased AUCs for V1, more prominent for Bonferroni than for Benjamini–Hochberg adjustment (Figs. S2 + S3). Median model fitting time requirements were lower for V1 (Fig. S1).
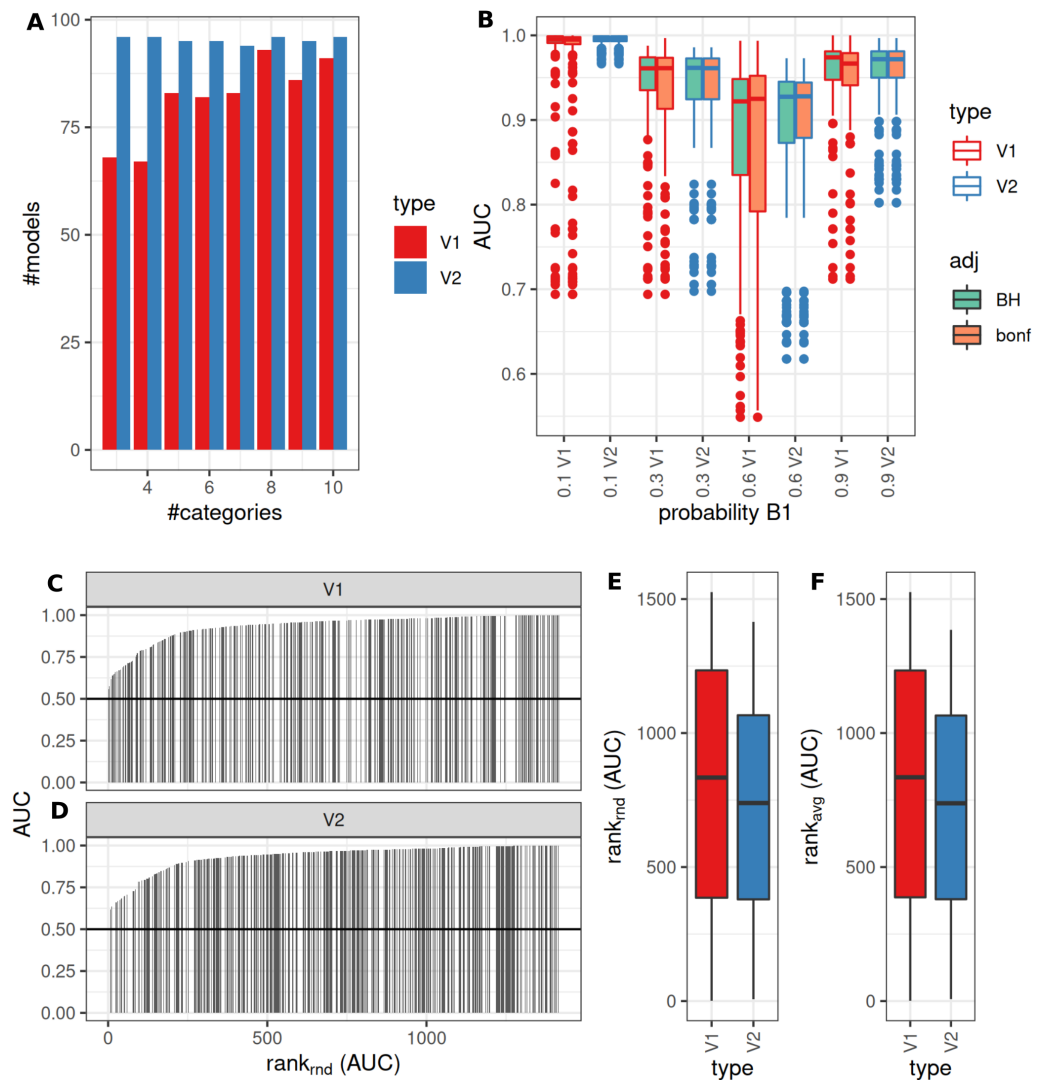
Figure 4 Comparison of feature preselection methods for semi-quantitative classification with true underlying equidistant differences between groups. (A) Number of successfully trained models (V1, Benjamini-Hochberg multiplicity adjustment). (B) AUCs of model predictions tested against ground truths for varying classification errors and different multiplicity adjustment methods. AUCs (C–D) and AUC ranks (E–F) of V1 and V2, rank-ties methods: avg, average, rnd, random.

Binary classification and two true underlying groups

The number of successful model fits ranged between 139 and 144 for V2 and 0 and 144 for V1 (Benjamini–Hochberg adjustment, Fig. 5A). Processing times were lower for V1 (<5 vs >20 s, Fig. 5B, Fig. S1). Minimum observed AUCs were >0.5 for all combinations evaluated with V1, 13% of models lead to an AUC <= 0.5 for V2 (Fig. S5). Separate analysis for combinations of prob1/prob2 showed that 0.1/0.1; 0.3/0.1; 0.1/0.3; 0.3/0.3; 0.6/06; 0.6/0.9; 0.9/0.9 did not yield any models for V1 (Fig. 5A), V2 identified models with low AUCs in these cases (Figs. S4 + S5). No general difference in AUCs between Bonferroni and Benjamini–Hochberg adjustment could be detected when separating

Peer
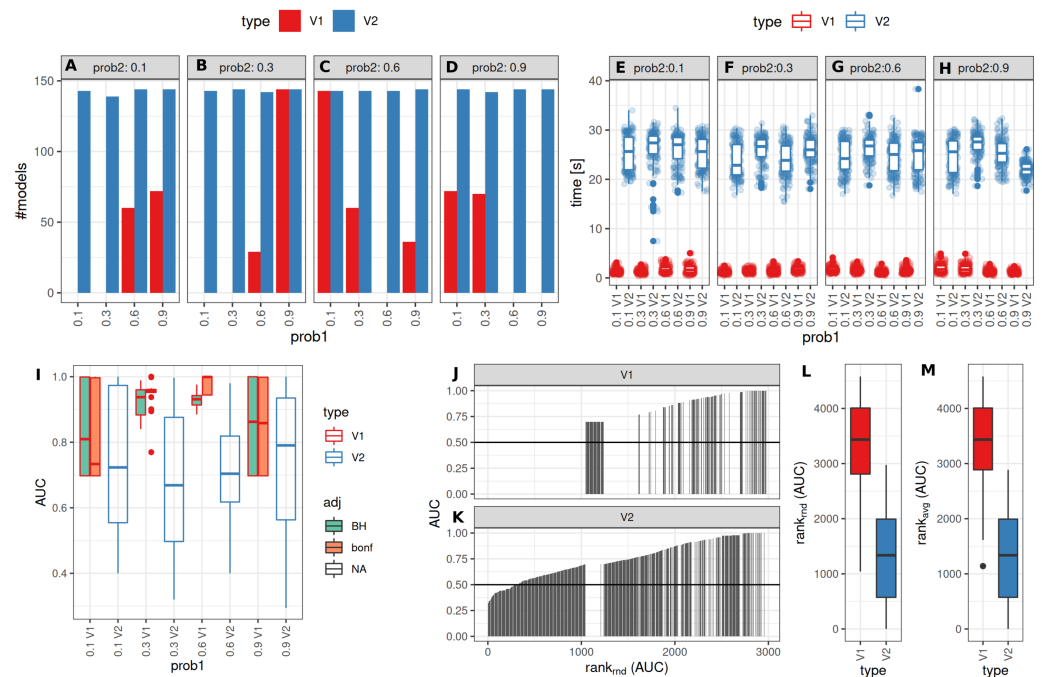
Figure 5 Comparison of feature preselection methods for binary outcomes with underlying true dichotomous groups. (A–D) Number of successfully trained models (Benjamini-Hochberg multiplicity adjustment). (E–H) Time requirements for model fitting. (I, J, K) AUCs of model predictions tested against ground truths for varying classification probabilities. (L–M) AUC ranks for V1 and V2, rank-ties methods: avg, average, rnd, random.

results by probability (Figs. S4 + S5). Increasing mean differences allowed model fitting for larger standard deviations with V1, with higher median AUCs for Bonferroni adjustment for most evaluated combinations (Figs. S4 + S5, Fig. 5C). Higher median AUC ranks were observed for V1 (Fig. 5D), as well as higher median AUCs (Fig. 5C).

Performance of both approaches were additionally tested for larger numbers of features (up to 10,000) and higher number of samples per group (same size, up to 100) and with probabilities 0.1 and 0.3 with Bonferroni p-value adjustment. Results are shown in Fig. S7. For $n = 25$ samples per group, no models could be fitted with V1. Both AUCs and AUC ranks were higher for V1, up to AUCs of 1 where the corresponding models from V2 reached AUCs not above 0.8. Minimum observed AUCs for V1 were 0.6. V2 did not show a clear influence of distribution parameters from which the data was sampled on AUCs as opposed to V1. In summary, V1 outperforms V2 also in larger datasets.

Semi-quantitative classification and two true underlying groups

An intermediate between the two previously analyzed conditions was evaluated next. Classification was allowed to be graduated (semi-quantitative), but the underlying grouping was assumed to be dichotomous. V1 lead to model fits of median 35% (range: 16–48%, Benjamini–Hochberg p-value adjustment) of successful model fits using V2 (Fig. 6A). Processing time was lower for V1 (Fig. S1). V1 yielded higher AUC ranks as compared to
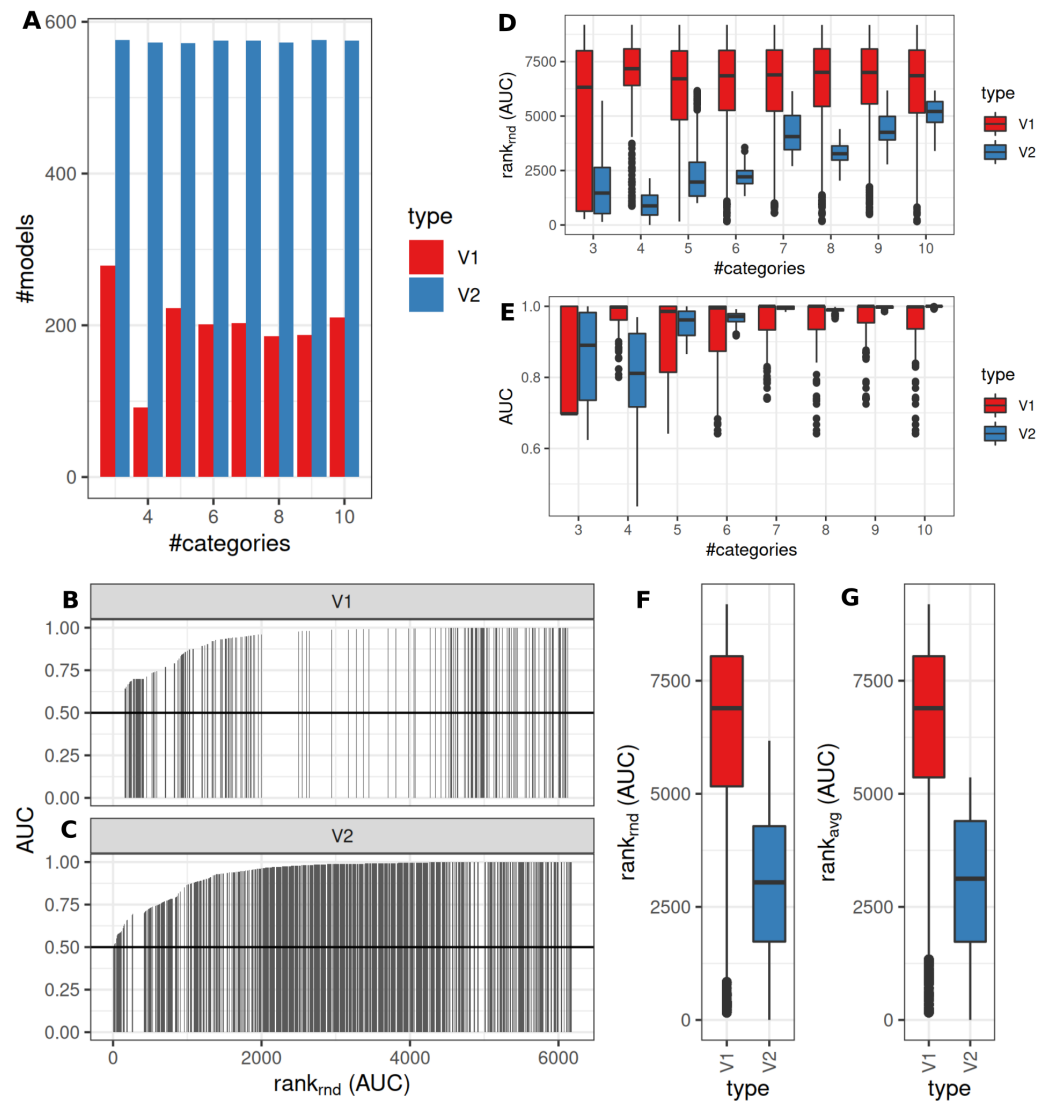
Figure 6 **Comparison of feature preselection methods for semi-quantitative equidistant classification with true dichotomous underlying groups.** (A) Number of successfully trained models (Benjamini-Hochberg multiplicity adjustment). AUCs (B–C) and AUC ranks (F–G) ranks of V1 and V2, rank-ties methods: avg, average, rnd, random. (D) AUCs and ranks (E) split by numbers of semi quantitative categories.

V2 (Fig. 6B). Minimum observed AUCs were 0.64 for V1 and 0.44 for V2. V2 yielded 5 models with AUCs <= 0.5. Stratification by numbers of categories showed higher median AUC ranks for V1 and increasing median ranks for V2 (Fig. 6C), as well as increases in AUCs for >= 4 categories for V2 (Fig. 6C). Stratification of AUCs by error probability (prob1) and number of categories showed decreases of AUCs for increasing error probabilities especially for V2 (Fig. S6). Dependency of AUCs on numbers of categories, mean difference and standard deviations between the two underlying groups showed higher AUCs for lower standard deviations especially for 4 groups for V1 (Fig. S6).

Use case: methylation based identification of prognostically different CIMP-glioblastomas

To comparatively evaluate the proposed heuristic, we assessed a number of methods for their ability to identify/retrieve two assumedly true groups of prognostically different G-CIMP- GBM tumors present in the TCGA-GBM 450k methylation array data cohort (Fig. 7). Two prognostically different groups (long-term survivors, LTS and short-term survivors, STS) were defined as outlined in Fig. S7 and Suppl. Methods. An umap representation was calculated from methylation array data, distribution of LTS/STS samples is shown in Fig. 7A. LTS tumors are rather located in the lower right part, STS tumor in the upper part of the graph. Data-driven separation of samples, based on the umap representation, was performed manually with a straight line (Fig. 7B). The resulting grouping of samples (above, below the line, grp1 and grp2) was used to train a random forest classifier, a lasso regression and a logistic regression with the proposed heuristic. For the random forest classifier, an additional analysis was performed by selecting the highest ranked CpG probes (importance, mean decrease Gini, Fig. 7E, 2nd to 4th column) for subsequent hierarchical cluster analysis. Additional methods for feature selection are shown in Figs. 7F–7H. Predictions from the random forest classifier, two main clusters for approaches involving hierarchical [consensus] clustering and optimal prognostic separation of continuous values (predictions from lasso and the novel heuristic, minimum p-values) were compared w.r.t. their ability for prognostic separation (Figs. 7D–7H and Table 1). Best separation was achieved with the novel heuristic (median survival difference of 7.51 months), followed by random forest classifier with hierarchical cluster analysis of most important CpGs (6.04 months). Selection of BIC, AIC or CV approach in fitM() yielded the same model (Table S1).

DISCUSSION

Omics-data, e.g., expression or methylation data is often used to gain insights the underlying biology (*Capper et al., 2018*). In translational research, molecular data from patients is often compared against a given binary classification (e.g., tumor subtype A vs B) or a graduated semi-quantitative rating, e.g., of an immune-histochemical straining of intensity classes 1 to 5 (*Balermpas et al., 2017*; *Knoll et al., 2016*). However, prospectively measured inter-rater agreement for classification of grade and histotype of ovarian cancer by specialists (pathologists) has been reported with only 85–95% (*Barnard et al., 2018*). Thus, a fraction of misclassifications of 5–15% might be considered a conservative estimation even for trained raters.

Model training for classification might hint towards underlying biological mechanisms as the model training step is assumed to select features which robustly allow to infer groups. Evaluation of all possible combinations of features for model training is not feasible for typical datasets and for standard modelling approaches also often not possible (numbers of features >> numbers of samples), ridge regression and lasso (*Santosa & Symes, 1986*; *Tibshirani, 1996*) allow to deal with such data. Alternatively, features can be pre filtered with a wide variety of methods (*Lazar et al., 2012*). Binary outcomes can be
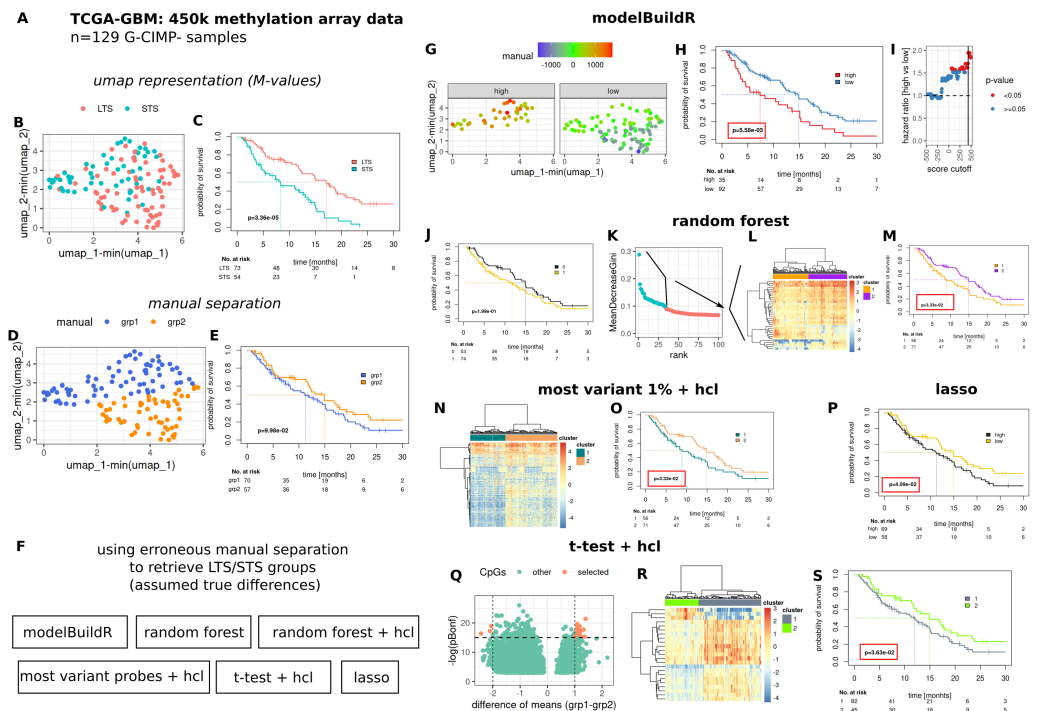
Figure 7 **Comparative evaluation of the novel proposed heuristic to identify prognostically different G-CIMP-tumors from methylation array data.** (A) Evaluated data. (B) Umap representation of M-values with LTS/STS classification (see Fig. S7 and Suppl.-Methods) and corresponding survival curves (C). Manual separation of prognostically different tumors, umap (D) and survival curves (E). (F) Evaluated approaches to detect groups of prognostically different tumors. hcl: hierarchical cluster analysis. (G–I) modelBuildR heuristic, (G) color-coded model scores in umap representation of methylation data and separated by prognostic group (high/low, see H). (H) Survival curves correspond to best achievable separation (minimal *p*-value, vertical line, I). Random forest predictions (J), random forest derived ranking of CpG probes (K, importance, mean decrease Gini), hierarchical cluster analysis of selected probes (L, blue, ward.D2, Euclidean distance), survival curves of two main clusters (M). (F) Hierarchical cluster analysis (ward.D2, Euclidean distance) of 1% of most variant probes (median absolute deviation) and corresponding survival curves of two main clusters (O). (P) Survival curves for lasso regression model predictions, analogously to I. (Q) Volcano plot of differentially regulated probes (*t*-test, Bonferroni adjustment), selected probes were used for consensus clustering (R, hcl, ward.D2, Euclidean), prognostic separation of two main clusters (S). Kaplan-Meier survival curves, likelihood ratio test p-values (Cox-PH models).

modeled with logistic regressions and graduated, equidistant classifications with linear models. For a discussion of currently applied methods for the analysis of real-world clinical data—starting from simple ROC based analyses to complex models and feature selection approaches—see *Chen et al. (2019)* and *Deo (2015)*. Even though deep-learning models might show extraordinary high performance for specific tasks in biomedical research, their application if often limited by sparsity of data or low quality (*Chen et al., 2019*) and might be vulnerable to small adversarial perturbations (*Yuan et al., 2019*). Therefore, novel methods are needed with specifically enable analysis using poorer quality data. Furthermore, highly complex and powerful deep-learning methods lack transparency (*Holzinger et al., 2019*), but explainability and interpretability often is a crucial point needed to gain a more mechanistic understanding of underlying processes.

Table 1 Comparison of different methods for prognostic separation of G-CIMP negative glioblastoma tumors based on methylation array data.

Method	Median survival difference [months]	HR, 95% CI	*p*-value
Random Forest	3.2	1.3 [0.86–2.03]	0.2
Random Forest + hcl	6.04	0.5 [0.41–0.96]	0.03
Most variant + hcl	4.5	0.8 [0.54–1.30]	0.4
t-test + hcl*	3.0	0.6 [0.39–0.98]	0.04
Lasso	3.7	0.6 [0.42–0.99]	0.04
modelBuildR	7.51	0.5 [0.33–0.81]	0.004

Notes.
Hcl, hierarchical cluster analysis; *, consensus clustering; HR, hazard ratio, Cox-PH models.

Methods often aim to explain the observed data (classification) as good as possible, e.g., by using goodness of fit tests or sufficient differences in information criteria, while addressing overfitting e.g., by incorporating cross-validation. However, selection of features is still based on (probable) erroneous classification.

We aimed to evaluate if a heuristic which inverts the roles of dependent (classification) and independent (features) variables in a pre-filtering step might help to retrieve features associated with the true underlying structure/grouping (testing for significant differences, cross validation). Therefore, we simulated data for two or more distinct classes, added an error on the classification and tried to retrieve the original classification as quantified by (multiclass) ROC analyses.

Evaluation of true different populations encoded semi-quantitatively showed no global preference for V1 or V2 except for lower time requirements for V1. AUC ranks were still higher for V1, thus making V1 a reasonable analysis approach. The presence of only small differences between populations, however, might impair performance in this setting.

The presence of two true groups can be encoded binary by a (human) rater or, e.g., for immune histochemical stainings, graduated even though only two groups are present. Both combinations were evaluated, showing a clear overall benefit of the proposed heuristic for binary encodings. This was not only true for systematic analyses with few numbers of features ($n = 100$) and 25 samples per group, but also for larger datasets with up to 10,000 features and 100 samples per group. For semi-quantitative encodings, a better performance was seen for lower numbers of semi-quantitative categories. Thus the heuristic can be recommended for binary classified data, and if only few categories (∼4) are used for classification if a binary ground truth might be present. Due to large time requirements, only the combination yielding a clear benefit (two groups, binary classification), was tested with larger numbers of features and samples.

The proposed heuristic for feature pre-filtering leads to a number of combinations where no model could be fit. These combinations, however, would have led to models with low AUCs (compared to the ground truth) using the cross validation only feature selection strategy (V2). More liberal *p*-value adjustment strategies were not always beneficial, thus

performance of different multiplicity adjustment procedures with varying *p*-value cutoffs while considering their respective power should be evaluated in future work. Lower time-requirements of the heuristic might prove useful especially for larger datasets.

We utilized the TCGA-GBM 450k methylation array data cohort of G-CIMP negative tumors to demonstrate the ability of the proposed heuristic to retrieve features able to separate probable true different underlying groups of tumors. Direct comparison with additional methods, even for only a small number of approaches, showed a superior performance of the novel heuristic. Without interpreting too much into the potential biological meaning (no independent validation), it is worth noting that methylation array data is used to detect and classify separate subgroups of glioma and G-CIMP- glioblastoma, which also show differences in prognosis (*Capper et al., 2018*; *Knoll et al., 2019*; *Hwang et al., 2019*).

In summary, the proposed heuristic proved most beneficial for the identification of two groups encoded in two or few categories. Identified features were then more probable to represent true associated characteristics. However, future work is needed to validate these findings in more complex/real-world data with e.g., unbalanced groups, larger sample sizes and multiple (non-)correlated true effects in underlying data. For an easy application of such benchmarks, our modelBuildR package can be used and is made publicly available on github.

CONCLUSIONS

In biomedical research, misclassification is not negligible with reported error rates up to 15%. Classical feature selection methods, however, assume that a provided labeling is correct and select features best explaining potentially erroneous data, even though interest lies in true underlying groups. We propose a novel feature selection heuristic which inverts roles of dependent and independent variables in an initial feature selection step and proceeds with standard methods. Its superior performance in identifying features associated with the ground truth even for wrongly labeled samples is demonstrated in synthetic data arising from two true groups and binary manual encoding. A use case with methylation array omics data shows promising results. Further work is needed to better characterize applications for which the proposed heuristic might be beneficial.

ACKNOWLEDGEMENTS

We thank the Heidelberg Medical Faculty at Heidelberg University for their financial support within the scope of the MD/PhD program to MK and JF.

Funding

This work was supported by the National Center for Tumor diseases (NCT PRO-2015.21), Deutsche Forschungsgemeinschaft (UNITE SFB13-89) and the German Cancer Research Center (iMED). Maximilian Knoll and Jennifer Furkel received financial support from

the Heidelberg Medical Faculty of Heidelberg University within the scope of the MD/PhD program. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Center for Tumor diseases: NCT PRO-2015.21.
Deutsche Forschungsgemeinschaft: UNITE SFB13-89.
German Cancer Research Center: iMED.
Heidelberg Medical Faculty of Heidelberg University within the scope of the MD/PhD program.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Maximilian Knoll conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Jennifer Furkel, Juergen Debus and Amir Abdollahi conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The R package is available at Github: http://github.com/mknoll/modelBuildR.
Code for the simulation study are accessible at CodeOcean:
Maximilian Knoll (2020) modelBuildR: An R package for model building and feature selection with erroneous classifications. [Source Code]. https://doi.org/10.24433/CO.3805663.v1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10849#supplemental-information.

REFERENCES

Akaike H (ed.) 1973. *Information theory and an extension of the maximum likelihood principle.* Budapest: Akadémiai Kiadó.

Balermpas P, Rodel F, Krause M, Linge A, Lohaus F, Baumann M, Tinhofer I, Budach V, Sak A, Stuschke M, Gkika E, Grosu AL, Abdollahi A, Debus J, Stangl S, Ganswindt U, Belka C, Pigorsch S, Multhoff G, Combs SE, Welz S, Zips D, Lim SY, Rodel C, Fokas E, Dktk ROG. 2017. The PD-1/PD-L1 axis and human papilloma virus in patients with head and neck cancer after adjuvant chemoradiotherapy: a multicentre study of the German Cancer Consortium Radiation

Oncology Group (DKTK-ROG). *International Journal of Cancer* **141(3)**:594–603 DOI 10.1002/ijc.30770.

Barnard ME, Pyden A, Rice MS, Linares M, Tworoger SS, Howitt BE, Meserve EE, Hecht JL. 2018. Inter-pathologist and pathology report agreement for ovarian tumor characteristics in the Nurses' Health Studies. *Gynecologic Oncology* **150(3)**:521–526 DOI 10.1016/j.ygyno.2018.07.003.

Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE, Kratz A, Wefers AK, Huang K, Pajtler KW, Schweizer L, Stichel D, Olar A, Engel NW, Lindenberg K, Harter PN, Braczynski AK, Plate KH, Dohmen H, Garvalov BK, Coras R, Holsken A, Hewer E, Bewerunge-Hudler M, Schick M, Fischer R, Beschorner R, Schittenhelm J, Staszewski O, Wani K, Varlet P, Pages M, Temming P, Lohmann D, Selt F, Witt H, Milde T, Witt O, Aronica E, Giangaspero F, Rushing E, Scheurlen W, Geisenberger C, Rodriguez FJ, Becker A, Preusser M, Haberler C, Bjerkvig R, Cryan J, Farrell M, Deckert M, Hench J, Frank S, Serrano J, Kannan K, Tsirigos A, Bruck W, Hofer S, Brehmer S, Seiz-Rosenhagen M, Hanggi D, Hans V, Rozsnoki S, Hansford JR, Kohlhof P, Kristensen BW, Lechner M, Lopes B, Mawrin C, Ketter R, Kulozik A, Khatib Z, Heppner F, Koch A, Jouvet A, Keohane C, Muhleisen H, Mueller W, Pohl U, Prinz M, Benner A, Zapatka M, Gottardo NG, Driever PH, Kramm CM, Muller HL, Rutkowski S, Hoff Kvon, Fruhwald MC, Gnekow A, Fleischhack G, Tippelt S, Calaminus G, Monoranu CM, Perry A, Jones C, Jacques TS, Radlwimmer B, Gessi M, Pietsch T, Schramm J, Schackert G, Westphal M, Reifenberger G, Wesseling P, Weller M, Collins VP, Blumcke I, Bendszus M, Debus J, Huang A, Jabado N, Northcott PA, Paulus W, Gajjar A, Robinson GW, Taylor MD, Jaunmuktane Z, Ryzhova M, Platten M, Unterberg A, Wick W, Karajannis MA, Mittelbronn M, Acker T, Hartmann C, Aldape K, Schuller U, Buslei R, Lichter P, Kool M, Herold-Mende C, Ellison DW, Hasselblatt M, Snuderl M, Brandner S, Korshunov A, Deimling Avon, Pfister SM. 2018. DNA methylation-based classification of central nervous system tumours. *Nature* **555(7697)**:469–474 DOI 10.1038/nature26000.

Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, Naessens JM, Larson DW, Liu H. 2019. Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digital Medicine* **2**:Article 43 DOI 10.1038/s41746-019-0122-0.

dataAnalysisMisc. 2020. Collection of functions for daily tasks. R package version 0.99.11. *Available at http://github.com/mknoll/dataAnalysisMisc*.

Deo RC. 2015. Machine learning in medicine. *Circulation* **132(20)**:1920–1930 DOI 10.1161/CIRCULATIONAHA.115.001593.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33(1)**:1–22.

Hastie T, Tibshirani R. 1986. Generalized additive models. *Statistical Science* **1(3)**:297–310 DOI 10.1214/ss/1177013604.

Holzinger A, Langs G, Denk H, Zatloukal K, Muller H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews* **9(4)**:e1312.

Hwang T, Mathios D, McDonald KL, Daris I, Park SH, Burger PC, Kim S, Dho YS, Carolyn H, Bettegowda C, Shin JH, Lim M, Park CK. 2019. Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta Neuropathologica Communications* **7(1)**:Article 88 DOI 10.1186/s40478-019-0744-0.

Knoll M, Debus J, Furkel J, Warta R, Bougatf N, Rapp C, Brors B, Wick W, Unterberg A, Herold-Mende C, Abdollahi A. 2019. Glioblastoma evolution pattern under surgery and radio(chemo)therapy (RCHT) to identify novel methylome based glioma subtypes. *Journal of Clinical Oncology* **37(suppl 15)**:2012–2012 DOI 10.1200/JCO.2019.37.15_suppl.2012.

Knoll M, Macher-Goeppinger S, Kopitz J, Duensing S, Pahernik S, Hohenfellner M, Schirmacher P, Roth W. 2016. The ribosomal protein S6 in renal cell carcinoma: functional relevance and potential as biomarker. *Oncotarget* **7(1)**:418–432 DOI 10.18632/oncotarget.6225.

Konopka T. 2020. unap: Uniform manifold approximation and projection. *Available at https://CRAN.R-project.org/package=umap*.

Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, Schaetzen VD, Duque R, Bersini H, Nowe A. 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9(4)**:1106–1119 DOI 10.1109/TCBB.2012.33.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2(3)**:18–22.

Maindonald JH, Braun WJ. 2020. DAAG: Data analysis and graphics data and functions. *Available at https://CRAN.R-project.org/package=DAAG*.

McCullagh P, Nelder JA. 1989. *Generalized linear models*. Second Edition. New York: Chapman & Hall/CRC.

McInnes L, Healy J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv 2018.

pvclust. 2019. Hierarchical clustering with *P*-values via multiscale bootstrap resampling. *Available at https://CRAN.R-project.org/package=pvclust*.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**:77 DOI 10.1186/1471-2105-12-77.

Santosa F, Symes WW. 1986. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing* **7(4)**:1307–1330 DOI 10.1137/0907087.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58(1)**:267–288.

Yuan X, He P, Zhu Q, Li X. 2019. Adversarial examples: attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems* **30(9)**:2805–2824 DOI 10.1109/TNNLS.2018.2886017.