Research Techniques Made Simple: Feature Selection for Biomarker Discovery

Rodrigo Torres[1] and Robert L. Judson-Torres[1,2,3]

Molecular biomarkers can be powerful tools for aiding in the efficiency and precision of clinical decision-making. Feature selection methods, machine-learning, and biostatistics have been applied to discover subsets of molecular markers that identify target classes of clinical cases. For example, in the field of dermatology, these approaches have been used to develop predictive models that identify skin diseases, ranging from melanoma to psoriasis, based upon a variety of biomarkers. However, a continuous increase in the variety and size of datasets from which candidate biomarkers can be derived, and limitations in the computational tools used to analyze them, have hindered the interpretability of biomarker discovery studies. In this article, the various methods of feature selection are described along with the important steps needed to properly validate the performance of the selected methods. Limitations and suggestions toward uses of these methods are discussed.

Journal of Investigative Dermatology (2019) **139,** 2068–2074; doi:10.1016/j.jid.2019.07.682

CME Activity Dates: 19 September 2019
Expiration Date: 18 September 2020
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: All authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationships with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:
- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of Beaumont Health and the Society for Investigative Dermatology. Beaumont Health is accredited by the ACCME to provide continuing medical education for physicians. Beaumont Health designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the Journal of Investigative Dermatology website: http://www.jidonline.org/current. Tests for CME credits may only be submitted online at https://beaumont.cloud-cme.com/RTMS-Oct19 – click 'CME on Demand' and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

INTRODUCTION

Is bigger data always better data?

Biomarkers play an important role in helping to improve early diagnosis of disease and prognosis of treatments. For diseases where early diagnosis greatly improves survival rates, such as melanoma, the identification of improved biomarker-based predictive models that are quantitative, reliable, economic, and easy to interpret would substantially improve patient well-being. Consequently, a myriad of reports connecting patient diagnosis or outcome to quantitative measurements—ranging from gene expression level or methylation status to clinical images analyzed by artificial intelligence—have emerged in recent years (Conway et al., 2019; Esteva et al.,2017; Shen et al., 2018).

[1]*Department of Dermatology, University of California, San Francisco, California, USA;* [2]*Department of Dermatology, University of Utah School of Medicine, Salt Lake City, Utah, USA; and* [3]*Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA*

Correspondence: Robert L. Judson-Torres, Huntsman Cancer Institute, 2000 Circle of Hope Drive, Room 2713, Salt Lake City, Utah 84112. E-mail: judsontorreslab@gmail.com

Abbreviation: miRNA, microRNA

SUMMARY POINTS

- Feature selection can be used to help with biomarker studies by providing more interpretable and possibly more relevant targets that could improve classification performance.
- There are many feature selection methods that can be used, each with unique benefits and weaknesses to be considered varying from balancing speed and simplicity with performance and complexity of interactions used.
- Regardless of how features are selected, proper validation is needed to avoid overfitting and to obtain the best real-world estimates of the performance of the model.

LIMITATIONS

This review provides an introduction to the use of feature selection for biomarker discovery, intended to aid in the assessment of published reports, but does not capture all the complexities involved in applying the methods. It is important to be cautious of overfitting when conducting any form of feature selection as overoptimistic estimates of performance can easily be obtained.

Given the number and size of datasets, and the range of advanced analytical methods available, it can be difficult to assess the degree of accuracy and generalizability of each individual study. The purpose of this review is to provide a practical guide for the assessment of biomarker discovery using feature selection for researchers and clinicians who are not specialized in data science.

In the age of big data, biomedical researchers are often presented with the challenge of "wide" datasets—the consequence of studies where sample size is dwarfed by the number of measured characteristics from each sample. Wide datasets are usually associated with high throughput—omics approaches, such as next generation genomic or transcriptomic sequencing, metabolomics, proteomics, and lipidomics, but can also include other types of datasets, such as clinical images or information from electronic health records. Each measured characteristic for a given sample, whether categorical values (e.g., sex, race) or numerical values (e.g., age, gene expression level), is considered a feature of that sample (Figure 1a). The full set of different features collected from a set of samples is the feature space. When searching for candidate biomarkers, investigators identify features or sets of features from the feature space that are associated with one specific target feature (also called target class) of interest, such as clinical outcome. Predictive models can then be generated by first training a function to best map these features to a target class. The trained model can then be tasked to identify the target class of new samples based only upon the selected features that provided the most accurate mapping. This process of feature selection can involve the application of common statistical tests (with which most researchers are familiar) or more computationally demanding machine-learning classifiers (often considered a "black box"). In all cases, the use of wide datasets for biomarker discovery runs the risk of false discovery via overtraining, and external datasets are required for validation. This review highlights the strengths, weaknesses, and appropriate application and validation of approaches to feature selection.

WHAT IS FEATURE SELECTION?

When dealing with wide datasets, a variety of dimensionality reduction techniques are available, such as feature selection and feature extraction. The purpose of these techniques is to remove irrelevant and redundant features. Feature selection is the process of refining a large set of variables or characteristics to a subset that optimally separates two or more target classes of samples. Feature extraction transforms the features into a smaller feature space by combining them into new features that still represent the full feature set, instead of taking a subset. Whereas both feature selection and extraction serve the goal of optimizing the feature space, feature selection maintains the original identity of the features, which allows for better interpretability and, therefore, more readily accessible targets for a biomarker study. Both methods are powerful tools for identifying the features that best map samples to specific target classes.

As performance of predictive models tends to decrease when too many features are considered (Figure 1b), successful biomarker discovery relies on appropriate feature selection approaches. Feature selection can help improve the performance and utility of a predictive model by eliminating confounding variables, providing simpler models less prone to overfitting, optimizing the efficiency and reducing the cost of future data collection, and highlighting variables that can be examined for causal links to the predicted class (Guyon and Elisseeff, 2003). The fundamental goal of all feature selection methods is to remove features that are, at best, irrelevant or redundant for a predictive model or, at worse, add noise into the model. The process can use a broad range of computational and statistical methods and be automatic through algorithm-based selection or manual through user-selected thresholding.

Features can be ranked based upon a desired criterion and are often combined with various methods of predictive modeling. Strategies for using feature selection in modeling can generally be divided into three subtypes: feature filtering, feature wrappers, and embedded methods (Saeys et al., 2007; Figure 1c). Criteria used to establish feature rank depends on the method used. Feature filtering utilizes traditional univariate statistical methods such as *t* test or correlation to rank features based on relevance using a *P*-value or degree of information gain. An arbitrary user-defined threshold is then often used to select top ranked

a

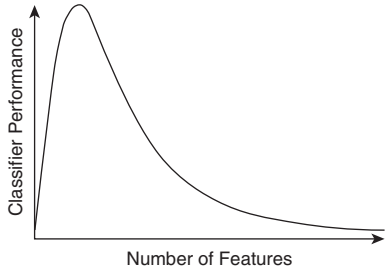
What is a feature?

	Categorical		Numerical		Target
	Sex	Location	Age	Exp	Pathology
1	M	Head	31	5	Melanoma
2	F	Trunk	63	25	Nevus
3	M	Limb	44	10	Melanoma
4	M	Limb	82	400	Nevus
5	F	Trunk	67	35	Nevus

Independent Dependent

b

What are too many features?



c

Feature selection methods

	Classification Independent	Classification Dependent	
	Filter Methods	Wrapper Methods	Embedded Methods
Feature Subset Classification Algorithm	1. Feature Rank 2. Classification model	Feature Search Classification model	Feature Search + Classification model
Advantage	• Fast and simple	• Uses feature dependencies • Developed with classifier performance	• Faster than wrapper • Uses feature dependencies • Combines feature selection and classifier
Disadvantage	• Ignores feature dependencies • Made independent of classifier	• Slowest • Higher risk of overfitting	• Limited choice of classifier
Examples	• t-test • Information Gain • Chi Squared • Correlation	• Backward Selection • Forward Selection • Genetic Algorithm	• LASSO • Decision Tree

Figure 1. Features and feature selection. (**a**) Any measured characteristic or variable can be considered a feature. Within this example dataset, samples contain both categorical features (red), such as the sex of the patient or location of the lesion, and numerical features (blue), such as the age of the patient or the expression level of an RNA biomarker. Together, these features represent the independent variables for the dataset, whereas the target variable (diagnosis from pathology; orange, also a feature) is the dependent variable. (**b**) As the number of features increases, the performance of predictive models increases up to an optimal point. However, the addition of more features with a limited-sized training set will degrade performance. (**c**) Differences between the following three feature selection methods: filter, wrapper, and embedded.

features for use in training a predictive model. Feature filtering methods have the benefit of being fast, simple, and straightforward to interpret. An example of successful feature filtering was reported in a recent article identifying a microRNA (miRNA) signature for cutaneous T-cell lymphoma by comparing the top differentially expressed miRNAs shared between two datasets (Shen et al., 2018). In this report, the authors identify five miRNAs of over 1000 from Agilent and Affymetrix miRNA microarrays, the expression of which predicts a diagnosis of cutaneous T-cell lymphoma with 96% sensitivity and 72% specificity upon validation. However, because each feature is considered independently, feature dependencies are not detected when using this approach. For example, two highly ranked features may provide redundant information, or, in contrast, two poorly ranked interdependent features may become highly predictive only when both are considered (Guyon and Elisseeff, 2003). Feature filtering allows initial estimates of predictive performance, but other feature selection methods often outperform filter methods (Kohavi and John, 1997).

Unlike feature filters, wrappers and embedded methods couple feature ranking and selection directly with the training of a classifier. The types of classifiers used in these approaches are myriad, and their application has been reviewed extensively (Aggarwal, 2014; Maglogiannis, 2007). Some popular examples include linear models, such as logistic regression and support vector machines; decision tree-based models, such as classification and regression trees and random forest; and probabilistic based

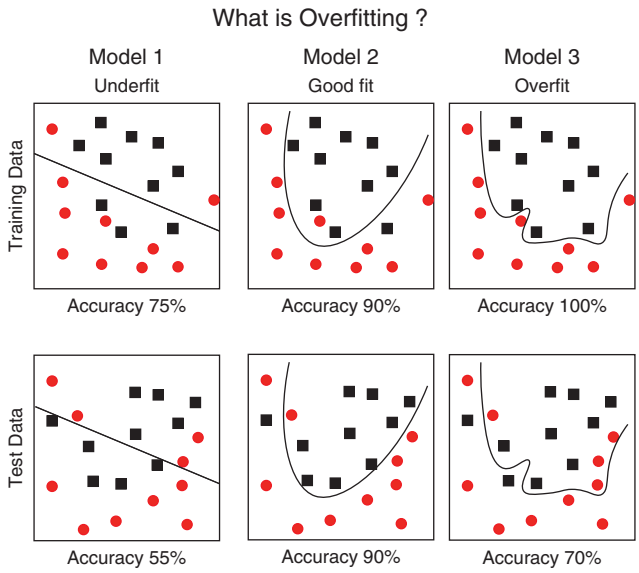


Figure 2. What is overfitting? When conducting feature selection and model building, it is crucial to avoid overfitting. Overfitting occurs when the model is built to best describe the training data but fails to fit any new data. In this example, three models are built on the training data to separate two classes. When evaluating the performance of the models on the test dataset, it is clear that model 1 is underfit (because it does poorly on both sets), model 2 is properly fit (because it has similar performance on both sets), and model 3 is overfit (because the performance is overestimated in the training set). Overfitting can be avoided by generally using simpler models and minimizing the number of input factors in a model. Regardless of the model or feature selection method used, it is important to evaluate the generalization of the model with proper validation.

models, such as naïve bayes. Wrappers and embedded methods rank and select features using the performance of these classification models trained with subsets of features. Thus, features are ultimately selected if they consistently permit optimal performance of a predictive model. Wrappers provide a method to iteratively test feature subsets with any classifier that ranks features based upon a measure of prediction performance, thus identifying the optimal performing subset of features. Considering that an exhaustive search of all subsets of features would be too computationally expensive, search algorithms are implemented, such as backward and/or forward selection and genetic algorithms. This combination of a classifier with a search algorithm constitutes a wrapper method. Wrapper methods evaluate feature dependencies and permit flexibility in the classifier; however, because each iteration in the feature selection creates a new subset to test, it grows exponentially with the number of features, resulting in methods that are both computationally demanding and have a higher chance of overfitting the data (Figure 2).

Embedded methods are classifiers that have a built-in process of feature selection, which means that unlike wrappers, they do not separate the learning from the feature selection steps. These methods include random forest and other decision tree-based models, as well as regularized methods like LASSO. In a recent study, an embedded method was used to identify a 40-CpG classifier for distinguishing primary invasive melanoma from nevi, with sensitivity of 96.6% and specificity of 100% upon validation (Conway et al., 2019). The investigators used a method called elastic net. After each iteration of this method, features that are not important for the model are moderately penalized, such that their potential contribution to the model during the next iteration is reduced. Over many iterations, the contribution of features that are consistently unimportant is reduced to zero and, conversely, the contribution of features that are consistently important is amplified in comparison. By applying this method, 40 methylation sites of 41,448 probes were found to be predictive of a malignant diagnosis, when considered in aggregate. Embedded methods have the benefit of being faster than wrapper methods while still incorporating feature subsetting and feature dependencies into classifier construction; however, they are limited to specific prebuilt classifiers, such that features determined from one classifier might not work with another. Each of these models and documentation for their application are available for investigators trained in the R or python languages. Several recent and excellent reviews provide detailed descriptions of how to apply these methods (Perez-Riverol et al., 2017). A case study, including sample data, code, and instructions, is provided as supplementary materials for this review. An investigator with basic training in R can use these materials to gain an introduction to recursive feature elimination, and the concepts of redundancy and noise in a feature space.

Feature selection can be a powerful method for improving the classification performance of candidate biomarkers (Hemphill et al., 2014). However, there are important limitations that need to be acknowledged when choosing a specific method and evaluating the feature subsets and performance. First, it is important to know the limitations and advantages of each method (Figure 1c). Second, an investigator must appreciate that there is no single feature selection method that is guaranteed to improve performance. Third, scalability can also be a problem especially with very large or small datasets as a small dataset might not properly determine the relevance of features and a large dataset can require too much computation to use some feature selection methods. Finally, when comparing models, it is also critical to understand the concept and role of stability. Stability refers to the robustness of the feature subset selection to differences in the training set sampling. In an unstable method, small differences in the training set sampling can result in completely different sets of selected features, each of which produce equivalent performance. More stable results can be obtained from ensemble methods that have been developed to minimize this problem by combining ranks from multiple tests or using bootstrapping to aggregate results (He and Yu, 2010). In a recent study, we used an ensemble method Boruta (Kursa and Rudnicki, 2010) to find a consistent set of miRNAs for classification of primary melanoma from nevi (Torres et al., in press), addressing the previously described poor stability of miRNA biomarkers for melanoma detection (Jayawardana et al., 2016). With this method, during each iteration, a

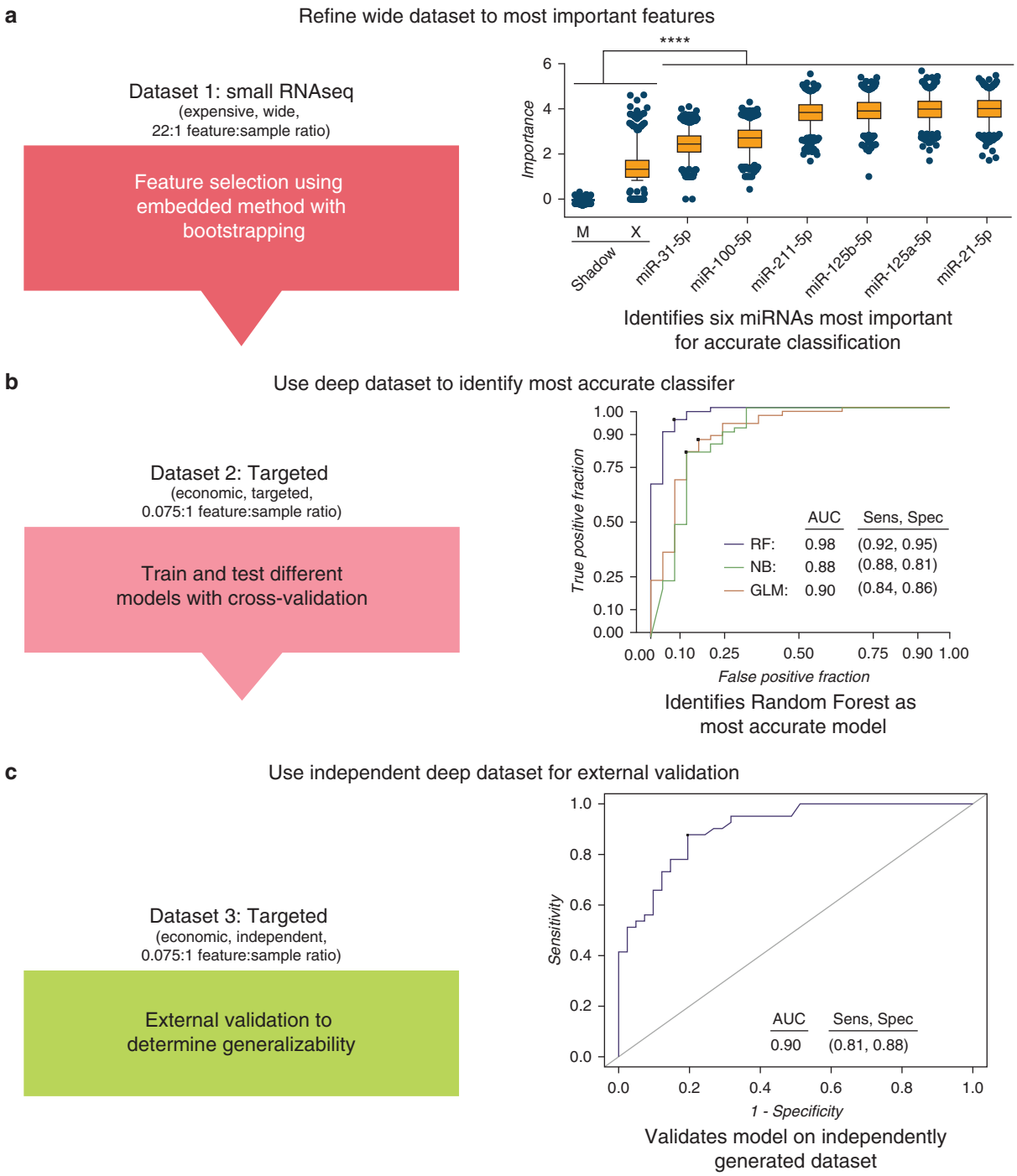


Figure 3. Feature selection workflow. Example of feature selection workflow from a study of miRNAs as biomarkers for melanoma. (**a**) The study generated a wide dataset, with many more features than samples, using next generation sequencing. A feature selection method called Boruta was used to test the importance of each miRNA for accurate classification over 1000 iterations, as compared to randomly generated artificial features (shadow features). Six miRNAs were significantly more important than both the median and the top performing shadow feature. (**b**) The study then considered only the expression of the feature selected miRNAs over a larger sample set. Cross-validation was used to compare the accuracies of several models. ROC curves and AUC for each model are shown. Sensitivity and specificity corresponding to black points are shown. (**c**) A third external dataset was generated to validate the optimally performing random forest model. AUC, area under the curve; GLM, generalized linear model; M, median performing; miRNA, microRNA; NB, naïve bayes; RF, random forest; ROC, receiver operating characteristic; Sens, sensitivity; Spec, specificity; X, top performing.

number of artificial features equal to the number of measured features are generated by randomizing the values of each feature. A model is then trained using the random forest method, and both real features and artificial features (called shadow features) are ranked based upon importance for predictive accuracy. Only those features that consistently score better than all shadow features are retained for further iterations. In doing so, features less likely to add to performance are eliminated and features that stably contribute to performance are retained. In this

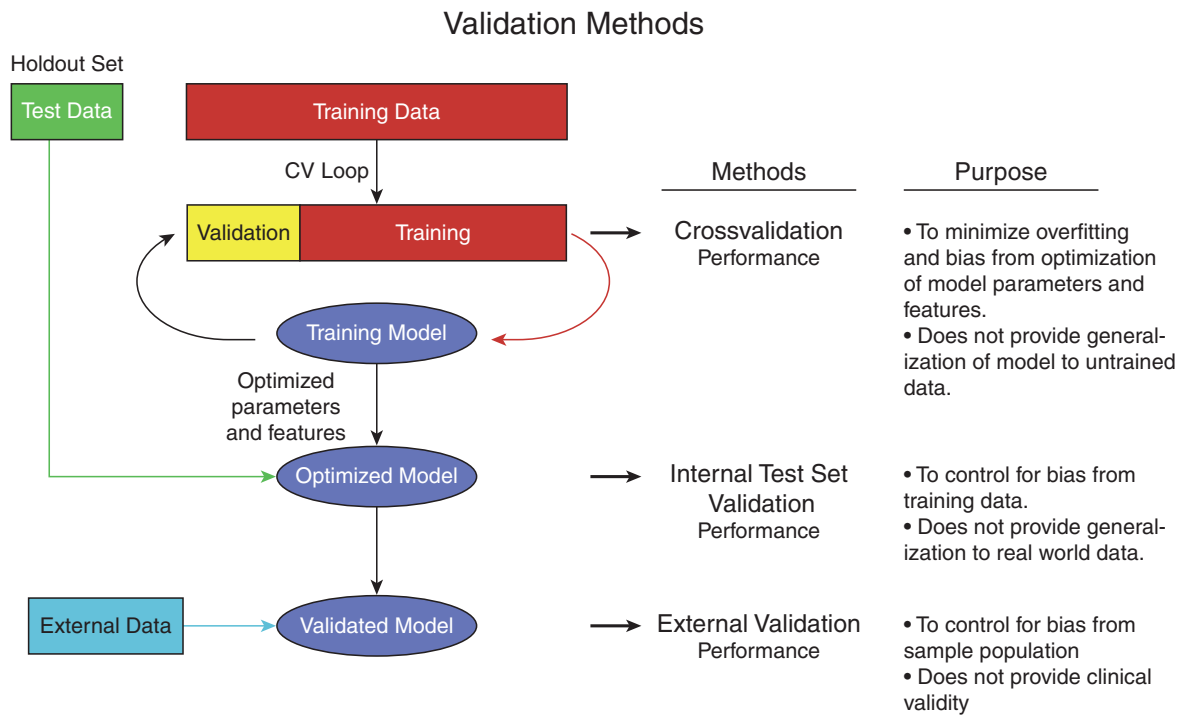


Figure 4. Validation methods. Summary of the important uses of validation methods when building a model regardless if feature selection is used. Cross-validation or holdout set splits can both be used to test performance of model parameters and feature sets, but a separate validation with an external dataset is needed outside of cross-validation to determine the generalizability of the model. CV, cross-validation.

study, the investigators used this approach to first identify six microRNAs with diagnostic value from a wide dataset (Figure 3a). Just these six features were then measured in a much larger cohort and used to identify an optimal classification model using cross-validation to test performance (Figure 3b). Finally, the generalizability of the model was tested using a third external validation cohort (Figure 3c). These types and purposes of validation are discussed in the next section.

VALIDATION APPROACHES IN BIOMARKER DISCOVERY

After wide datasets are refined to just those features that optimally separate samples into classes of interest using a trained model, validation processes are used to determine the generalizability of the model. Standard practice with a sufficiently sized dataset is to split samples into training and test sets (Figure 4). The training set is used to perform feature selection, classifier selection, and ultimately train a predictive model, whereas the test set is held in reserve to validate the trained model. Variations on this method provide non-biased performance data from samples that were not used to train the model. However, since all samples originally came from the same dataset, this method of validation is not a strong indicator of real-world performance. Biases, noise or confounding variables introduced by the method of sample collection, the location or time the samples were collected, the technologies used to measure features, or the researchers collecting the samples are not taken into account when using this approach. Further testing on an external validation set with a completely separate population sample can better estimate real-world performance of the model. If conducted with a large dataset, this external validation can also allow for interrogation of performance that is dependent on specific subgroups and used to reassess the utility of the model in specific clinical settings (Figure 4). It is important to note that these described strategies for validation represent a baseline for candidate biomarker identification for the assessment of discovery-based reports. Further clinical validation requires extensive assay development and subsequent testing at different sites (Voskuil, 2015).

Worth discussion is the practice of cross-validation. Cross-validation involves removing a portion of the training data, optimizing with the remainder, and testing with the withheld data. Unlike the training set/test set division described above, many iterations of data withholding and testing are conducted, and average aggregated performance measurements are used to identify an optimal model. When conducting feature selection, classifier selection, and model training, cross-validation is an excellent strategy for optimizing model parameters with minimal bias and is commonly applied. However, as the process inherently resamples the same data set for both training and testing, cross-validation should not be considered a substitution for the validation strategies described earlier.

SUMMARY

Feature selection can be a powerful method for improving classification performance. When evaluating methods, it is important to note that no one method will always perform better than others. Knowing the limitations of the dataset and using multiple methods can increase the chance of finding an optimal model. Regardless of the feature selection methods used, proper validation prevents false assumptions from the dataset and maximizes the

MULTIPLE CHOICE QUESTIONS

1. Which of the following is NOT a benefit of feature selection?
   A. It can reduce overfitting
   B. It can help remove irrelevant and redundant features
   C. It can transform all features into a smaller feature space
   D. It allows for the selection of more biologically relevant targets associated with target class
2. The advantages of embedded feature selection methods include all of the following, EXCEPT:
   A. Take advantage of feature dependencies
   B. Combine features subsetting and classifier construction
   C. Are more computationally efficient than wrapper methods
   D. Can be used with any classification method
3. Which of the following is NOT a possible cause of overfitting the model?
   A. Using too many model parameters for the size of the dataset
   B. Using cross-validation during model building
   C. Having many irrelevant features
   D. Using less of the data
4. The external validation dataset is used for which of the following?
   A. Validating model parameters optimized during training
   B. Testing the generalizability of a training model on data not used for training
   C. To remove all bias from training data
   D. Testing the generalizability of a model to an outside population
5. Which of the following is an appropriate use of cross-validation?
   A. Validating model parameters optimized during training
   B. Testing the generalizability of a training model on data not used for training
   C. To remove all bias from training data
   D. Testing the generalizability of a model to an outside population

Note: See online version of this article for a detailed explanation of correct answers.

generalization of the model on future data. Researchers, dermatologists, and dermatopathologists interested in assessing the generalizability of reported predictive models should consider the rationales behind each applied method and the types of validation sets tested. The combination of these methods can help ensure the reliability of new diagnostic and prognostic biomarkers for dermatological diseases.

CONFLICT OF INTEREST

The authors state no conflict of interest.

AUTHOR CONTRIBUTIONS

Visualization: RT, RLJT; Writing - Original Draft Preparation: RT; Writing - Review and Editing: RLJT

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Aggarwal CC, editor. Data classification: algorithms and applications. New York: Chapman and Hall/CRC; 2014.

Conway K, Edmiston SN, Parker JS, Kuan PF, Tsai YH, Groben PA, et al. Identification of a robust methylation classifier for cutaneous melanoma diagnosis. J Invest Dermatol 2019;139:1349–61.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

He Z, Yu W. Stable feature selection for biomarker discovery. Comput Biol Chem 2010;34:215–25.

Hemphill E, Lindsay J, Lee C, Măndoiu II, Nelson CE. Feature selection and classifier performance on diverse bio- logical datasets. BMC Bioinformatics 2014;15:S4.

Jayawardana K, Schramm SJ, Tembe V, Mueller S, Thompson JF, Scolyer RA, et al. Identification, review, and systematic cross-validation of microRNA prognostic signatures in metastatic melanoma. J Invest Dermatol 2016;136:245–54.

Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36:1–13.

Maglogiannis IG, editor. Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. Amsterdam, Netherlands: IOS Press; 2007.

Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz M-P, Audain E. Accurate and fast feature selection workflow for high-dimensional omics data. PLOS ONE 2017;12:e0189875.

Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17.

Shen X, Wang B, Li K, Wang L, Zhao X, Xue F, et al. MicroRNA signatures in diagnosis and prognosis of cutaneous T-cell lymphoma. J Invest Dermatol 2018;138:2024–32.

Torres R, Lang UE, Hejna M, Shelton SJ, Joseph NM, Shain AH, et al. MicroRNA ratios distinguish melanomas from nevi. J Invest Dermatol, in press.

Voskuil J. How difficult is the validation of clinical biomarkers? F1000Res 2015;4:101.

DETAILED ANSWERS

1. **Which of the following is NOT a benefit of feature selection?**

Answer: C. It can transform all features into a smaller feature space.

Feature extraction transforms the features into a lower space, whereas feature selection only takes a subset of the features but does not change their representation.

2. **The advantages of embedded feature selection methods include all of the following, EXCEPT.**

Answer: D. Can be used with any classification method.

Only a subset of classifiers has the intrinsic ability to combine feature subsetting and classification to perform embedded feature selection methods, as this function is classifier specific and not usually part of many classification functions.

3. **Which of the following is NOT a possible cause of overfitting the model?**

Answer: B. Using cross-validation during model building.

By using cross-validation during model building the chance of overfitting is reduced because parameters are trained on a broader sampling of the data.

4. **The external validation dataset is used for which of the following?**

Answer: D. Test the generalizability of a model to an outside population.

External validation is best used to test the real-world performance of a model on a population that was not biased by the original feature selection or methods.

5. **Which of the following is an appropriate use of cross-validation?**

Answer: E. Validating model parameters optimized during training.

During cross-validation, model parameters can be optimized by testing the performance of the model over each cross-validation fold with changing hyperparameters to find the best performing set.