Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers

Timo M. Deist[1,2,†,*], Frank J.W.M. Dankers[2,3,†], Gilmer Valdes[4], Robin Wijsman[3], I-Chow Hsu[4], Cary Oberije[2], Tim Lustberg[7], Johan van Soest[7], Frank Hoebers[7], Arthur Jochems[1,2], Issam El Naqa[5], Leonard Wee[7], Olivier Morin[4], David R. Raleigh[4], Wouter Bots[3,8], Johannes H. Kaanders[3], José Belderbos[6], Margriet Kwint[6], Timothy Solberg[4], René Monshouwer[3], Johan Bussink[3], Andre Dekker[7] and Philippe Lambin[1]

[1]The D-lab: Decision Support for Precision Medicine, GROW - School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Universiteitssingel 40, 6229 ER, Maastricht, The Netherlands; [2]Department of Radiation Oncology, GROW, School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands; [3]Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands; [4]Department of Radiation Oncology, University of California San Francisco, California, USA; [5]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan; [6]Department of Radiation Oncology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [7]Department of Radiation Oncology (MAASTRO), GROW, School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands; [8]Institute for Hyperbaric Oxygen (IvHG), Arnhem, The Netherlands.

[†]Authors contributed equally

[*]Corresponding author: t.deist@maastrichtuniversity.nl, The D-lab, GROW, Universiteitssingel 40 / 4.549, 6229 ER Maastricht, The Netherlands

Abstract

Purpose

Machine learning classification algorithms (classifiers) for prediction of treatment response are becoming more popular in radiotherapy literature. General machine learning literature provides evidence in favor of some classifier families (random forest, support vector machine, gradient boosting) in terms of classification performance. The purpose of this study is to compare such classifiers specifically for (chemo)radiotherapy datasets and to estimate their average discriminative performance for radiation treatment outcome prediction.

Methods

We collected 12 datasets (~~3496~~ 3484 patients) from prior studies on post-(chemo)radiotherapy toxicity, survival, or tumor control with clinical, dosimetric, or blood biomarker features from multiple institutions and for different tumor sites, i.e. (non-)small cell lung cancer, head and neck cancer, and meningioma. Six common classification algorithms with built-in feature selection (decision tree, random forest, neural network, support vector machine, elastic net logistic regression, LogitBoost) were applied on each dataset using the popular open-source *R* package *caret*. The *R* code and documentation for the analysis are available online[1]. All classifiers were run on each dataset in a 100-repeated nested 5-fold cross-validation with hyperparameter tuning. Performance metrics (AUC, calibration slope and intercept, accuracy, Cohen's kappa, and Brier score) were computed. We ranked classifiers by AUC to determine which classifier is likely to also perform well in future studies. We simulated the benefit for potential investigators to select a certain classifier for a new dataset based on our study (*pre-selection* based on other datasets) or estimating the best classifier for a dataset (*set-specific selection* based on information from the new dataset) compared to uninformed classifier selection (random selection).

Results

Random forest (best in 6/12 datasets) and elastic net logistic regression (best in 4/12 datasets) showed the overall best discrimination but there was no single best classifier across datasets. Both classifiers had a median AUC *rank* of 2. Pre-selection and set-specific selection yielded a significant average AUC improvement of 0.02 and 0.02 over random selection with an average AUC *rank* improvement of 0.~~42~~ 52 and 0.~~66~~65, respectively.

Conclusion

Random forest and elastic net logistic regression yield higher discriminative performance in (chemo)radiotherapy outcome and toxicity prediction than other studied classifiers. Thus, one of these two classifiers should be the first choice for investigators when building classification models or to benchmark one's own modelling results against. Our results also show that an informed pre-selection of classifiers based on existing datasets can improve discrimination over random selection.

Keywords: radiotherapy; classification; outcome prediction; machine learning; predictive modelling

Introduction

Machine learning algorithms for predicting (chemo)radiotherapy outcomes (e.g., survival, treatment failure, toxicity) are receiving much attention in literature, for example in decision support systems for precision medicine[2,3]. Currently, there is no consensus on an optimal classification algorithm. Investigators select algorithms for various reasons: the investigator's experience, usage in literature, data characteristics and quality, hypothesized feature dependencies, availability of simple implementations, and model interpretability. One objective criterion for selecting a classifier is to maximize a chosen performance metric, e.g., discrimination (expressed by the area under the receiver operating characteristic curve, AUC). Here, we discuss the performance of binary classifiers in (chemo)radiotherapy outcome prediction, i.e. algorithms that predict whether or not a patient has a certain outcome. We empirically study the behaviour of existing simple implementations of classifiers on a range of (chemo)radiotherapy outcome datasets to possibly identify a classifier with overall maximal discriminative performance. This is a relevant question for investigators who search for a rational basis to support their choice of a classifier or who would like to compare their own modelling results to established algorithms.

We employ various open-source *R* packages interfaced with the *R* package *caret*[4] (version 6.0-73) that is readily available for investigators and has shown to produce competitive results[5]. With our results, we also wish to provide guidance in the current trend to delegate modelling decisions to machine learning algorithms.

Large scale studies in the general machine learning literature[5–7] provide evidence in favor of some classifier families (random forest (*rf*), support vector machine (*svm*), gradient boosting machine (*gbm*)) in terms of classification performance. In our study, we investigate how these results translate to (chemo)radiotherapy datasets for treatment outcome prediction/prognosis. To the best of our knowledge, this is the first study to investigate classifier performance on a wide range of such datasets. The studied features are clinical, dosimetric, and blood biomarkers.

Within the framework of existing classifier implementations, we attempt to answer three research questions:

(1) Is there a superior classifier for predictive modelling in (chemo)radiotherapy?
(2) How dataset-dependent is the choice of a classifier?
(3) Is there a benefit of choosing a classifier based on empirical evidence from similar datasets (*pre-selection*)?

Parmar et al. (2015)[8] compared multiple classifiers and feature selection methods (i.e. *filter*-based feature selection) on *radiomics* data using the *caret* package. We build upon this work and extend the analysis to 12 datasets outside the *radiomics* domain. We omit *filter* methods because all classifiers in our study comprise built-in feature selection methods (i.e. *embedded* feature selection) and the main advantage of *filter* methods, i.e. low computational cost per feature, is not relevant for our datasets with only modest numbers of features.

Material and Methods

Data collection

Twelve datasets (~~3496~~ 3484 patients) with treatment outcomes described in previous studies were collected from public repositories (www.cancerdata.org) or provided by collaborators. Table 1 characterizes these datasets. Given availability, some datasets consist of subsamples of or contain fewer/more patients and/or features than the cohorts described in the original studies. Two datasets were excluded after a preliminary analysis (these datasets are also not mentioned in table 1) where none of the studied classifiers resulted in an average AUC above 0.51, which is evidence that they contain no discriminative power. Datasets without discriminative power are not suitable for this analysis as we would be unable to determine differences in discriminative performance across classifiers. The patient cohorts of 2 datasets, Wijsman et al. (2015 and 2017), partially overlap but each dataset lists a different outcome (esophagitis and pneumonitis). Datasets were anonymized in the analysis because their identity is not relevant for interpreting the results and to encourage investigators to share their datasets.

Non-binary outcomes were dichotomized, e.g., overall survival was translated into 2-year overall survival in the dataset of Carvalho et al. (2016). Missing data was imputed for training and test sets (the splitting of datasets into training and test sets is described in section *Experimental Design*) by medians for continuous features and modes for categorical features based on the training set. Basing the imputation on the training set avoids information leakage from test to training sets. Categorical features in training and test sets were dummy coded, i.e. representing categorical features as a combination of binary features, based on the combined set for classifiers that cannot handle categorical features (see table 2). Dummy coding on the combined set ensures that the coding represents all values observed in a dataset. Features with zero variance in training sets were deleted in the training set and in the corresponding test set. Additionally, we removed near-zero variance features for *glmnet* to avoid the classifier implementation from crashing during the fitting process. Features in training sets were rescaled to the interval [0,1] and the same transformation was applied to the corresponding test sets. Rescaling is needed for certain classifiers, e.g., *svmRadial*. All these operations (imputation, dummy coding, deleting (near-)zero variance features, rescaling) were performed independently for each pair of training and test sets (step 2 in figure 1).

Classifiers

Six common classifiers were selected and their implementations were used via their interfacing with the open-source *R* package *caret*. The selection includes classifiers frequently used in medical data analysis and advanced classifiers such as random forests or neural networks.

- Elastic net logistic regression is a regularized form of logistic regression, which models additive linear effects. The added shrinkage regularization (i.e. feature selection) makes it is suitable for datasets with many features while maintaining the interpretability of a standard logistic regression.
- Random forests generate a large number of decision trees based on random subsamples of the training set while also randomly varying the features used in the trees. Random forests allow modelling non-linear effects. A random forest model is an ensemble of many decision tree models and is therefore difficult to interpret.
- Single-hidden-layer neural networks are simple versions of multi-layer perceptron neural network models, which are currently popularized by deep neural network applications in machine learning. In the hidden layer, auxiliary features are generated from the input features which are then used for classification. The weights used to generate auxiliary features are derived from the training set. The high number of weights require more training data than other simpler algorithms and reduce interpretability. However, if sufficient data is available, complex relationships between features can be modelled.

- Support vector machines with a radial basis function (RBF) kernel transform the original feature space to attain a better separation between classes. This transformation, however, is less intuitive than linear SVMs where a separating hyperplane is in the original feature space.
- LogitBoost (if used with decision stumps as in this paper) learns a linear combination of multiple single feature classifiers. Training samples that are misclassified in early iterations of the algorithm are given a higher weight when determining further classifiers. The final model is a weighted sum of single feature classifiers. Similar to random forests, it builds an ensemble of models which is difficult to interpret.
- A decision tree iteratively subdivides the training set by selecting feature cutoffs. Decision trees can model non-linear effects and are easily interpretable as long as the tree depth is low.

Classifier details can be found in general machine learning textbooks[23,24]. Table 2 further characterizes these classifiers. We use the option in *caret* to return class probabilities for all classifiers, including non-probabilistic classifiers like *svmRadial*. Classifier hyperparameters, i.e. model-intrinsic parameters that need to be adjusted to the studied data prior to modelling, were tuned for each classifier using a random search: 25 randomly chosen points in the hyperparameter space are evaluated and the point with the best performance metric (we chose the AUC in this study) is selected. The boundaries of the hyperparameter space are given in *caret*.

Experimental Design

For each classifier, test set (or *out-of-sample*) performance metrics (AUC, Brier score, accuracy, and Cohen's kappa) were estimated for each of the 12 datasets. The performance metric estimator was the average performance metric computed from the outer test folds in a nested and stratified 5-fold cross-validation (CV). The experiment was repeated 100 times. The 100 times repeated nested cross-validation yields a better estimate of the true test set performance by randomly simulating many scenarios with varying training and test set compositions.

The experimental design is depicted in figure 1: Each dataset was split into 5 random subsamples stratified for outcome classes (step 1 in figure 1), each of them acting once as a test set and 4 times as a part of a training set. The number of inner and outer folds was set to 5 following standard practice[24(p242)]. Data pre-processing is done per pair of training and test sets (step 2; see details in section *Datasets*). The models were trained on the training set (step 6) and applied on the test set (step 7) to compute the performance metrics for the test set (step 8) resulting in 5 estimates per performance metric (i.e. 1 per outer fold). During the training in each outer fold, the best tuning parameters were selected from the random search (see section *Classifiers*) according to the maximum AUC of an inner 5-fold CV. In the inner CV, the training set was again split into 5 subsamples and models with different tuning parameters were compared (steps 3-5). The nested 5-fold CV was repeated 100 times with different randomization seeds which are used, e.g., for generating the outer folds in step 1. Note that the performance metrics computed on the outer test folds of any two classifiers can be analysed by pairwise comparison because the classifiers were trained (step 6) and tested (step 7) on the same training and test sets for a specific dataset within each of the 100 repetitions.

The mean AUC, Brier score, accuracy, and Cohen's kappa were computed from the 5 estimates of the 5 folds in the outer CV. Calibration intercept and slope were computed from a linear regression of outcomes and predicted outcome probabilities for each of the 5 outer folds. To attain aggregated calibration metrics over the 5 outer folds of the CV, the mean absolute differences from 0 and 1 were computed for the calibration intercept and slope, respectively. Classifier rankings were computed per dataset and repetition by ordering the classifiers' CV-mean AUC (i.e. the average AUC for 5 test sets) in descending order and then assigning the ranks from 1 to 6. Using CV-mean AUCs and CV-mean AUC *ranks*, we answer research questions 1 & 2. We chose AUC for the analysis following Steyerberg et al. (2010)[31]. They emphasize the importance of discrimination and calibration metrics when assessing prediction models. For the simplicity, we restricted the extended analysis to discrimination (AUC) but also report results for calibration and other metrics in appendix A.

To address the question of pre-selection (research question 3), we assess the advantage of choosing a classifier based on performance metrics from similar datasets, which we call *pre-selection* below. To estimate the benefit of our classifier pre-selection for a new dataset and to compare it to alternative strategies, the results of the experiment above were used as input for a simulation. For each outer fold of the 1200 5-fold CVs (12 datasets * 100 repetitions * 5 folds = 6000 folds), 3 classifier selections were made and tested on the test set that belongs to the specific outer fold:

- pre-selecting the classifier according to the average AUC *rank* in all other datasets (excluding all folds from the current dataset),
- selecting the classifier that performed best in the inner CV on the training set,
- randomly selecting a classifier.

Pre-selecting the classifier for one dataset that had the best average AUC *rank* in the other datasets simulates the scenario in which an investigator bases their classifier choice on empirical evidence as is reported in this manuscript. Randomly selecting a classifier represents the case where an investigator chooses a classifier without any prior knowledge about the dataset that (s)he is about to analyze. Selecting the tuned classifier with best inner CV performance corresponds to evaluating multiple classifiers on the training dataset and thus including dataset-specific information in the classifier selection. The performance metrics are averaged over all 500 outer folds (5 folds * 100 repetitions) for each of the 12 datasets.

The documented *R* code used for the analysis is available online[1].

Results

Running 1 nested 5-fold cross-validation and computing the metrics on 1 dataset for all 6 classifiers allows 1 comparison of classifiers. This was applied on 12 different datasets, with each run repeated 100 times for a total of 1200 comparisons. The total computation time was approximately 6 days on an Intel Core i5-6200U CPU (or 15 seconds per classifier per dataset per outer fold, on average).

The results are presented and discussed threefold:

(1) results aggregated over all datasets and repetitions to determine the presence of a superior classifier,

(2) separate results for each dataset but aggregated over repetitions to determine dataset dependency,

(3) a simulation of classifier selection methods in new datasets to estimate the relative effect of classifier pre-selection.

The detailed analysis is restricted to the classifiers' discriminative performance according to the AUC. Results for the remaining metrics (Brier score, calibration intercept/slope, accuracy, and Cohen's kappa) are reported in appendix A.

Results aggregated over all datasets

Figure 2 shows the distribution of classifier rankings based on the average AUC (12 datasets * 100 repetitions = 1200 data points per classifier). Figure 3 depicts pairwise comparisons for each classifier pair (1200 comparisons per pair). The numbers in the plot indicate how often classifier A (y-axis) achieved an AUC greater than classifier B (x-axis). Coloring indicates whether the increased AUCs of classifier A are statistically significant (violet) or not (light violet). Untested pairs are colored grey. The significance cutoff was set to the 0.05-level (one-sided Wilcoxon signed-rank test, Holm-Bonferroni correction for 15 tests).

rf and *glmnet* showed the best median AUC *rank*, followed by *nnet*, *svmRadial*, *LogitBoost*, and *rpart* (figure 2). At the low end of the ranking, *rpart* showed poor discriminative performance. Manual inspection of the *rpart* models showed that *rpart* frequently returns empty decision trees for particular sets (for 34%, ~~19%, 68~~67%, 35%, 58% of all outer folds for *sets D*, ~~*E, G*~~*F*, *K*, *L*, respectively). In pairwise comparisons, *rf* and *glmnet* significantly outperformed all other classifiers (figure 3). *rf* exhibited a small but statistically insignificant better AUC *rank* than *glmnet*.

The results in figures 2 and 3 indicate the existence of a significant classifier ranking for these datasets. However, the considerable spread per classifier in figure 2 and the low pairwise comparison percentages (between 57% and ~~91~~88% in figure 3) also suggest a yet unobserved dependency for classifier performance. To this end, the relationship between datasets and varying classifier performance is investigated.

Results separate for each dataset

Figure 4 shows the average AUC for each pair of classifier and dataset (100 repetitions = 100 data points per pair). Figure 5 depicts the average *rank* derived from the AUC (100 data points per pair).

rf and *glmnet* generally yielded higher AUC values and AUC *ranks* per dataset (figures 4 & 5). However, this observation is not consistent over all datasets: e.g., *nnet* outperforms *rf* in *sets* ~~*H*~~*G*, *J*, and *K*, and *svmRadial* outperformed *glmnet* in *sets A* and *C*.

The results in the figures 4 and 5 indicate that dataset-specific properties impact the discriminative performance of classifiers. These results challenge our proposition that one can pre-select classifiers for predictive modelling in (chemo)radiotherapy based on representative datasets from the same field.

Effects of empirical classifier pre-selection on discriminative performance

Table 3 lists, for each dataset, the name and average AUCs, i.e. averaged over all 100 repetitions, for random classifier selection, classifier pre-selection, and set-specific classifier selection.

The pre-selection procedure always results in *rf* or *glmnet*. The mean benefit of empirically pre-selecting a classifier is small: the AUC improvement ranges between -0.~~02~~ 01 and 0.~~06~~ 07 with a mean of 0.02. In a pairwise comparison over all datasets (p < 0.05, one-sided Wilcoxon signed-rank test), the AUC values by pre-selection were significantly larger than the AUC values by random selection. The AUC *rank* improves by 0.5~~4~~2 on average. Including dataset-specific information by inner CV yields a mean AUC improvement of 0.02 and improves the *rank*, on average, by 0.65~~6~~. In a pairwise comparison of set-specific and random classifier selection over all datasets (p < 0.05, one-sided Wilcoxon signed-rank test), the AUC increase was also statistically significant.

Given this simulation, the expected benefit of pre-selecting a classifier for a new dataset based on results from (chemo)radiotherapy-specific numerical studies is limited with an average increase in AUC of 0.02.

Discussion

Our results suggest that there is indeed an overall ranking of classifiers in (chemo)radiotherapy datasets, with *rf* and *glmnet* leading the ranking. However, we also observe that the performance of a classifier depends on the specific dataset. Pre-selecting classifiers based on evidence from related datasets would, on average, provide~~s~~ a benefit for investigators because it increases discriminative performance. An increase in average discriminative performance is desirable in that an investigator would be less likely to discard their data because of a perceived absence of predictive or prognostic value. The estimated 0.02 mean AUC improvement might appear small but it comes ‘for free’ with classifier selection based on empirical evidence from multiple radiotherapy datasets. Furthermore, the 0.02 AUC improvement is relative to random classifier selection. If an investigator had initially chosen *rpart*, which is the overall worst performing classifier in our study, switching to the preselected classifier would result in an average AUC increase of 0.07. Switching from LogitBoost, which is the second worst performing classifier in our study, to the preselected classifier would result in an average AUC increase of 0.04.

The results in table 3 show that classifier pre-selection and set-specific classifier selection, on average, yield the same AUC increase. We think that the usefulness of set-specific classifier selection is dependent on the size of the training set: classifier pre-selection is preferable for small datasets, set-specific classifier selection is better for larger datasets. Classifier pre-selection represents choosing classifiers using evidence from a large collection of similar datasets from the general radiotherapy outcome domain. Set-specific classifier selection represents choosing classifiers based on the training set, which is a considerably smaller evidence base but comes from the patient group under investigation. If the training dataset is too small, selecting classifiers based on results from other datasets might be less-error prone. On the contrary, if an investigator has collected a large dataset, they have the option to conduct set-specific classifier selection (with all 6 classifiers) for their training data using our documented *R* code[1].

In table 3, one can observe that the pre-selected classifier is mostly *rf* and sometimes *glmnet*. To understand this behaviour, consider dataset *A*: *glmnet* was pre-selected for *set A* by selecting the classifier with the best average AUC *rank* in all other sets (excluding *set A*). Note that, for all 12 datasets together, the average AUC *rank* for *rf* is only slightly better than for *glmnet* (2.29~~8~~ for *rf* and 2.~~43~~ 45 for *glmnet*; the average of the rows in figure 5). Since *glmnet* performs badly while *rf* performs best in *set A*, excluding this information leads to a better average AUC *rank* for *glmnet* and a worse average AUC *rank* for *rf* in the remaining 11 datasets. As a consequence, *glmnet* becomes the pre-selected classifier for this dataset. A similar behaviour is observed for *set*~~*s*~~ ~~*C*~~ *I* ~~and *E*~~ but not in *sets* *C, D, E, H,* ~~*D, F, I,*~~ where *glmnet* also performs worse than *rf* but the difference between both classifiers is smaller and does not induce a switch in the pre-selected classifier.

The result that classifier pre-selection is as good as set-specific selection in the studied datasets does *not* imply that one *cannot* determine a better classifier for a new dataset. Our implementation of set-specific classifier selection only evaluates the performance of various classifiers but does not directly take into account properties of the dataset itself. For example, if an investigator collected a dataset in which the outcome has a quadratic dependency on a feature, *glmnet* would not be able to capture this relation (since it models only linear effects) but *rf* would. However, pre-selecting a classifier based on results from other (chemo)radiotherapy datasets works well on average. Furthermore, including set-specific classifier selection complicates the modelling process and therefore might not be desirable.

In this study, we collected 12 datasets for different treatment sites, i.e. (non-) small cell lung cancer, head and neck cancer, meningioma with different outcomes, i.e. survival, pneumonitis, esophagitis, odynophagia, regional control. However, this collection is certainly not a complete representation of treatment outcome datasets analyzed in the field of radiotherapy. Furthermore, we only studied one implementation of classifiers while classifier performance may vary between implementations. Past studies, however, indicate that classifier implementations in *R* interfaced with *caret* are competitive[5]. Given the apparent lack of comparative classifier studies in radiotherapy, our intention has been to provide numerical evidence for classifier selection to investigators even though our analysis is not exhaustive.

We intentionally limited the analysis to classifier selection while ignoring factors such as the investigator's experience, usage in literature, hypothetical feature dependencies, and model interpretability. This restriction imitates the current trend to delegate modelling decisions to machine learning algorithms and/or non-domain experts. Nonetheless, we feel the need to emphasize that including these factors has merit. Furthermore, expertise on a specific classifier could warrant its selection: Lavesson and Davidsson (2006)[32] observed in a study on 8 datasets from different research domains that the impact of hyperparameter tuning exceeds that of classifier selection. Therefore, the investigator could tune a classifier for better performance by also tuning the hyperparameters outside the subset of hyperparameters tuneable inside *caret*. Even in those cases, however, we suggest comparing these results to simpler implementations of *rf* and *glmnet* as these classifiers on average have the best discriminative performance according to this study.

Finally, for the clinical implementation of classifiers, model interpretability is arguably a major requirement[33]: this view is also convincingly motivated by Caruana et al.[34]. Fortunately, our study shows that *glmnet*, which is an intuitive classifier, is also one of the best performing classifiers.

Conclusion

We have modelled treatment outcomes in 12 datasets using 6 different classifier implementations in the popular open-source software *R* interfaced with the package *caret*. Our results provide evidence that the easily interpretable elastic net logistic regression and the complex random forest classifiers generally yield higher discriminative performance in (chemo)radiotherapy outcome and toxicity prediction than the other classifiers. Thus, one of these two classifiers should be the first choice for investigators to build classification models or to compare one’s own modelling results. Our results also show that an informed pre-selection of classifiers based on existing datasets improves discrimination over random selection.

Disclosure of Conflicts of Interest

Andre Dekker, Johan van Soest, Tim Lustberg are founders and shareholders of Medical Data Works B.V., which provides consulting on medical data collection and analysis projects. Cary Oberije is CEO of ptTheragnostic B.V. Philippe Lambin is member of the advisory board of ptTheragnostic B.V.

Acknowledgements

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno) and the QuIC-ConCePT project, which is partly funded by EFPI A companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. This research is also supported by the Dutch technology Foundation STW (grant n° 10696 DuCAT & n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from the EU 7th framework program (ARTFORCE - n° 257144, REQUITE - n° 601826), SME Phase 2 (RAIL - n°673780), EUROSTARS (SeDI, CloudAtlas, DART, DECIDE), the European Program H2020 (BD2Decide - PHC30-689715, ImmunoSABR - n° 733008, PREDICT - ITN - n° 766276, CLEARLY - TRANSCAN-FP-045), Interreg V-A Euregio Meuse-Rhine ("Euradiomics"), Kankeronderzoekfonds Limburg from the Health Foundation Limburg, Alpe d'HuZes-KWF (DESIGN), the Zuyderland-MAASTRO grant and the Dutch Cancer Society, KWF- TraIT2HealthRI, Province Limburg-LIME-Personal Health Train, NFU-Data4LifeSciences, Varian Medical Systems-SAGE & ROO.

References

1. Deist TM, Dankers FJWM, Valdes G, et al. Code for: Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. https://github.com/timodeist/classifier_selection_code.

2. Lambin P, van Stiphout RGPM, Starmans MHW, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10(1):27-40. doi:10.1038/nrclinonc.2012.196

3. Lambin P, Roelofs E, Reymen B, et al. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol*. 2013;109(1):159-164. doi:10.1016/j.radonc.2013.07.007

4. Kuhn M, Wing J, Weston S, et al. *Caret: Classification and Regression Training*.; 2016. https://CRAN.R-project.org/package=caret.

5. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J Mach Learn Res*. 2014;15:3133-3181.

6. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. *ArXiv160600930 Cs*. June 2016. http://arxiv.org/abs/1606.00930. Accessed April 8, 2017.

7. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. In: *Biocomputing 2018*. WORLD SCIENTIFIC; 2017:192-203. doi:10.1142/9789813235533_0018

8. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*. 2015;5:13087. doi:10.1038/srep13087

9. Belderbos J, Heemsbergen W, Hoogeman M, Pengel K, Rossi M, Lebesque J. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. *Radiother Oncol*. 2005;75(2):157-164. doi:10.1016/j.radonc.2005.03.021

10. Bots WTC, van den Bosch S, Zwijnenburg EM, et al. Reirradiation of head and neck cancer: Long-term disease control and toxicity. *Head Neck*. 2017;39(6):1122-1130. doi:10.1002/hed.24733

11. Carvalho S, Troost EGC, Bons J, Menheere P, Lambin P, Oberije C. Prognostic value of blood-biomarkers related to hypoxia, inflammation, immune response and tumour load in non-small cell lung cancer – A survival model with external validation. *Radiother Oncol*. 2016;119(3):487-494. doi:10.1016/j.radonc.2016.04.024

12. Carvalho S, Troost E, Bons J, Menheere P, Lambin P, Oberije C. Data from: Prognostic value of blood-biomarkers related to hypoxia, inflammation, immune response and tumour load in non-small cell lung cancer – a survival model with external validation. http://doi.org/10.17195/candat.2016.04.1. Published 2016.

13. Janssens GO, Rademakers SE, Terhaard CH, et al. Accelerated Radiotherapy With Carbogen and Nicotinamide for Laryngeal Cancer: Results of a Phase III Randomized Trial. *J Clin Oncol*. 2012;30(15):1777-1783. doi:10.1200/JCO.2011.35.9315

14. Jochems A, Deist TM, El Naqa I, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. *Int J Radiat Oncol*. 2017;99(2):344-352. doi:10.1016/j.ijrobp.2017.04.021

15. Kwint M, Uyterlinde W, Nijkamp J, et al. Acute Esophagus Toxicity in Lung Cancer Patients After Intensity Modulated Radiation Therapy and Concurrent Chemotherapy. *Int J Radiat Oncol • Biol • Phys*. 2012;84(2):e223-e228. doi:10.1016/j.ijrobp.2012.03.027

16. Egelmeer AGTM, Velazquez ER, Jong JMA de, et al. Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: A cohort study based on 994 patients. *Radiother Oncol*. 2011;100(1):108-115. doi:10.1016/j.radonc.2011.06.023

17. Lustberg T, Bailey M, Thwaites DI, et al. Implementation of a rapid learning platform: Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting. *Oncotarget*. 2016;7(24):37288-37296. doi:10.18632/oncotarget.8755

18. Oberije C, De Ruysscher D, Houben R, et al. A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. *Int J Radiat Oncol Biol Phys*. 2015;92(4):935-944. doi:10.1016/j.ijrobp.2015.02.048

19. Oberije C, De Ruysscher D, Houben R, et al. Data from: A validated prediction model for overall survival from Stage III Non Small Cell Lung Cancer: towards survival prediction for individual patients. 2015. https://www.cancerdata.org/id/10.5072/candat.2015.02.

20. Olling K, Nyeng DW, Wee L. Predicting acute odynophagia during lung cancer radiotherapy using observations derived from patient-centred nursing care. *Tech Innov Patient Support Radiat Oncol*. 2018;5:16-20. doi:10.1016/j.tipsro.2018.01.002

21. Wijsman R, Dankers F, Troost EGC, et al. Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-)radiotherapy. *Radiother Oncol*. 2015;117(1):49-54. doi:10.1016/j.radonc.2015.08.010

22. Wijsman R, Dankers F, Troost EGC, et al. Inclusion of incidental radiation dose to the cardiac atria and ventricles does not improve the prediction of radiation pneumonitis in advanced stage non-small cell lung cancer patients treated with intensity-modulated radiation therapy. *Int J Radiat Oncol*. doi:10.1016/j.ijrobp.2017.04.011

23. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer-Verlag; 2013. //www.springer.com/gp/book/9781461471370. Accessed March 4, 2018.

24. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. New York: Springer-Verlag; 2009. //www.springer.com/gp/book/9780387848570. Accessed March 4, 2018.

25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1–22.

26. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22.

27. Venables WN, Ripley BD. *Modern Applied Statistics with S*. Fourth. New York: Springer; 2002. http://www.stats.ox.ac.uk/pub/MASS4.

28. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 Package for Kernel Methods in R. *J Stat Softw*. 2004;11(9):1–20.

29. Tuszynski J. *CaTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, Etc.*; 2014. https://CRAN.R-project.org/package=caTools.

30. Therneau T, Atkinson B, Ripley B. *Rpart: Recursive Partitioning and Regression Trees*.; 2017. https://CRAN.R-project.org/package=rpart.

31. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiol Camb Mass*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

32. Lavesson N, Davidsson P. Quantifying the Impact of Learning Algorithm Parameter Tuning. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. Boston, Massachusetts: AAAI Press; 2006:395–400. http://dl.acm.org/citation.cfm?id=1597538.1597602. Accessed April 9, 2017.

33. Valdes G, Luna JM, Eaton E, Ii CBS, Ungar LH, Solberg TD. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Sci Rep*. 2016;6:37854. doi:10.1038/srep37854

34. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. New York, NY, USA: ACM; 2015:1721–1730. doi:10.1145/2783258.2788613

Appendix A

Table A1 lists performance metrics per classifier. These values are averaged over all repetitions and datasets (100 repetitions * 12 datasets = 1200 data points each). Accuracy and Cohen’s kappa were computed at the 0.5-cutoff. Calibration fails in some outer folds for every classifier resulting in either large or undefined values for intercept and/or slope. This failure occurs frequently with *nnet* and *rpart*. Undefined (NaN) values are excluded when calculating the median.

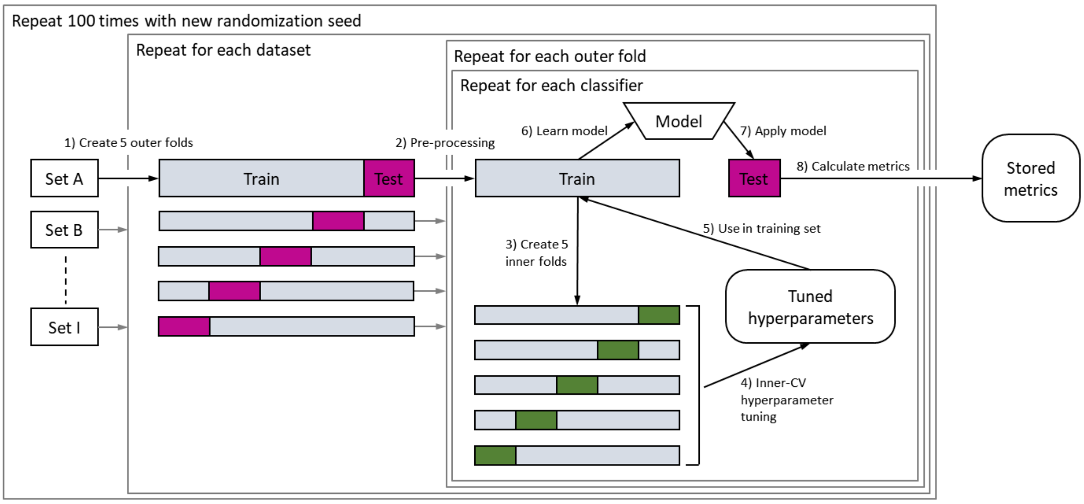

Figure 1. Experimental design: each dataset is split into 5 stratified outer folds (step 1). For each of the folds, the data is pre-processed (imputation, dummy coding, deleting zero variance features, rescaling) (step 2). The hyperparameters are tuned in the training set via a 5-fold inner CV (steps 3-5). Based on the selected hyperparameters, a model is learned on the training set (step 6) and applied on the test set (step 7). Performance metrics are calculated on the test set (step 8) and stored for all outer folds. This process is repeated 100 times for each classifier. Randomization seeds are stable across classifiers within a repetition to allow pairwise comparison.

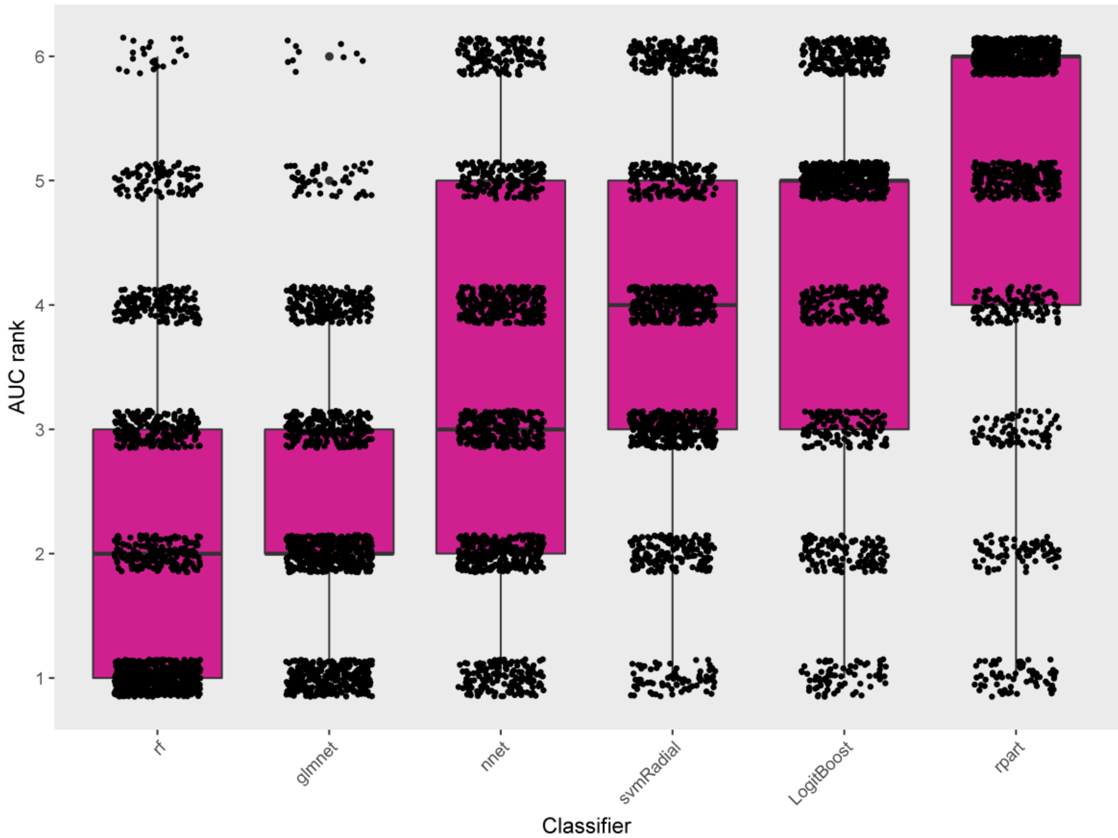

Figure 2. Box- and scatterplot of the AUC *rank* (lower being better) per outer 5-fold CV aggregated over all datasets and repetitions (12 datasets * 100 repetitions = 1200 data points per classifier).

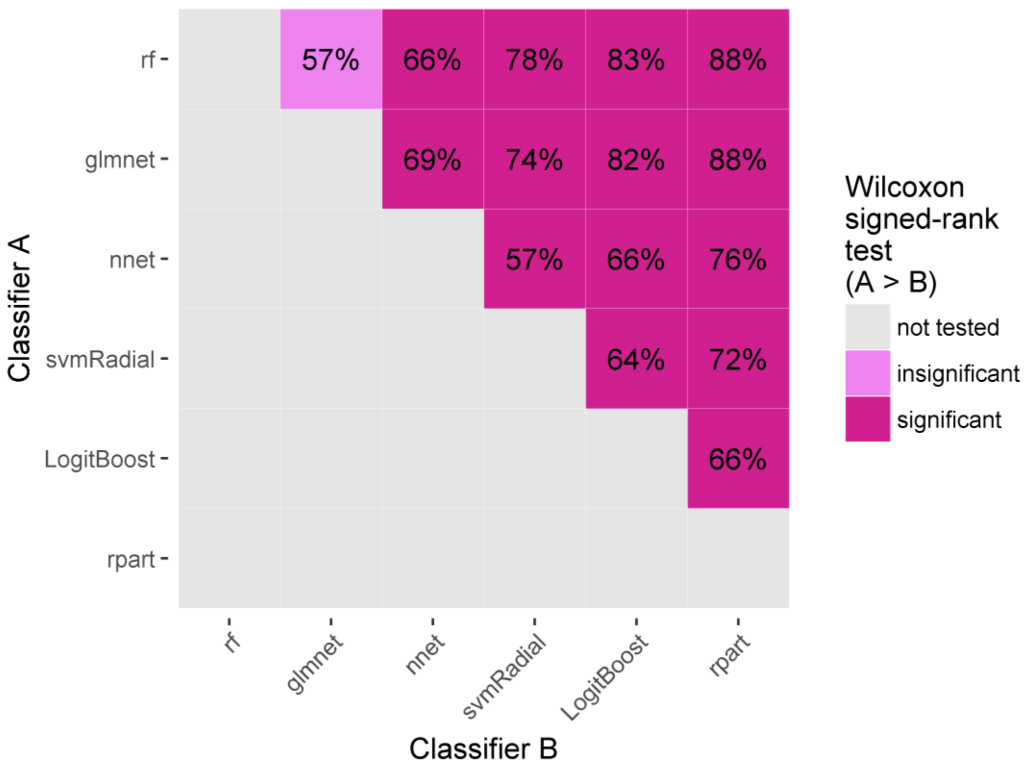

Figure 3. Pairwise comparisons of each classifier pair (12 datasets * 100 repetitions = 1200 comparisons per pair). The numbers in the plot indicate how often classifier A (y-axis) achieved an AUC greater than classifier B (x-axis). The color indicates whether the increased AUCs by classifier A are statistically significant (violet), insignificant (light violet), or have not been tested (grey). The significance cutoff was set to the 0.05-level (one-sided Wilcoxon signed-rank test, Holm-Bonferroni correction for 15 tests).

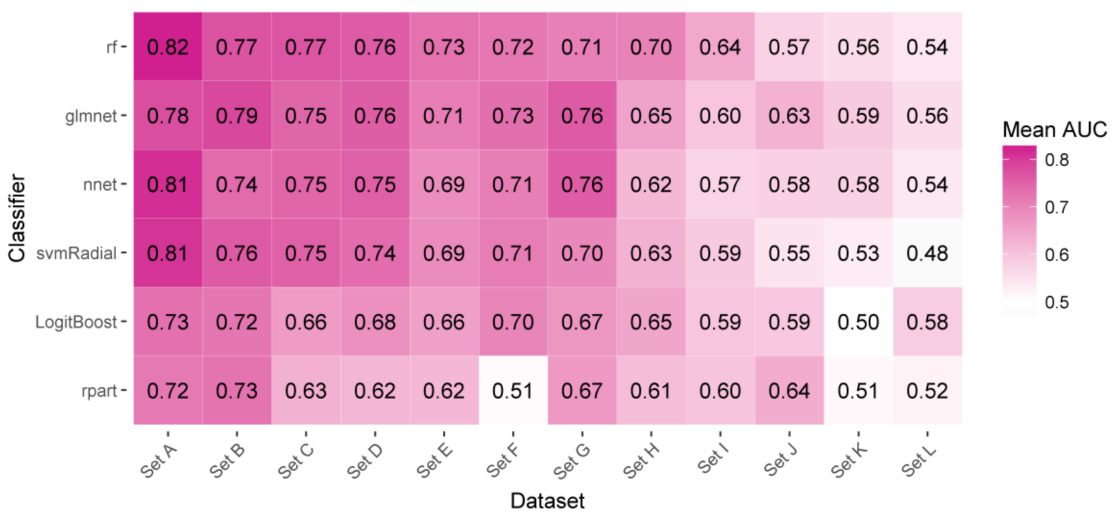

Figure 4. The mean AUC for each pair of classifier and dataset (100 repetitions = 100 data points per pair).

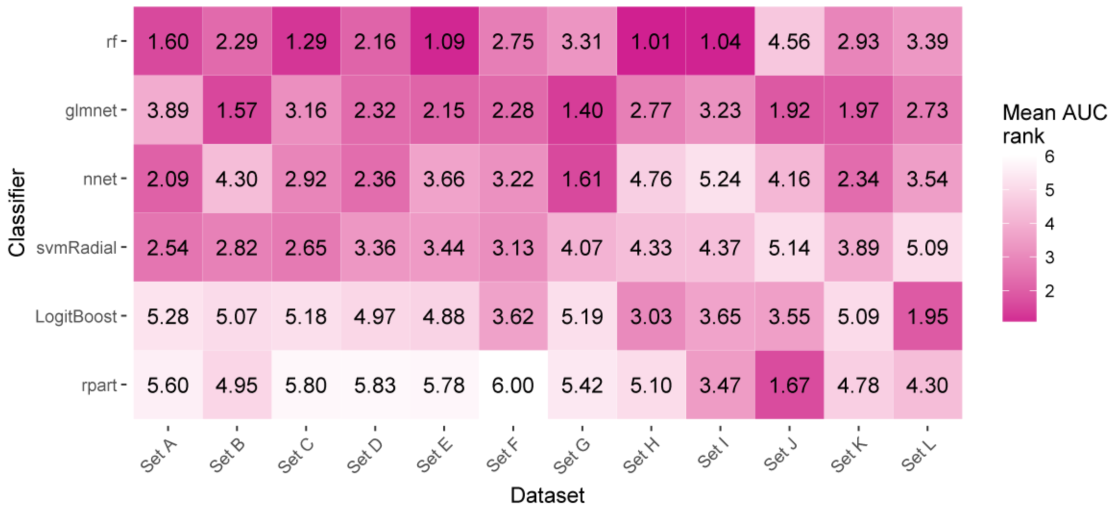

Figure 5. The mean *rank* derived from the AUC (100 repetitions = 100 data points per pair).

Table 1. Dataset characteristics. The number of features is determined before pre-processing.

Dataset	**Disease**	**Outcome**	**Prevalence (in %)**	**Patients**	**Features**	**Feature types**	**Source**
Belderbos et al. (2005)[9]	Non-small cell lung cancer	Grade ≥2 acute esophagitis	27	156	22	Clinical, dosimetric, blood	Private
Bots et al. (2017)[10]	Head and neck cancer	2-year overall survival	42	137	10	Clinical, dosimetric	Private
Carvalho et al. (2016)[11]	Non-small cell lung cancer	2-year overall survival	40	363	18	Clinical, dosimetric, blood	Public[12]
Janssens et al. (2012)[13]	Laryngeal cancer	5-year regional control	89	179	48	Clinical, dosimetric, blood	Private
Jochems et al. (2016)[14]	Non-small cell lung cancer	2-year overall survival	36	327	9	Clinical, dosimetric	Private
Kwint et al. (2012)[15]	Non-small cell lung cancer	Grade ≥2 acute esophagitis	61	139	83	Clinical, dosimetric, blood	Private

Lustberg et al. (2016)[16,17]	Laryngeal cancer	2-year overall survival	83	922	7	Clinical, dosimetric, blood	Private
Morin et al. (forthcoming)	Meningioma	Local failure	36	257	18	Clinical	Private
Oberije et al. (2015)[18]	Non-small cell lung cancer	2-year overall survival	36~~17~~	~~548~~536	20	Clinical, dosimetric	Public[19]
Olling et al. (2017)[20]	Small and non-small cell lung cancer	Odynophagia prescription medication	67	131	47	Clinical, dosimetric	Private
Wijsman et al. (2015)[21]	Non-small cell lung cancer	Grade ≥2 acute esophagitis	36	149	11	Clinical, dosimetric, blood	Private
Wijsman et al. (2017)[22]	Non-small cell lung cancer	Grade ≥3 radiation pneumonitis	14	188	18	Clinical, dosimetric, blood	Private

Table 2. Classifier characteristics.

Classifier	***caret*[4] label**	***R* package**	**Requires dummy coding**	**Tuned hyper-parameters**
Elastic net logistic regression	*glmnet*	*glmnet*[25]	Yes	α, λ
Random forest	*rf*	*randomForest*[26]	No	$mtry$
Single-hidden-layer neural network	*nnet*	*nnet*[27]	No	$size, decay$
Support vector machine with radial basis function (RBF) kernel	*svmRadial*	*kernlab*[28]	Yes	σ, C
LogitBoost	*LogitBoost*	*caTools*[29]	Yes	$nIter$
Decision tree	*rpart*	*rpart*[30]	No	cp

Table 3. For each dataset, the AUC *rank* averaged over all repetitions when (a) randomly selecting a classifier (Random classifier), (b) pre-selecting the classifier with the average best AUC *rank* in all other datasets, i.e. without any information about the current dataset (Pre-selected classifier), (c) selecting the classifier that yielded the highest AUC in the inner CV (Set-specific classifier). Improvements in average AUC and average AUC *rank* compared to (a) are reported. The average AUC improvements by pre-selection and set-specific selection were tested for statistical significance ($p < 0.05$, one-sided Wilcoxon signed-rank test) and found to be statistically significant (*). No other statistical tests besides the two aforementioned tests were conducted.

	Random classifier	Pre-selected classifier				Set-specific classifier		
	Rank		Rank		AUC	Rank		AUC
Dataset	**Mean**	**Name**	**Mean**	**Increase**	**Increase**	**Mean**	**Increase**	**Increase**
Set A	3.43	glmnet	3.64	-0.21	0.00	3.10	0.33	0.02
Set B	3.44	rf	2.92	0.52	0.02	3.31	0.13	0.00
Set C	3.49	rf	1.94	1.55	0.05	2.78	0.71	0.03
Set D	3.59	rf	2.60	0.99	0.05	3.31	0.28	0.02
Set E	3.53	rf	1.89	1.63	0.05	2.58	0.94	0.03
Set F	3.57	rf	2.99	0.58	0.04	3.52	0.05	0.01
Set G	3.43	rf	3.81	-0.39	0.00	1.70	1.73	0.05
Set H	3.65	rf	1.59	2.06	0.07	1.71	1.93	0.06
Set I	3.49	glmnet	3.50	0.00	0.00	2.08	1.42	0.03
Set J	3.52	rf	4.18	-0.67	-0.01	3.41	0.11	0.01
Set K	3.59	rf	3.33	0.26	0.02	3.20	0.39	0.02
Set L	3.44	rf	3.50	-0.06	0.00	3.66	-0.22	-0.01
Mean	**3.51**		**2.99**	**0.52**	**0.02***	**2.86**	**0.65**	**0.02***

Table A1. Median performance metrics per classifier aggregated over repetitions and datasets (1200 data points each). Undefined (NaN) values are excluded when calculating the median.

Classifier	AUC	Brier score	Accuracy	Cohen's kappa	Calibration intercept error	Calibration slope error
rf	0.71	0.19	0.70	0.14	0.12	0.38
glmnet	0.71	0.20	0.70	0.14	0.26	0.66
nnet	0.69	0.22	0.67	0.11	0.31	0.87
svmRadial	0.69	0.19	0.70	0.06	0.32	0.82
LogitBoost	0.66	0.24	0.66	0.18	0.24	0.60
rpart	0.62	0.23	0.67	0.17	0.22	0.55