# 1 Introduction

Even though multivariate biomarker discovery has already made its debut in the realm of biomedical research, it is poised to become a crucial facet of personalized medicine, which will prompt the demand for a myriad of novel biomarkers representing distinct "omic" biosignatures. So, what is a *multivariate biomarker*? Stating that it is a set of more than one variable (such as genomic variations, gene or protein expression levels, or metabolite concentrations), whose combined pattern of values can be used for predicting the value of a target (or response) variable (such as a disease state, response to treatment, or therapy outcome) would not be sufficient. A very important characteristic of a truly multivariate biomarker is that it has to be identified with the use of *multivariate methods*, that is, methods that evaluate each variable in the context of other variables (and thus, consider correlations and interactions among the variables). That means that a set (of variables) identified by selecting its members via univariate methods – for example, by evaluating the correlation of each variable with the target – should not (and herein, will not) be considered a multivariate biomarker. We are interested in the association between a *set of variables* and the target variable, whereas members of a multivariate biomarker may or may not be individually associated with the target variable.

One may insist that there could exist efficient biomarkers that are composed of several variables, which were selected in a univariate way. That is, of course, possible. However, if a set of univariately-identified variables is used as a biomarker, then it is virtually guaranteed that a better (and truly multivariate) biomarker could have been identified were a multivariate approach used.[1]

Another important characteristic of a properly-identified multivariate biomarker is its parsimony – not only to keep Ockham happy. An optimal multivariate biomarker should consist of as few variables as possible. As will later be discussed, such parsimony is extremely important for biomarkers based on high-dimensional data.[2] For such data, it is

[1] The only exception would be a situation where the variables are independent of each other (i.e., where there are no correlations among them), which is definitely not the case for high-dimensional biomedical data (such as gene or protein expression data).

[2] One may argue that a group of variables (such as genes) that are highly correlated should be included in a biomarker if any one of them is selected into the biomarker. Although such an argument may be supported by the view that including groups of highly correlated variables in a multivariate biomarker (instead of only one of them) would allow for easier biological interpretation of the biomarker, we will see, in Part IV, that there are better approaches, such that will allow both for the identification of parsimonious biomarkers as well as facilitating their biological interpretation.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

quite trivial to identify a multivariate biomarker that would overfit the training data, but much more difficult to find one that would be well generalizable to unseen data. Accordingly, one of the main goals of this book is to explain and show how to identify multivariate biomarkers that represent the real patterns in the target population, rather than the easily identifiable spurious patterns that are present in the training data but not in the target population from which the data were sampled. The principle of parsimony is one of the crucial aspects of this approach and it requires that we apply a combined criterion of minimizing the size of the biomarker (its cardinality) while simultaneously maximizing its predictive power; hence, such optimization will represent a compromise between these two seemingly contradictory criteria. It may appear that increasing the number of variables included in a multivariate biomarker would increase its predictive power; this, however, is a misconception – such a perceived increase in predictive power, when improperly estimated on the training data would, actually, decrease the generalizability of the biomarker.

There are other misconceptions in performing predictive modeling, which, if followed in a biomarker discovery project, would render its results – at the very best – suboptimal. In this book, we will discuss such misconceptions, as well as describe proper methodologies for each and every step of the predictive modeling process for multivariate biomarker discovery based on high-dimensional biomedical data.

## 1.1 Biomarkers and Multivariate Biomarkers

There seems to be some confusion and inconsistency in using terminology and interpreting concepts related to biomarkers and biomarker discovery. This was a reason for the FDA-NIH to decide in 2015 to create "The Biomarkers, EndpointS, and other Tools (BEST) Resource", which is intended to be a dynamic glossary that will be periodically updated and revised. A consequence of this initiative, for example, is that the general definition of a biomarker has evolved from:

> *A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.* (Biomarkers Definitions Working Group 2001),

to that provided by the BEST resource in November 2021:

> *A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions. Biomarkers may include molecular, histologic, radiographic, or physiologic characteristics. A biomarker is not a measure of how an individual feels, functions, or survives.* (FDA-NIH Biomarker Working Group 2021).[3]

[3] www.ncbi.nlm.nih.gov/books/NBK338448/#IX-B

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

Though such initiatives should be applauded, the confusion about terminology related to the emerging *multivariate* biomarker research remains, nevertheless, abundant. It seems quite likely that the main reason for this confusion is the still enduring (and often followed, consciously or not) traditional, historical one-gene-at-a-time (or, more generally, one-variable-at-a-time) approach to discovering and describing biomarkers. However, **a multivariate biomarker is *a biomarker***. It is not a "set of biomarkers", nor a "panel of biomarkers", nor any similar univariately-biased concept.[4] A properly identified multivariate biomarker is a set of variables, which together – as a set – constitute *a single biomarker*. If any of the variables included in the set were a biomarker for the researched goal, there would be no need for the multivariate biomarker. This also means that a proper biological interpretation of a multivariate biomarker should interpret it as a set; interpretation of the individual genes, proteins, or metabolites that compose the biomarker does not translate into the biological interpretation of the set. Hence, a proper multivariate interpretation of biological processes underlying the predictive abilities of a multivariate biomarker is a challenging task and still quite uncharted territory.

Included in the categories of biomarkers listed in the BEST resource are:

- Diagnostic biomarkers
- Prognostic biomarkers
- Predictive biomarkers
- Response biomarkers
- Monitoring biomarkers
- Safety biomarkers
- Risk biomarkers

*Diagnostic biomarkers* are used to recognize – or confirm the presence of – a specific disease, or to support differential diagnosis among subtypes or stages of a disease. *Prognostic biomarkers* estimate the probability of a specific future clinical event, such as relapse, disease progression, or recovery. *Predictive biomarkers* evaluate a patient's sensitivity, as well as the probability of a positive response to specific exposure, such as ionizing radiation or a particular chemotherapy. *Response biomarkers* evaluate a patient's response to a medical product or therapy. *Monitoring biomarkers* are those that are repeatedly measured to assess a patient's status, such as disease progression or response to a treatment, which may be expected to change over a period of time. *Safety*

[4] Such names are appropriate only in situations when independently identified and validated univariate biomarkers are used together to achieve a more comprehensive view of the patient's condition. However, even in such situations, they are still independent single-variable biomarkers that could be used individually or in combinations with other biomarkers or other diagnostic or prognostic evaluations, and they do not constitute a multivariate biomarker. Combinations of such independent biomarkers are sometimes called *composite biomarkers*. On the other hand, multivariate biomarkers, especially the omic ones, have been recently also called *biosignatures* ('omic' biosignatures, molecular signatures), which is a valid alternative description of them.

*biomarkers* are used to evaluate the probability or presence of adverse reactions to a drug or other treatment. *Risk* (or risk profiling) *biomarkers* indicate predisposition, or risk, to develop a specific disease or medical condition.

Although taxonomies of biomarkers may be useful, one may observe that there are overlaps between these biomarker categories, and thus a particular biomarker may be classified to more than one of them. For example, a diagnostic biomarker that assigns a patient to one of the disease subtypes may, by this very assignment, play an additional role of a prognostic biomarker. Furthermore, in the context of data science and predictive modeling for biomarker discovery, biomarkers are the means to predict the value of the target variable (such as a disease state or the level of toxicity). Consequently, in this context, they are all *predictive* vehicles, regardless of the BEST category (or categories) to which they would be assigned. On the other hand, some terms that are commonly used to describe some types of biomarkers are not explicitly included in the BEST list; for example, *screening* biomarkers, which target at-risk populations of asymptomatic individuals, in the hope of the early detection of specific conditions, for which early treatment may be crucial for positive outcomes.[5]

## 1.2 Biomarkers and Personalized Medicine

The main goal and promise of *personalized medicine* (also called individualized or – more recently – precision medicine) is to tailor treatment to the various individual characteristics of a patient, such as the patient’s genomic, transcriptomic, proteomic, epigenetic, or metabolomic profiles. More generally, personalized medicine will try to determine an optimal approach to a particular patient’s care (prediction, prevention, diagnosis, or therapy) by integrating comprehensive information about the patient’s condition, which will include specific profiles identified by specific biomarkers. Hence, moving toward this goal will require discovery, validation, and clinical implementation of many new biomarkers; it is quite likely that most of them, especially the omic ones, will be multivariate biomarkers.

Although the paradigm of Western medicine can be seen as treating all patients with a specific diagnosis in the same – or very similar – way, one may nevertheless argue that conventional medicine has been – at least to some extent – personalized for a long time, as some individual characteristics of a patient were always taken into account and, thus, that personalized medicine is not a novel concept. However, developments in molecular biology and recent technological advances allow for the detection or quantification of not only many new physiological or pathological characteristics of patients, but also – and foremost – new types and scopes of such information (like those based on

[5] They are treated by the BEST biomarker taxonomy as a subtype of diagnostic biomarkers that are used for screening. Although this may be reasonable, the term *screening biomarkers* has been and probably will continue to be used to describe this specific type of biomarker, for which a low prevalence of the screened condition requires focusing on biomarkers with a very high specificity (see the example in Chapter 4).

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

the analysis of whole genome or transcriptome data, or on the simultaneous analysis of large numbers of proteins, metabolites, or non-coding regulatory RNAs). This facilitates the "personalization" of a patient's care in a much more specific and individualized way. And this process has just begun. How are (and will) such high-throughput technologies be utilized? By facilitating the discovery of new biomarkers. Since it is likely that we will need, and be able, to identify profiles associated with more and more complex biological processes, it is quite safe to assume that many, if not most, such new biomarkers will be multivariate ones.[6]

A step toward truly personalized medicine is *stratified medicine*, which assigns patients to specific subpopulations (or strata) based on molecular profiles as well as clinical and environmental variables characteristic for those subpopulations. Patients with similar "signatures" that, for example, may include similar genotype, lifestyle, and environmental factors, are assigned to a particular stratum and thus, considered for similar treatment options.

## 1.3 Biomarker Studies

Although the design of biomarker studies may differ from study to study,[7] there are some general steps that may be associated with any of them:

- Sample acquisition and preparation
- Processing samples in a laboratory (technologies may include genomic, proteomic or transcriptomic microarrays, next generation sequencing (NGS), liquid chromatography, mass spectrometry, nuclear magnetic resonance (NMR) spectroscopy, etc.)
- Low-level data preprocessing (depends on technology)
- Quality control (which may include batch and confounding factors adjustment)
- Normalization
- Presenting data in a form ready for analysis (usually, as a data matrix of variables times samples)
- Exploratory data analysis (to gain some general insight into the data)

[6] Personalized medicine has many aspects; for example, wearable devices transmitting real-time data from patients (and possibly from healthy individuals) as well as various "online medicine" tools and services (which may be associated with the idea of personalized medicine, but are better described as "personalized healthcare"), raise not only privacy concerns, but also concerns about the ethical, social, legal, and political implications of personalizing medicine (Nuffield Council on Bioethics 2010; Prainsack 2017). Here, we are focusing on personalized medicine only in the context of biomarker discovery.

[7] Biomarker discovery studies are primarily retrospective studies (using data collected from patients, for whom the outcome was known before the study was designed). Their main goal is to identify an optimal biomarker and design a predictive model implementing the biomarker. However, after the biomarker has been discovered, a prospective study (enrolling patients for whom the outcome is unknown, and who are then observed over a period of time) using the biomarker may follow, to validate and evaluate the biomarker predictions in clinical settings.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

- Multivariate biomarker discovery (applying data science methods to identify an optimal biomarker for a particular target variable)
- Initial clinical evaluation
- Development of the test implementing the biomarker and the predictive model
- Independent clinical validation (predictive accuracy, reproducibility, relevance to clinical application)

The focus of this book is on the *multivariate biomarker discovery* step that involves application of appropriate data science methods in order to identify (and test) an optimal parsimonious multivariate biomarker for a particular target variable. Nevertheless, every step of this process is very important, and mistakes or deficiencies in any of them may severely impact the overall quality of a study.

For example, the collection of biological samples (of solid tissue, blood or other biofluids) should be preceded by a careful selection of patients (and controls) to promote homogeneity of the training data with regard to all variables except the target variable. Of course, absolute homogeneity is impossible, but as many factors as possible (such as age, gender, and ethnicity) should be taken into account in order to increase the probability that the main differences between the observations in the study are in fact due to the researched target variable.[8] The number of patients (or observations) included in the study is also a very important factor. Generally, the larger this number, the better and more robust the results expected. Although it may depend on what target is researched, we should not treat seriously those biomarker studies that are based on a dozen or so observations, as their results may be – at best – anecdotal. With the availability and the continually decreasing costs of high-throughput omic technologies, the *preferable* number of observations (or observations in a class, in the case of classification) should be at least in the hundreds. Furthermore, if the goal of a biomarker is to differentiate among specific conditions (or classes), then patients included in the study must be diagnosed into one of those classes with a very high level of confidence. Decisions on the type of collected samples should also depend on the prospective use of the biomarker. For example, if the biomarker is to be used for screening large populations, it should preferably be based on samples collected in the least invasive way possible (such as saliva, urine, or peripheral blood). Sample preparation, laboratory processing, and low-level data preprocessing (if necessary) depend heavily on the type of biological samples and the technology used. Normalization (as well as batch adjustment) may be necessary to make the data – which may have been prepared and processed in different labs, at different times, and not necessarily under the exactly the same conditions – comparable.

Analysis of the data may start once the good quality data is presented in the typical format of a data matrix with rows representing variables (such as the gene or protein expression level) and columns representing observations (patients, cell lines,

[8] In cases when this would be difficult to achieve – and when heterogeneity of the target population is not an intended aspect of the study design – we may decide to pursue segmentation predictive models based on separate biomarkers identified for different segments (or strata) of the target population (see Chapter 3 for more on this subject).

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

compounds, etc.). Each observation, say a patient, is also associated with the value of the target variable, whose prediction is the goal of the study. For classification biomarkers, this is the value of the categorical target variable, that is, a class label (such as the name of one of the differentiated disease states or subtypes). For estimation biomarkers, this is the value of the continuous target variable (such as the toxicity level). Data analysis typically starts with the exploratory data analysis (EDA), which – in addition to providing basic statistical information about the data – may identify some issues with the data quality, if they were not identified and corrected at the earlier steps. As a part of EDA, unsupervised data science approaches (such as clustering, principal component analysis, or self organizing maps) may be used to *visualize* the data and to look at how the observations may be grouped. If, for example, clusters of observations are aligned with the different laboratories the data were processed in, then this would indicate a serious problem with the data quality.

Multivariate biomarker discovery will be described in detail throughout this book; here, we will take a look at its main components:

- Identifying training and test data sets.

  It would be preferable if the training and test sets are independent statistical samples from the target population. However, if we only have one data set available for the analysis, then the data are randomly split into training and test sets. The latter (which, in this situation, is also called the holdout set) is set aside and not used or "seen" during the analysis. Hence, in either case, the test set will only be used to test the final predictive model resulting from the analysis.[9]

- Performing many multivariate feature selection experiments (with each of them using a different random subsample of the training data).

  This is the most important step of biomarker discovery based on high-dimensional data.

- Aggregating results of the feature selection experiments in order to:
  - identify an optimal size of the multivariate biomarker, and
  - select an optimal subset of variables to be included in the biomarker.
- Building and evaluating a predictive model (or models) that implements the identified multivariate biomarker.
- Testing the model on the test data set (that was never "seen" or used during the entire analysis).
- Facilitating biological interpretation of the multivariate biomarker.

A multivariate biomarker discovery project results in a multivariate biomarker (an optimal set of variables) with properly estimated predictive power and generalization

[9] Sometimes it may be convenient or necessary to split the available data into training, validation, and test sets. For example, the validation set may be used for tuning parameters of a predictive model implementing an already identified biomarker, or to choose from among a number of candidate multivariate biomarkers before the final biomarker is tested on the test data.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

abilities, as well as a predictive model implementing the biomarker. Preferably, the results should also include information that could facilitate biological interpretation of the biomarker.[10] If the identified biomarker and the predictive model have sufficiently high predictive power (which may be evaluated, for example, by sensitivity, specificity, and accuracy for classification biomarkers, or by the mean squared error for estimation biomarkers), then subsequent stages of the biomarker study may be initiated.[11]

To have any clinical value, a biomarker must have sufficient predictive performance, must be reproducible, and must be acceptable and relevant for clinical use; its biological interpretation may also play an important role in its acceptance. Hence, each new biomarker and a test implementing it have to be subject to thorough independent clinical validation. It appears, however, that the major obstacle in the adoption of new biomarkers is not the fact that some of them are failing the clinical validation stage, but the fact that many of them are not even entering this stage. One of the possible reasons for this situation is that biomarker discovery projects have a good chance of being published; however, clinical validation of their results may be seen as having limited scientific value and thus, lower chances for funding and publication (Kumar and Van Gool 2013).

Furthermore, if a multivariate biomarker discovery project was based on outdated paradigms, such as employing the one-variable-at-a-time (univariate) approach, or basing the results only on their statistical significance, then such intrinsic deficiencies would render its results unlikely to pass thorough validation. The algorithms and data science methods that are appropriate for multivariate biomarker discovery – and free from such deficiencies – will be described and discussed in detail throughout this book.

## 1.4 Basic Terms and Concepts

Although the terms and concepts used in the book are explained in places where they are introduced or discussed, short descriptions of a few of them are, nevertheless, provided here for ease of reference.

### *Multivariate Biomarker*

A multivariate biomarker is a set of variables (representing, for example, the gene or protein expression levels), whose combined pattern of values can – with high accuracy – predict the value of the dependent variable (such as disease state or the probability of relapse) for new or future observations. To be considered a multivariate biomarker, the set has to be identified via multivariate methods. Such methods consider sets of variables, rather than individual ones. Hence, individual relations between each independent variable and the dependent variable are irrelevant. It is also very important to understand that ***a multivariate biomarker* is a *biomarker***. It is not a “set of biomarkers”

[10] Which may potentially answer the questions of why and how the biomarker works.

[11] Perhaps assuming also that the predictive power of the new biomarker is significantly higher than that of existing biomarkers (if any) for the same target variable.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

and should not be referred to by using such univariately-biased terms. The variables included in the multivariate biomarker – all of them together, as a set – constitute a single biomarker. None of them individually is a biomarker; if any of them were a biomarker for the researched goal, there would be no need for the multivariate biomarker. Therefore, in this book, whenever the terms *biomarker* or *biomarker discovery* are used without qualifications, they should be treated as synonymous with *multivariate biomarker* and *multivariate biomarker discovery*, respectively.

### Optimal Multivariate Biomarker

An optimal multivariate biomarker should be parsimonious and well generalizable. This means that it should consist of as few variables as possible, while simultaneously trying to maximize its predictive power. Hence, its optimality should be based on a compromise therebetween.

### Independent Variables

Independent variables are characteristics of the observations (say, patients) included in the training data (for example, the gene expression levels for 20,000 or so genes). Only some of them will be included in the multivariate biomarker and will be used to predict the value of the dependent variable.[12] It has to be emphasized that the term ***independent variables*** should be understood only in the context of independent variables versus the ***dependent variable***. It does not provide *any* information about relations and correlations among the independent variables. Whenever the term “variables” is used, in this book, without qualification, it will refer to independent variables.

### Dependent Variable

Predicting values of the dependent variable is the goal of biomarker discovery (or, more generally, of any predictive modeling). If the dependent variable is categorical, we are solving a classification problem; if it is continuous, we are solving a regression problem. The synonyms for the dependent variable are ***target*** variable and ***response*** variable.

### Biomarkers for Classification

If the dependent variable is categorical, the goal of biomarker discovery is to identify a biomarker that will be able to classify new patients (or, more generally, new observations) into one of the differentiated categories (such as two or more disease states). In this case, predictive modeling is using classification learning algorithms; the

[12] This is why we do not call all of them “predictors” (as they are often referred to in the literature, especially in statistical texts). With thousands of variables in typical high-dimensional data, most represent noise, and only a few of them will be selected into a multivariate biomarker and eventually used for prediction.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

identified multivariate biomarker can be called a ***classification biomarker***, and the predictive model that implements it – a ***classification model***.

### Biomarkers for Estimation

If the dependent variable is continuous, the goal of biomarker discovery is to identify a biomarker capable of predicting (estimating) the value of this dependent (target, response) variable (for example, the probability of relapse). In this case, predictive modeling is using regression learning algorithms; the identified multivariate biomarker can be called an ***estimation biomarker***, and the predictive model implementing it – a ***regression model***.

### Personalized Medicine

Personalized medicine attempts to determine an optimal and individualized approach to a patient's care (prediction, prevention, diagnosis, or therapy) by integrating comprehensive information about the patient's condition, especially by considering the patient's omic profiles (such as genomic, transcriptomic, proteomic, or metabolomic signatures). *Stratified medicine* is a step toward personalized medicine – patients' omic signatures (and treatment options) are not yet considered at a personalized level, but are instead matched to the profiles characteristic for specific subpopulations (or strata).

### Predictive Modeling

Although, in the context of this book, *predictive modeling* can be seen as synonymous with *multivariate biomarker discovery*, one may argue that the goal of biomarker discovery is to find an optimal biomarker, and that predictive modeling also includes building a predictive model (or models) based on such an optimal biomarker. Therefore, even if in practice both of these goals are essential parts of a properly designed biomarker discovery process, we will also, to avoid any confusion, use such phrases as *predictive modeling for biomarker discovery*.

### Multivariate Methods

Multivariate methods consider all independent variables simultaneously; thus, interactions and correlations among the variables are taken into account. Even if a multivariate algorithm sometimes focuses on a specific variable (for example, a feature selection method may consider which of the variables should be added or removed from the currently considered subset of variables), such a variable is always evaluated in the context of other variables.

### Univariate Methods

Univariate methods consider only one independent variable at a time, and evaluate the relationship between each of them and the dependent variable individually, ignoring

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

any and all relationships among the independent variables. As such, univariate methods are not used for multivariate biomarker discovery.

### *Supervised Learning Algorithms*

Supervised learning algorithms are methods used for predictive modeling (thus, including biomarker discovery). They are supervised by the values of the dependent variable that are associated with the observations in the training data, and they are used to identify a multivariate biomarker and predictive model (implementing the biomarker), which can then be used to predict the value of the dependent (response) variable. The goal of supervised learning is to maximize the accuracy of predicting the value of the response variable (which may be achieved via maximizing class separation – for classification, or via maximizing the proportion of the explained variation in the response variable – for regression) for *new* observations.

### *Unsupervised Learning Algorithms*

Unsupervised learning algorithms are used to identify groups of similar observations or groups of similar variables. They can provide visualization of high-dimensional data (and the grouping results) in a low-dimensional space, and are thus valuable tools for exploratory data analysis. However, unsupervised methods are blind to the dependent variable and are therefore inappropriate for biomarker discovery. It should be stressed – as this is still a quite common misconception – that they should not be used for decreasing the dimensionality of the training data that would be then used for biomarker discovery.

### *Training Data*

The training data set is used to perform all steps of predictive modeling for biomarker discovery. It includes $N$ observations (e.g. patients) and $p$ variables (e.g. gene or protein expression levels). Each of the observations may be represented by a $p \times 1$ vector $\mathbf{x}_i = \left[x_{1i}, \ldots, x_{pi}\right]^T$, and the value of the response variable, $y_i$, $i = 1, \ldots, N$, associated with the observation. Hence, the entire training data may be represented by a $p \times N$ matrix $\mathbf{X}$ and a $N \times 1$ vector $\mathbf{y}$.

### *Test Data*

The test data set is used exclusively for testing the performance of the optimal multivariate biomarker identified by the predictive modeling analysis. It has to be unavailable during this analysis. Ideally, the test data would be independent of the training data (for example, collected in a different geographical region and processed in a different lab). If such independent data are unavailable, then the data available to the project need to be randomly split into the training and test data, and the test data set aside, and used only after the analysis is completed.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

### Target Population

Both training and test data sets are statistical samples from the target population, which is the population for which we are building a predictive model. For classification modeling, when the dependent variable is categorical – and, for example, represents several disease states – each category could be considered a separate population; however, while keeping this in mind, we will often refer to all of them by the singular term.

### High-dimensional Data

High-dimensional biomedical data sets, for which the number of variables is greater, or often much greater, than the number of biological samples, are routinely generated by current high-throughput omic technologies. Applying – to such data – traditional statistical or predictive modeling methods that have been successfully used in low-dimensional settings will virtually guarantee overfitting. It is very likely that most of the variables in such high-dimensional data sets represent noise, and to be able to extract a true signal therefrom, more sophisticated heuristic multivariate search algorithms or regularization methods need to be used.

### Feature Selection

The goal of feature selection for multivariate biomarker discovery is to identify a small subset of independent variables whose combined pattern of values allows for the accurate prediction of the value of the response variable for new observations. In this context, feature selection is synonymous with ***multivariate feature selection***. Furthermore, since we are interested in biomarkers consisting of some of the original variables (rather than any "engineered" features), these terms should also be understood as being synonymous with *variable selection*, and *multivariate variable selection*.

### Parallel Feature Selection Experiments

Performing only a single feature selection experiment, when analyzing high-dimensional data, will virtually guarantee overfitting (that is, finding a spurious pattern that exists in the training data, but does not exist in the target population). To overcome this curse of dimensionality, we should perform many parallel feature selection experiments, with each of them using a random subsample of the original training observations. By aggregating the results of such feature selection experiments, we can identify a multivariate biomarker that is much more likely to be generalizable (to the target population) than those resulting from a single feature selection run.

### Hyperparameters

In the context of predictive modeling, the term *parameter* may refer to a predictive model's parameter as well as to its *hyperparameter*; however, a distinction between the

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press

two is usually made. *Parameters* are such internal characteristics of a predictive model that are estimated directly from the training data. *Hyperparameters* are, however, external to the model, and their values are either set manually or require tuning. Hence, we can say that hyperparameters are *tunable* parameters.

### *Bias-Variance Tradeoff*

The bias-variance tradeoff refers to finding an optimal balance between fitting the model to the training data and its ability to be generalizable to the target population. If a predictive model perfectly fits the training data, it is likely that the identified pattern is a spurious one, which exists in the training data by chance, but does not exist in the target population. In such situations, the model performance in predicting new observations would be poor, and the model would have low bias and high variance. A high-variance model is very sensitive to changes in the training data, while the opposite is true for a high-bias model. The goal of a proper tradeoff is to find a model, which is not too complex to overfit the training data, but complex enough to provide accurate predictions for new observations from the target population.

https://doi.org/10.1017/9781009006767.003 Published online by Cambridge University Press