A Master Pipeline for Discovery and Validation of Biomarkers

Sebastian J. Teran Hidalgo[1], Michael T. Lawson[2], Daniel J. Luckett[2], Monica Chaudhari[2], Jingxiang Chen[2], Arkopal Choudhury[2], Arianna Di Florio[3], Xiaotong Jiang[2], Crystal T. Nguyen[2], and Michael R. Kosorok[2](✉)

[1] Department of Biostatistics, Yale University, Hew Haven, CT, USA
sebastian.teranhidalgo@yale.edu

[2] Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
jgxchen@email.unc.edu,
{mtlawson,luckett,mcunc12,arkopal,xiaotong,ctn92}@live.unc.edu,
kosorok@bios.unc.edu

[3] Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, Wales, UK
diflorioa@cardiff.ac.uk

Abstract. A major challenge in precision medicine is the development of biomarkers which can effectively guide patient treatment in a manner which benefits both the individual and the population. Much of the difficulty is the poor reproducibility of existing approaches as well as the complexity of the problem. Machine learning tools with rigorous statistical inference properties have great potential to move this area forward. In this chapter, we review existing pipelines for biomarker discovery and validation from a statistical perspective and identify a number of key areas where improvements are needed. We then proceed to outline a framework for developing a master pipeline firmly grounded in statistical principles which can yield better reproducibility, leading to improved biomarker development and increasing success in precision medicine.

Keywords: Biomarker discovery · Reproducibility · Data mining · Machine learning

1 Introduction

Biomarkers occupy a position of fundamental importance in biomedical research and clinical practice. They can be employed for a variety of tasks, such as diagnostic tools or surrogate endpoints for clinical outcomes; Table 1 gives several examples of biomarkers and their uses. In this chapter, we will primarily focus on **prognostic biomarkers**, which provide information on the natural history of a disease and help in estimating a patient's overall outcome or prognosis, and **predictive biomarkers**, which provide information on the likelihood that a patient will respond to a therapeutic intervention and help in identifying the

A. Holzinger (Ed.): ML for Health Informatics, LNAI 9605, pp. 259–288, 2016.
DOI: 10.1007/978-3-319-50478-0_13

Table 1. Examples of biomarkers and their significance in medicine

Biomarker	How it is measured	Relevance
Body mass index (BMI)	Person's weight in kilograms divided by the square of height in meters	Associated with a number of health outcomes, including obesity [6] and death [7]
Periodic variation of R-R intervals (heart rate variability)	Calculated from continuous electrocardiogram record	Indicator of the activity of the autonomic nervous system [8], predictor of survival after heart attack [8,9]
Glycosolated hemoglobin (HbA1c)	Assayed from blood samples	Diagnostic marker for diabetes [10]; an indicator of glycemic control in patients with diabetes [11]
KRAS Somatic mutations	Assayed from tumor samples	Associated with treatment response in colorectal cancer [12]
BRCA1 Germline mutations	Assayed from human buccal cells or blood	Associated with the risk of breast and ovarian cancer [13]

most effective course of treatment. Note that some biomarkers, such as estrogen receptor status in breast cancer, can be both prognostic and predictive [1].

The emergence of "-omics" approaches has enabled new biomarkers to step into the limelight, holding promise for precision medicine, an emerging field that builds individual variability in biological and environmental factors into its approach to treating disease [2]. Parallel advances in high-throughput technologies have generated an unprecedented amount of data ("big data"). The sheer scale and variety of information available, along with its structural and functional heterogeneity and often its inconsistencies, have led to the current paradox: biomarker discovery is more possible than it has ever been before, but it is also more problematic and inefficient. Of the hundreds of thousands of disease-associated markers that have been reported, only a small fraction have been validated and proven clinically useful [3–5].

It has become abundantly clear, given the current difficulties, that research practices in biomarker discovery must be firmly grounded in statistical and biomedical practices. In this chapter, we outline a framework for developing a master pipeline for biomarker discovery and validation that is aimed at increasing the reliability and reproducibility of biomarker discovery experiments.

We will first review various approaches to study design, highlighting those that are relevant to biomarker discovery trials. We will then discuss the challenges of ensuring data quality in the world of "big data" and propose strategies for data collection and curation. We will introduce several statistical analysis techniques, devoting special attention to the role of machine learning techniques. We will emphasize the role of traditional statistical considerations, such as power analysis, in biomarker studies, regardless of the specific analysis technique. We will then mention several approaches to the validation and evaluation of biomarkers. We will conclude by discussing the clinical interpretation of biomarkers and the central role it plays, and by providing some directions for future research.

2 Glossary

Accelerated Failure Time (AFT) Model: specifies the regression model $\lambda(t|Z) = e^{-\beta' Z(t)}\lambda_0\left(e^{-\beta' Z(t)}\right)$ for the hazard function.

Biomarkers or *biological markers*: quantifiable, objectively measured and evaluated indicators of physiological and pathogenic processes, responses to interventions, and environmental exposures.

Biclustering: a clustering method which considers groupings of rows (experimental subjects) and columns (covariates) both together and independently.

Classification: a supervised learning method with a binary, ordinal, or multi-category outcome variable. The focus of classification is placing observations into the correct class based on covariates.

Cluster Analysis: aims to group together individuals who are more similar to each other than the individuals assigned to other clusters. Examples include k-means, hierarchical, and spectral clustering.

Cox Proportional Hazard Model: specifies the semiparametric regression model $\lambda(t|Z) = \lambda_0(t)e^{\beta' Z(t)}$ for the hazard function.

Dimension Reduction: reduces the number of covariates and converts data to a lower dimensional space that is easier to analyze. Examples include principal component analysis (PCA), linear discriminant analysis (LDA), and classical multidimensional scaling (MDS).

False Discovery Rate (FDR): The average proportion of false discoveries (V) among all discoveries R, or rejections of the null hypothesis, in a study, i.e. $FDR = E[V/R]$.

Family-wise Error Rate (FWER): The probability that even one false discovery (V) will be made in a study, i.e. $FWER = \Pr(V \geq 1)$.

G-Estimation: a method for estimating causal effects in structural nested models, while accounting for time-varying confounders and mediators.

Least Absolute Shrinkage and Selection Operator (LASSO): an L_1-penalized regression technique for the linear model $Y = X\beta + \epsilon$. The L_1 penalization causes the estimates of coefficients for unimportant covariates to shrink to exactly zero, thereby performing model selection.

Machine Learning Methods: flexible, nonparametric methods derived from the field of computer science, which arose from the study of pattern recognition in artificial intelligence. Machine learning methods are well-suited for prediction in a variety of complex data scenarios, but many do not have well-studied inferential properties.

Negative predictive value (NPV): the probability that a subject is disease negative given that they test negative.

Nonparametric Methods: assume that the data arise from a complicated process whose set of explanatory parameters is not fixed. Offer flexibility at the expense of interpretability and efficiency.
Non-penalized Methods: estimate parameters by directly maximizing a likelihood function.

Outcome Weighted Learning: a machine learning method suited to identifying predictive biomarkers within a randomized trial or observational study for binary treatments with substantial treatment heterogeneity.

Parametric Methods: assume that the data arise from a known probability distribution that is determined by a small, fixed number of parameters. Offer interpretability and efficiency at the expense of strong assumptions.

Penalized Methods: estimate parameters by maximizing a likelihood function that is modified by a penalty term. Penalized methods are used to regularize parameter estimates, which aids in prediction by reducing overfitting.

Positive predictive value (PPV): the probability that a subject is disease positive given that they test positive.

Power: the probability that a null hypothesis that is actually false will correctly be rejected.

Predictive Biomarker: a biomarker that helps in determining which of several possible treatments will be most beneficial to a patient. Causal in nature.

Prognostic Biomarker: a biomarker that helps in ascertaining or predicting disease status. Not causal in nature.

Q-Learning: a regression-based machine learning method that estimates optimal personalized treatment strategies by directly estimating the Q-functions.

Random Forests: nonparametric machine learning tools that combine decision trees, which provide low bias without strong assumptions, bootstrap aggregation, which reduces the variance of the tree-based estimate, and feature randomization, which reduces the correlation between trees for further variance reduction.

Receiver Operating Characteristic (ROC) Curve: a plot of the sensitivity and specificity of a diagnostic test over all possible cutoff values.

Regression: a supervised learning technique that poses a model for the mean of an outcome variable that depends on the covariates of interest and the inherent variability of the sample.

Reproducibility: the ability of a study's results to be corroborated or confirmed by similar experiments in similar settings.

Semiparametric Methods: assume that the data arise from a process containing a parametric piece and a nonparametric piece. Offer a middle ground between the flexibility of nonparametric methods and the efficiency and interpretability of parametric methods.
Sensitivity: the probability of a positive test given that a subject is disease positive; also called the true positive fraction.

Singular Value Decomposition (SVD): the factorization of a data matrix $\boldsymbol{X} = \boldsymbol{UDV}^T = \sum_{k=1}^{r} s_k \boldsymbol{u}_k \boldsymbol{v}_k^T$, where r is the rank of $\boldsymbol{X}$, $\boldsymbol{U}$ is a matrix of orthonormal left singular vectors, $\boldsymbol{V}$ is a matrix of orthonormal right singular vectors, $\boldsymbol{D}$ is a diagonal matrix with positive singular values on its diagonal. $\boldsymbol{X}$ can be approximated $\boldsymbol{X} \approx \boldsymbol{X}^{(K)} \equiv \sum_{k=1}^{K} \boldsymbol{u}_k \boldsymbol{s}_k \boldsymbol{v}_k^T$ where $\boldsymbol{X}^{(K)}$ is the closest rank-K approximation of $\boldsymbol{X}$ [14].

Specificity: the probability of a negative test given that a subject is disease negative; one minus specificity is also called the false positive fraction.

Supervised Learning: a class of learning methods that explicitly incorporate an outcome variable. Supervised learning methods can be used to predict future values of the outcome, assess the effect of covariates on the outcome, or both.

Support Vector Machine (SVM): a supervised learning method that classifies data points with a binary outcome based on the optimal separating hyperplane.

Unsupervised Learning: a type of machine learning that uses unlabeled data to conduct statistical inference, where the covariates of interest are known but the outcome variables are not given.

3 Biomarker Discovery and Validation Pipeline

3.1 Study Design

While the role a biomarker plays in a study—prognostic or predictive—is important, trouble can arise if investigators focus too much on the details specific to that role and lose sight of the fundamentals of study design. In *Anna Karenina*, Leo Tolstoy wrote that "Happy families are all alike; every unhappy family is unhappy in its own way." A similar statement can be made about biomarker studies. Successful studies will address similar minimal criteria at each phase of development, while unsuccessful studies can fail to do so in any number of unique and creative ways. Study objectives, outcome measures and their reliability, availability of appropriate analysis methods, and the biomarker's clinical context should all play an essential role in determining the study design.

Studies involving prognostic biomarkers tend to focus on the development and evaluation of clinical assays and screening tests for a disease. Pepe et al. (2001) [15] suggest five phases for prognostic biomarker studies:

1. **Phase 1: Pre-clinical Exploratory Studies** Phase 1 studies identify and prioritize potentially useful biomarkers from a large pool of candidates. A biomarker's utility is based on how significantly its levels differ between disease cases and healthy controls, which are often matched to account for patient heterogeneity.
2. **Phase 2: Clinical Assay Development** Phase 2 studies develop reliable clinical assays based on the biomarkers identified in phase 1. Clinical assays employ non-invasively obtained specimens that are simple to collect.
3. **Phase 3: Retrospective Longitudinal Repository Studies** Phase 3 studies assess how well a biomarker can be used for early disease detection by examining whether levels of the biomarker in clinical specimens differ significantly between disease cases and healthy controls during the time period before the cases were diagnosed. Phase 3 studies can be used to define criteria for a screening test, which is evaluated in future phases.
4. **Phase 4: Prospective Screening Studies** Phase 4 prospective studies evaluate the performance and determine the operating characteristics (see Sect. 3.6) of a screening test. Patients are screened with the proposed test, and true disease status is ascertained with a "gold standard" diagnostic test.
5. **Phase 5: Disease Control Studies** Phase 5 confirmatory randomized trials address whether biomarker-based screening reduces the actual burden of disease. There is a distinction between success in phases 4 and 5: a biomarker may screen for disease effectively but not lead to a decrease in mortality due to other factors, such as lack of appropriate treatment.

While we have presented these phases as a straightforward progression, not all prognostic biomarkers will progress linearly through the five phases, and some study designs will combine elements of multiple phases.

Predictive biomarkers are typically incorporated into pivotal phase III trials of experimental treatments. The reason is twofold. First, predictive biomarkers assist in determining which of several treatments is likely to be more effective for a given patient; this information can help clarify the treatment effect that a phase III trial intends to estimate. Second, predictive biomarkers are causal in nature, and the setting of a randomized clinical trial provides the most compelling evidence to support claims of causation. Two simple questions can assist in selecting the correct phase III biomarker design: how many candidate biomarkers are in consideration, and how strong is the evidence that supports them?

Trials that incorporate one biomarker supported by strong evidence often take the form of biomarker-enriched, biomarker-stratified, or biomarker-strategy studies [16,17]. In all three cases, the study population is assayed for the biomarker of interest before randomization. In biomarker-enriched trials, only patients testing positive for the biomarker proceed to randomization; this scheme is particularly appropriate when biological evidence suggests that the test-negative population will not benefit from the treatment, raising concerns of ethics and efficiency [18,19]. In biomarker-strategy trials, patients are randomized into either a biomarker-directed arm, in which their treatment is dictated by the biomarker, or a control arm, where all patients receive control treatment; this scheme may be preferable when the biomarker-directed treatment strategy is complex [16]. In biomarker-stratified trials, patients are split into two groups, test-positive and test-negative, and then randomized normally within these groups; biomarker-stratified designs, when logistically and ethically appropriate, offer a great deal of efficiency [16,17]. These designs can be combined to address complex research questions—for instance, when a complex experimental therapy is enriched by multiple biomarkers at once [17].

Extensions of these methods can accommodate weaker assumptions on the biomarkers of interest by performing inference on biomarker properties in tandem with estimation of treatment effect. Adaptive threshold designs use a single biomarker without a pre-specified test-positive threshold, which reduces reliance on phase II studies to correctly determine the threshold [20]. Adaptive biomarker designs consider a relatively small pool of candidate biomarkers rather than a single biomarker, selecting the most promising biomarker or biomarkers during the course of the trial [18,20].

Even in the most restrictive setting, where investigators have a large pool of candidate biomarkers with little to no prior evidence supporting them, clever study designs enable valid statistical inference. One such design, the adaptive signature design, employs two outcome stages [21]. The first outcome stage tests for treatment efficacy at the α_1 significance level in the overall population of size N; if this stage is successful, the drug is considered generally useful. If the first stage does not find overall efficacy, the second stage uses the first N_1 accrued patients to train a machine learning classifier that divides the final $N_2 = N - N_1$ patients into two groups: those who are likely to benefit from the experimental treatment, E, and those who are not likely to, C. Then treatment efficacy is tested at the α_2 significance level in the promising group E. If efficacy is shown in phase 2, the drug is considered effective for the biomarker-selected group, and the machine learning classifier can be used to predict group status for future patients. Choosing $\alpha_1 + \alpha_2 = \alpha$ controls type I error at the α level; see Sect. 3.4 for more details. Adaptive signature designs were proposed with a simple machine learning classifier that aptly handled a variety of simulation settings [21], but any number of the more sophisticated methods discussed in Sect. 3.3 may prove useful in extending adaptive signature designs.

The incorporation of biomarkers poses its fair share of challenges to common considerations in phase III trials. For one, biomarker-based designs may complicate interim analyses, encouraging flexible stopping rules over rigid ones [16]. For another, biomarkers may define subgroups of scientific interest to test as secondary outcomes, making multiple comparisons adjustments (see Sect. 3.4) especially relevant [18].

In addition, several study designs address the task of discovering predictive biomarkers outside of phase III trials. Sequential multiple assignment randomized trial (SMART) designs offer a framework for applying predictive biomarkers to dynamic treatment regimes, often in phase II trials [22]. Electronic health record data, used in concert with causal inference, could give rise to efficient observational and other non-randomized predictive biomarker studies [3,23].

3.2 Ensuring Data Quality

Unlike in the myth of Athena's birth, which depicts the goddess leaping from Zeus's forehead fully formed and armed for battle, a well-executed study yielding reliable results does not spontaneously arise from a well-designed study. Data management is a crucial step of a successful study—mishandling a study's data threatens the validity of everything that follows.

Often, when errors creep into a dataset, they do so at the phase of data collection. Although human and measurement error will always lie outside an investigator's control, investigators can limit the impact of these factors. A detailed study protocol that lists how data collection should be carried out reduces ambiguity and lessens reliance on subjective judgments. Studies that utilize multiple data collection sites carry an additional burden: measurements must be consistent not only within sites, but across sites. For a much more thorough treatment of the topic, see the Data Acquisition section of the Society for Clinical Data Management's GCDMP 4.0 [24].

Investigators must also pay close attention to how their data are stored and linked. The optimal data management plan will vary from study to study based on numerous factors, including the physical and institutional proximity of collaborators, the volume and frequency of data collected, and the data use guidelines put in place by participating institutions. Extra care must be taken with "big data," which may tax the computing resources available to a research team. What should not vary is the investigators' approach to data management: data management requires a clearly stated plan and thorough documentation of all steps taken throughout the process. A data flow diagram may help clarify the steps of data collection, processing, and storage, and potentially aid in identifying problems [25]. Overall, investigators should balance two guiding principles: ease of access and protection of privacy.

While the first of these principles is intuitively clear, the second deserves some elaboration. In the course of data collection, investigators will have access to sensitive personal information, and investigators have a solemn obligation to protect the privacy of their participants to the greatest extent possible. This obligation may be legal in addition to ethical, thanks to privacy-protecting statutes such as HIPAA [26]. Investigators should familiarize themselves with statutes that apply to their study and make sure their methods of data collection and storage comply with all relevant guidelines.

Investigators should also be wary of losing sight of what their data mean functionally—they should be able to describe the information contained in every column of every dataset. Informative file and variable names can help, but they are not enough. As standard practice, investigators should draw up a data dictionary that explains each dataset and each variable the datasets contain.

3.3 Statistical Methods for Biomarkers

In this section, we provide a brief overview of statistical methods appropriate for the analysis of biomarker data. We pay particular attention to machine learning methods, which offer an attractive combination of flexibility and desirable statistical properties.

While we present a variety of methods for both supervised and unsupervised learning below, we wish to emphasize that the two are rarely as disjoint as they may appear in this section. They are often used in concert: covariates are often pre-processed through an unsupervised method before being employed in a supervised method, for instance. The matter is further complicated by latent supervised learning, which posits intermediate ground between supervised and unsupervised methods based on latent subgroups [27], and semi-supervised learning, which trains a model on both data where the outcome variable is observed and data where it is not [28].

Supervised Methods. The taxonomy of supervised methods is expansive—supervised methods can accommodate data from a truly staggering variety of studies. While the details of a specific biomarker study and data type are invaluable in selecting the correct method, the search for the appropriate method in any study can be aided by two general questions. First, what is the goal of the method? Second, what type of information does the method need to provide? The questions are clearly related, and together they often point directly to a small class of methods. For instance, if the investigators primarily care about prediction of future values of the outcome, and they do not particularly care about interpreting the effect of covariates, a nonparametric machine learning method may prove their best option; but if their primary research question is quantifying the relationship between a biomarker of interest and the outcome, a parametric or semiparametric model will likely serve them better. The spectrum of supervised methods offers a trade-off between flexibility and interpretability, between what Kosorok (2009) [29] calls "the ability to discover and the ability to generalize."

1. **Parametric Methods** Parametric methods assume that the data are generated by a known probability distribution with a small, fixed number of parameters (e.g. the mean and variance of a Gaussian random variable, or the rate of an exponential). The parameters of that distribution provide a concise way to summarize and interpret the data, and they serve as the target of statistical inference. Another attractive property of parametric methods is their efficiency: when the assumptions for a parametric method are truly met, the estimates that method provides are highly precise.

   Non-Penalized Methods Non-penalized methods estimate the parameters directly from the form of the probability distribution, or likelihood function, by finding the parameter values that maximize the likelihood. Several popular and widely-used methods belong to the class of non-penalized parametric methods, among them linear regression, logistic regression, and the accelerated failure time model for survival analysis.

   Accelerated Failure Time (AFT) Model The AFT model poses a regression model on the scale of the hazard function, λ, of the failure time T. Let Z denote the (potentially time-varying) covariates, and let $\lambda(t|Z)$ denote the hazard function at time t conditional on the covariates. Then the AFT model is given by $\lambda(t|Z) = e^{-\beta' Z(t)} \lambda_0 \left(e^{-\beta' Z(t)} t \right)$ where λ_0 represents the unobserved baseline hazard function of T when all covariates equal zero. When λ_0's parametric distribution is specified in advance, the AFT model is fully parametric [30]; when λ_0 is estimated nonparametrically, the AFT model is semiparametric [31]. The parameter β is the target of inference, as it describes the effect of the covariates on survival time. As an example, Altstein and Li (2013) [31] used the AFT model to discover biomarker-based latent subgroups in a melanoma trial.

Penalized Methods Penalized methods estimate model parameters by maximizing a likelihood function that is modified by a penalty term. The penalty term is added in to regularize parameter estimates, which can reduce overfitting and aid the model's performance in prediction. Many penalty terms can be chosen, each of which offer their own benefits; we present only one in this chapter. For a more thorough treatment of penalized methods, see chapters 3 and 4 of Hastie, Tibshirani, and Friedman (2008) [32].

LASSO The least absolute shrinkage and selection operator (LASSO) is among the most popular penalized methods, and is particularly useful for high-dimensional data where the number of variables is much larger than the number of data points. A primary reason is that the LASSO, in addition to regularizing parameter estimates, also sets the estimates of many coefficients exactly equal to zero—hence, the LASSO performs both regularization and variable selection. If we let Y_i denote the continuous outcome for patient i and X_i denote that patient's covariates, then the LASSO estimate of β minimizes the function $\sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$, where $\beta = (\beta_1, ..., \beta_p)$ and $\lambda \geq 0$ is an L_1-constrained penalty parameter. When λ is small, the LASSO estimate of β resembles the result from ordinary least squares, but as λ grows, increasingly many components of β are set equal to zero. Cross-validation is typically used to specify the value of λ. Once λ is chosen, the optimization problem simplifies to a quadratic programming problem, which can be solved through an efficient sequential algorithm [33]. There are many extensions of the LASSO which accommodate categorical data [33,34], survival data [35], study designs with interactions [36], and mixed models [37,38].

2. **Nonparametric Methods** Nonparametric methods assume that the data are not generated by a probability distribution with a fixed number of parameters—rather, the number of parameters needed to explain the data is allowed to grow to infinity as the sample size grows. While some nonparametric methods are complex and computationally intensive, a number of convenient nonparametric methods are available for data analysis. Many **machine learning techniques**, which perform well at a variety of difficult prediction tasks and are becoming increasingly well-studied from a statistical perspective, fall under the umbrella of nonparametric methods.

   Random Forests Random forest approaches, which are fully nonparametric machine learning tools, offer great predictive power in both regression and classification. The technical details underlying random forests are rather complex, and a full treatment of them is beyond the scope of this chapter; we simply mention that they combine the strengths of two well-known techniques, decision trees and bootstrap aggregation [39]. Random forests offer a measure of covariate importance, albeit a much less interpretable measure than a regression coefficient in a parametric model. Random forests have inspired many extensions. Among them are Bayesian additive regression trees (BART), which marry the random forest and Bayesian nonparametric approaches [40], and reinforcement learning trees (RLT), which use reinforcement learning to select important variables while muting unimportant ones [41]. RLTs appear particularly promising, as several results about their statistical inference properties have been shown [41]. As an example, Gray et al. (2013) used random forests to classify patients into subgroups of Alzheimer's disease based on a variety of biomarkers, including MRI volumes, cerebrospinal fluid measures, and genetic markers [42].

Deep Learning Deep learning methods have proven quite powerful in prediction, both in the regression and classification setting, in a variety of difficult prediction contexts, such as speech recognition [43] and image processing [44]. The most commonly used deep learning methods are deep neural networks, which posit that the covariates are related to the outcome through multiple hidden layers of weighted sums and nonlinear transformations. Some deep neural networks, such as convolutional neural networks, build a spatial dependency into the structure of the hidden layers. Deep neural networks can be plagued by overfitting; the introduction of dropout, which reduces dependencies among nodes in the hidden layers, appears to greatly reduce this weakness [45]. Although deep learning techniques have met with great success in application, their inferential properties are, as of yet, not well-studied, though research in this area is currently active. As an example, Xiong et al. (2015) used deep neural networks to predict disease status based on alternative genetic splicing [46].

Support Vector Machine Another popular machine learning method for nonparametric classification is the support vector machine (SVM). The support vector machine considers each observation of covariates X_i as a point in d-dimensional space with a class label $Y_i \in \{-1, 1\}$. The SVM sets up a classification rule by finding the $d-1$-dimensional hyperplane that optimally separates points with $Y_i = 1$ and $Y_i = -1$. This problem can be formulated as a constrained optimization problem and analytically solved; for details, see chapter 7 of Cristianini and Taylor (2000) [47]. The SVM can be extended to nonlinear classification using reproducing kernel Hilbert spaces [48] and regression settings using support vector regression [49].

Q-Learning Q-learning is a regression-based method for estimating an optimal personalized treatment strategy, which consists of a sequence of clinical decisions over time. Q-learning estimates a set of time-varying Q-functions, Q_t, $t = 1, \ldots, T$, which take the current patient state S_t and the clinical decision D_t as inputs and give the value, which is based on the clinical outcome of interest, as an output. When the Q-functions have been estimated, the only information we need to determine the optimal future treatment is the patient's current state. Estimates of the Q-functions, $\{\hat{Q}_1, \ldots, \hat{Q}_T\}$, are obtained through a backwards iterative algorithm [50]. The estimated Q-functions allow us to estimate the optimal treatments:

$$\hat{\pi}_t = \underset{d_t}{\operatorname{argmax}}\ \hat{Q}_t(s_t, d_t) \qquad \text{for } t = 1, \ldots, T,$$

That is, we select the treatment sequence $\{\hat{\pi}_1, \ldots, \hat{\pi}_T\}$ that maximizes the sequence of Q-functions. Q-learning can be applied to complex, multi-stage trials, such as sequential multiple assignment randomized trials [51].

G-Estimation Predictive biomarkers are often used to provide information for several decisions over a period of time. Suppose that we are interested in discovering a causal relationship between an exposure and an outcome over time. If any time-varying confounder is also related to future exposure, standard methods for adjusting for confounders will fail. G-estimation, a method for estimating a causal effect in structural nested models while accounting for both confounders and mediators [52], is useful in this setting [53,54]. G-estimation has been applied in a number of settings where time-varying covariates are of interest, such as cardiovascular disease [55] and AIDS [56]. Vansteelandt et al. (2014) [52] give a more thorough overview of G-estimation and structural nested models.

Outcome Weighted Learning (OWL) OWL offers a method for identifying predictive biomarkers in a randomized trial or observational study testing binary treatments which have substantial treatment heterogeneity [57]. OWL estimates the optimal individualized treatment rule by formulating it as a weighted classification problem, which can be solved through a computationally efficient algorithm. For ease of notation, we present the case of a two-arm randomized trial in this chapter. Suppose we have a binary treatment $A \in \{-1, 1\}$, and that the p biomarkers of the n patients are recorded in the $n \times p$ covariate matrix X. Let R denote the clinical outcome, or reward, that we wish to maximize. In this framework, an individualized treatment rule (ITR) is a function that takes the covariates as an input and recommends one of the two treatments as an output. The optimal ITR, then, is the function that satisfies

$$\mathcal{D}^*(x) = \underset{\mathcal{D}}{\arg\min} \left\{ E\left(\frac{R \cdot 1\{A \neq \mathcal{D}(X)\}}{\Pr(A)} \right) \right\}, \tag{1}$$

where $\Pr(A)$ is the prime probability of being assigned to treatment A [58]. Essentially, OWL finds the optimal ITR by matching the treatments of patients with a high reward and mismatching patients who have small rewards. Equation 1 with 0–1 loss yields an optimization problem that is non-deterministic polynomial-time (NP) hard, and can be quite computationally intensive to solve. To alleviate this difficulty, OWL employs the hinge loss used in the Support Vector Machine. In addition, OWL uses regularization to stabilize the estimate of the ITR based on the observed sample (x_i, a_i, r_i),

$i = 1, \cdots, n$. Hence, OWL searches for the decision rule f that minimizes the regularized optimization problem

$$\frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\Pr(a_i)} \left(1 - a_i f(x_i)\right)_+ + \lambda ||f||^2, \tag{2}$$

where $||f||^2$ is the squared L_2-norm of f and λ is a tuning parameter used to balance model accuracy and complexity. Once we have estimated f, the OWL ITR is simply $\hat{\mathcal{D}}(X) = \text{sign}(f)$. OWL has many attractive inferential properties, including results for Fisher consistency and risk bounds [57].

Several methods extend the capabilities of OWL. Zhou et al. (2015) [59] improve OWL model accuracy by fitting the reward function with the covariates ahead of time and plugging the residuals into Eq. 2 instead of the reward. Xu et al. (2015) [60] add an L_1 penalty term to OWL, allowing OWL to perform variable selection. Zhao et al. (2015) [61] extend OWL from a single-stage trial, as described above, into multiple-stage clinical trials, allowing OWL to inform optimal dynamic treatment regimes.

OWL is not the only approach to finding predictive biomarkers in trials with treatment heterogeneity. Other approaches examine the interaction between treatment and candidate predictive biomarkers, including recent work in tree-based methods that provide flexible models for determining variable importance [62]. Zhang et al. (2012) take a similar approach based on a semiparametric model that uses inverse-probability weighting to analyze observational studies [63]. Tian et al. (2014) [64] model the interactions between the treatment and modified covariates in a variety of settings, including the setting with a large pool of biomarkers about which little is known and only a subset of patients expected to benefit from treatment.

3. **Semiparametric Methods** Semiparametric methods contain a parametric piece and a nonparametric piece, offering a trade-off between the flexibility offered by nonparametric methods and the efficiency offered by parametric methods. Although some semiparametric models are in wide use, others which offer the same attractive balance have found adoption much slower. In this section, we only discuss the most commonly used semiparametric model: the Cox proportional hazards model.

   Cox proportional hazards model The Cox proportional hazards model, like the AFT model, poses a regression model on the scale of the hazard function [65]. If we let Z denote the potentially time-varying covariates, and we let $\lambda(t|Z)$ denote the hazard function at time t conditional on the covariates, the Cox model can be expressed as $\lambda(t|Z) = \lambda_0(t)e^{\beta' Z(t)}$, where λ_0 denotes the unobserved baseline hazard function, which can be estimated nonparametrically or assumed to follow a parametric distribution. The former is more common, and leads to a semiparametric model. The parametric piece of the Cox model is $e^{\beta' Z(t)}$, and β is estimated through maximum partial likelihood. As an example, Kalantar-Zadeh et al. (2007) used the Cox model to show an association between levels of A1C and mortality risk after controlling for several demographic characteristics [66].

Latent Supervised Learning. Latent supervised learning, a novel approach that exists in the middle ground between supervised and unsupervised learning, simultaneously handles parameter estimation and the problem of unlabeled subgroups. To illustrate: suppose that the patient population in a clinical trial consists of several underlying subgroups, and treatment efficacy differs according to these latent subgroups. Ignoring these subgroups can cause a supervised method to produce poor estimates [27]. Let Y denote the outcome and X denote the covariates. Wei and Kosorok (2013) proposed the following model for binary classification using latent supervised learning [27]:

$$Y = \mu_{1,0} 1\{\omega_0^T X - \gamma_0 \geq 0\} + \mu_{2,0} 1\{\omega_0^T X - \gamma_0 < 0\} + \epsilon$$

The model posits that an unknown linear function of the covariates determines the mean value of the outcome—patients with different signs of $\omega_0^T X - \gamma_0$ have different means ($\mu_{1,0}$ vs. $\mu_{2,0}$). The underlying subgroup structure is assumed to be linear. When ϵ is Gaussian, model parameters can be estimated through maximum likelihood [27]. These model parameters provide not only an estimate of the treatment effect, but also subgroup predictions based on covariates.

The assumption of latent subgroups that depend on biomarkers is not only reasonable, but often of primary scientific interest. Methods that can accommodate this assumption and simultaneously provide estimates of its effect, as the emerging field of latent supervised learning does, offer a great deal of promise for future research.

Unsupervised Methods. In some settings, it may be impractical to observe the outcome due to logistics or cost, or the exact nature of the outcome may not be known. More commonly, investigators may wish to perform some sort of data pre-processing before plugging their data into a supervised method. In these settings, **unsupervised learning** techniques assist in conducting statistical inference about the underlying structure of the data. Identifying the underlying structure can provide valuable insight into different classes that exist among the data, and which subset of variables determines those classes [67].

Dimension reduction reduces the number of effective covariates and brings data into a lower-dimension space that is easier to analyze. The most commonly used dimension reduction techniques are principal component analysis (PCA), linear discriminant analysis (LDA), and classical multidimensional scaling (MDS). While these methods are widely used, they may fail to capture the underlying structure of complex data. In these situations, nonlinear dimension reduction methods may prove more fruitful. Isometric feature mapping, or Isomap, builds a weighted graph using data points as nodes and calculates the geodesic distance between data points as the sum of weights along the shortest path between points. The geodesic distance is then used in place of Euclidean distance in MDS, which allows for Isomap to handle points that lie on a nonlinear manifold [68]. Another popular technique is t-Stochastic Neighbor Embedding, or t-SNE, which finds a low dimensional mapping to minimize the Kullback–Leibler divergence between the distributions of the data in low and high dimensional spaces [69]. t-SNE commonly employs a normal distribution in the high dimensional space and a t-distribution in the low dimensional space. Many other nonlinear dimension reduction methods are available, such as Locally Linear Embedding (LLE) [70] and diffusion maps [71].

Traditional clustering methods, such as k-means and hierarchical clustering, treat covariates of interest as a monolithic collection, which proves ineffective if only a subset of the covariates is truly informative. **Biclustering** methods address this issue by considering clusters among both subjects and covariates simultaneously. We present several methods for biclustering.

Let X denote the overall data matrix. Large Average Submatrix (LAS) finds K constant, potentially overlapping submatrices of X via maximum likelihood, then poses X as the sum of these submatrices and random noise [72]. Sparse clustering imposes a Gaussian likelihood on the biclusters of X, where the biclusters have unique means and common variance. The means are estimated through L_1-penalized least squares, which sets many bicluster means identical to zero, inducing sparsity [73].

Several biclustering methods use **singular value decomposition (SVD)** for dimension reduction. Sparse SVD (SSVD) additionally shrinks small nonzero singular vectors to zero through an L_1 penalty on the squared Frobenius norm of X, meaning only a checkerboard pattern of influential rows and columns remains nonzero [14]. Heterogeneous sparse SVD (HSSVD) functions in the case where biclusters vary in both mean and variance. HSSVD has the advantages of scale and rotation invariance, and has an improved capacity for detecting overlapping biclusters compared to classic SVD, as well as improved performance relative to several methods even in the case where the biclusters have homogeneous variance [74]. Currently, HSSVD has limited utility in handling count data and data that arise from more than one "-omics" platform.

Example: The following example comes from a lung cancer dataset with the expression levels of 12,625 genes from 56 patients discussed in [74]. The investigators performed HSSVD, classifying patients' lung cancer subtype (normal lung, pulmonary carcinoid tumors, colon metasteses, and small-cell carcinoma). They then compared the results of HSSVD with those of FIT-SSVD, LSHM, and SVD. The comparison is visualized in the checkerboard plots in Fig. 1. Successful biclustering is expected to produce the checkerboard appearance exhibited by HSSVD and FIT-SSVD, but not LSHM and SVD. The biclusters are identified as the rows divided by the white lines.

3.4 Power

Machine learning techniques appear prominently in biomarker discovery studies, especially in genomics settings [75,76]. While machine learning techniques are well-suited to analyzing large, heterogeneous datasets, many core statistical concepts—such as power calculations—are essentially absent from the machine learning literature [77]. In this section, we emphasize that power should play a central role in any biomarker study, regardless of the analysis method selected.

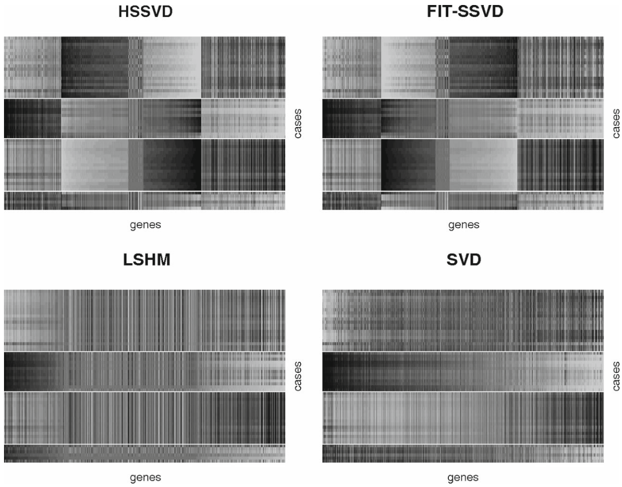

Fig. 1. Checkerboard plots produced by four different SVD biclustering methods on 12,625 genes from 56 patients with four levels of lung cancer [74].

The exact definition of power relevant to a biomarker study depends on the study's goal. Phase III trials incorporating predictive biomarkers revolve around one or more hypothesis tests that address whether the experimental treatment has a significant effect. In this setting, the traditional definition of power—the probability that, if the null hypothesis is truly false, it will be rejected—is appropriate. Trials evaluating a screening test based on a prognostic biomarker require a modified definition of power: the probability that the screening test correctly classifies a high proportion of patients, e.g. 90%. It is immediately apparent that both formulations of power are eminently desirable—without power, an unacceptably high number of results are likely to be false positives. Sample size is one of the drivers of power; in a typical study, investigators should aim for a sample size that enables at least 80% power.

Unacceptably high rates of false positives, or type I errors, threaten the validity of a study's conclusions. This concern is especially relevant when many hypotheses are tested simultaneously, as multiple comparisons inflate the type I error if they are not controlled for. Investigators can choose between many well-studied methods to limit type I error rate to a low level, conventionally 5%, in the presence of multiple tests, most commonly family-wise error rate (FWER) or false discovery rate (FDR). The FWER is the probability that we incorrectly reject even one true null hypothesis. FDR, meanwhile, is the expected proportion of falsely rejected hypotheses. FWER offers a stronger control than FDR, in the sense that if we control FWER at a certain level, we automatically also control FDR at that level. FWER and FDR may be preferred in different settings: FWER is used in many confirmatory studies [78], while FDR may be more logical in exploratory and other settings [79]. The most popular procedure for control of FWER is the Bonferroni correction [80]; despite this method's numerical simplicity, however, it is not recommended in light of Holm's step-down procedure [81], which is uniformly more powerful than the Bonferroni correction at the same level of control. The Benjamini-Hochberg step-up procedure offers control of FDR [82,83]. For a more detailed overview of multiple comparisons, we direct the reader to Dudoit and van der Laan's book [84].

We now present a general algorithm for calculating sample size in biomarker studies via simulation. The crux of the algorithm is the generation of *realistic data scenarios*, simulated datasets that incorporate information about the study design, method, and biomarkers in question. Information needed for all study designs includes the maximum allowable type I error rate, α, the multiple comparisons adjustment method, the desired power, β, the number of simulations to be run, B, an initial guess of the sample size, n_0, and the study's *minimal clinical measure of importance*. The specific measure of importance will vary by study. In a biomarker-stratified study, for instance, the measure of importance would be the expected change in effect size for the experimental treatment from the biomarker-positive to biomarker-negative groups. In an adaptive signature study, the measure of importance would be the change in effect size between overall and biomarker-specific groups, and additional necessary information would include the expected proportion of biomarkers that are true predictive biomarkers (likely below 1%). In a phase 4 prospective study of a screening test based on a prognostic biomarker, the measure of importance would be the misclassification proportion of the test compared to gold standard (e.g. 10%), and other necessary information would include the expected operating characteristics of the test, which may be based on information from a phase 3 prognostic biomarker study. While the details vary, the philosophy remains constant: investigators should not be excessively optimistic when generating realistic data scenarios. Most values should represent a worst-case scenario—e.g., the minimum effect size the investigators could observe and still conclude that a drug has a meaningful effect worth pursuing.

Algorithm 1. Power calculations through simulation

```
1 Run $B$ simulations adjusting for a type I error rate of at most $\alpha$, under a realistic data scenario with sample size $n_0$. This will entail simulating the correct number of biomarkers necessary for the study, under the assumption that they attain only the minimum clinical measure of importance.
2 Calculate the proportion of times the simulated biomarkers were detected and/or the hypotheses of interest were correctly rejected. This proportion is the estimated power of the test.
3 If the estimated power is $\epsilon$ less (or more) than the prespecified $\beta$, then increase (or decrease) $n_0$ by 1 and go back to step **1**.
4 Otherwise, stop the algorithm and the desired sample size is $n_0$.
```

3.5 Validation

When candidate biomarkers have been identified, they must be validated through an external dataset. At times, it is appropriate to consider truly exogenous data—data from a previous study that have evaluated the same outcome and biomarkers, for instance. At other times, this is infeasible or inappropriate; in these cases, researchers should plan to collect a secondary dataset. Researchers should consider the same issues listed above when determining the sample size for the validation set, but the set of assumptions should be less restrictive—namely, the number of biomarkers will be smaller and the proportion of biomarkers believed to be true will be higher. Algorithm 1 once again provides a convenient way to calculate sample size under these new assumptions.

3.6 Evaluation

In most real-world applications, it is not enough for a candidate biomarker to exist—it must also be useful. Once researchers have identified and validated a candidate biomarker, they can turn their attention to the issue of evaluating a biomarker's utility, whether that utility is in diagnosis, risk prediction, or any of a variety of functions in clinical practice. There are many statistical methods available for evaluating both the relationship between a biomarker and the disease area of interest and the usefulness of a biomarker when applied to specific populations; we outline only a few of these methods. Researchers should choose evaluation methods based on the specifics of their experiment while putting together their analysis plan, before data are collected. For a more in-depth introduction to some of the techniques mentioned in this section, see Pepe (2003) [85].

Measures of Accuracy. Biomarkers are often used in diagnostic medicine to classify patients as diseased or non-diseased. In this setting, evaluating the performance of a candidate biomarker is informed by the typical measures of accuracy for any diagnostic test. Let X be a candidate biomarker, and let $Y = 1\{X > c\}$ be an indicator equal to one when the biomarker exceeds a certain threshold, c, and zero otherwise. Suppose that a researcher intends to diagnose disease based on Y. Let D be an indicator of disease, so that D is equal to one for diseased subjects and zero otherwise. We define the **sensitivity** of Y as $se = \Pr(Y = 1|D = 1)$. The **specificity** of Y is $sp = \Pr(Y = 0|D = 0)$. Sensitivity is also referred to as the true positive fraction (TPF) and one minus specificity as the false positive fraction (FPF). Sensitivity and specificity give the probability of correct classification, conditional on disease status.

Two additional accuracy measures are the positive predictive value (PPV) and negative predictive value (NPV), given by $\text{PPV} = \Pr(D = 1|Y = 1)$ and $\text{NPV} = \Pr(D = 0|Y = 0)$. That is, the positive and negative predictive values give the probability of correct disease classification, conditional on test result. Although PPV and NPV are related to sensitivity and specificity, there is an important distinction between them. Sensitivity and specificity are functions of the test alone–they do not vary when the test is applied to different populations. Positive and negative predictive values, on the other hand, are highly dependent on the disease prevalence in the population the test is applied to: the same test may have wildly different positive and negative predictive values when applied to different populations. As such, the positive and negative predictive values cannot be estimated using data from studies that lack an estimate of disease prevalence, such as case-control studies. Accuracy measures are typically estimated empirically, and confidence regions for these measures can be constructed using the methods outlined in chapter 2 of Pepe (2003) [85].

Receiver Operating Characteristic Curves. Often, investigators will wish to evaluate a continuous biomarker. Receiver operating characteristic (ROC) curves are frequently used to do so, and they can be applied to a wide variety of tasks: comparing two biomarkers, constructing single-number summaries of biomarker performance, and selecting the screen positive threshold, c, among others. Roughly speaking, the ROC curve for a biomarker, X, is a plot of the true positive fractions and false positive fractions as functions of the threshold c. That is, if $\text{TPF}(c)$ and $\text{FPF}(c)$ are the true and false positive fractions for the test $1\{X > c\}$, respectively, then $\text{ROC}(\cdot) = \{(\text{FPF}(c), \text{TPF}(c)), -\infty < c < \infty\}$. If we let S_D and $S_{\bar{D}}$ denote the survival functions for X in the diseased and non-diseased populations, respectively, then the ROC curve can be equivalently defined as $\text{ROC}(t) = S_D(S_{\bar{D}}(t))$ for $0 < t < 1$. ROC curves boast well-developed theory, and the literature contains many procedures for ROC curve estimation and inference, both parametric and semiparametric [86–90]. Pepe (2000) [91] proposes a method to adjust ROC curves for covariates.

An advantage of the ROC framework is the ability to construct single number summaries that can be used to evaluate biomarkers. One such summary is the area under the ROC curve (AUC). A test that classifies perfectly yields AUC $= 1$, while a test that is no better than random chance has AUC$= 0.5$. If we let X_D and $X_{\bar{D}}$ denote observed biomarkers from the diseased and non-diseased populations, respectively, then AUC has an attractive interpretation as the probability of correctly ordered biomarkers—that is, $\text{AUC} = \Pr(X_D > X_{\bar{D}})$. Two candidate biomarkers can be compared by testing whether the corresponding AUCs differ, as in chapter 5 of Pepe (2003) [85].

Alternative Methods. While AUC is often a useful summary of a biomarker's performance, other methods may be more relevant for risk prediction models. Consider a risk prediction model that contains several biomarkers, and suppose we add a new biomarker to the model. Ware (2006) [92] observed that doing so may result in many subjects being placed in new risk categories, even if the change in AUC is small. Pencina et al. (2008) [93] propose using reclassification statistics to more adequately capture the effect of the new biomarker. Specifically, net reclassification improvement (NRI) measures how much an additional biomarker improves model-predicted probabilities of disease. When a strongly predictive biomarker enters the model, subjects with disease are reclassified into higher risk categories while subjects without disease are reclassified into lower risk categories, resulting in a large NRI. Pencina et al. (2011) [94] present a number of extensions to reclassification statistics.

Pencina et al. (2008) [93] propose a second alternative to AUC for evaluating biomarkers called integrated discrimination improvement (IDI). IDI measures how much a new biomarker increases the values of sensitivity and specificity, integrated over all possible thresholds. IDI also has a useful interpretation as the change in average TPF corrected for any increase in average FPF.

Predictive Biomarkers. The evaluation of predictive biomarkers requires special care. Predictive biomarkers are often evaluated by considering the interaction between the biomarker and treatment in a regression analysis [95,96]. While a strong interaction between a biomarker and treatment is consistent with the role of a predictive biomarker, it is not sufficient: predictive biomarkers are causal in nature, and causal evidence is needed to truly support them. The potential outcomes framework may help overcome this pitfall. Huang et al. (2012) [97] use the potential outcomes framework to evaluate a predictive biomarker under the assumption of monotone treatment effect, while Zhang et al. (2014) [98] propose a method that relaxes the assumption of monotonicity. For a general discussion of the evaluation of predictive biomarkers, see Polley et al. (2013) [99].

3.7 Reproducibility

"Non-reproducible single occurrences are of no significance to science."

- Karl Popper

Reproducibility is not just a criterion for a research study or manuscript to be accepted—it is a central, guiding tenet of the scientific method, an aim that every study should seek to attain. The Oxford English Dictionary defines reproducibility as "the extent to which consistent results are obtained when produced repeatedly." Applied to the setting of biomarker studies, reproducibility is the principle that experiments conducted under similar conditions should give similar results. Reproducibility is a crucial goal: if a study is not reproducible, its results will be difficult, and perhaps inappropriate, to apply to other settings.

In practice, to validate whether results can be reproduced, researchers may carry out a new experiment under similar conditions. The results of the new experiment can be compared to those of the original using a variety of methods, such as the Pearson correlation coefficient, the paired t-test, the intraclass correlation coefficient or the concordance correlation coefficient (CCC) [100]. When researchers wish to compare multiple outcomes, they should be careful to correct for multiple comparisons, as described in Sect. 3.4.

To clarify a subtle point: there is a distinction between reproducibility and replicability. An experiment need not be exactly replicated to qualify as reproducible–in fact, it may be very difficult to achieve perfect replication. A reproducible study's results can recur, or be reaffirmed, under similar but not identical settings–researchers at different labs testing the same biomarker, for instance [101]. A replicable study, on the other hand, would lead to identical results when it is conducted under the same conditions [102]. Whether strict replicability is necessary or not is currently the topic of some debate. To return to the biomarker example, Drummond [101] notes that such experiments can easily be affected by gene-deficient variants of the biomarkers in question, as these could lead independently synthesized gene segments to have significantly differing effects. As a consequence, it is unlikely for researchers to achieve a precise replication of the original experiment–but it may be unimportant to do so as long as the results of the experiments are consistent.

A natural question to ask is how many times an experiment should be repeated before it is considered trustworthy enough to publish. The answer to this question depends on expense and ethical considerations. Generally speaking, an experiment should be repeated several times before it is reported. This recommendation can be relaxed in some settings when it would prove overly restrictive, for instance when replication is excessively costly or when a repeated experiment's ethics would be questionable [102]. An illustration of the latter comes from the guidelines for experimentation on vertebrate animals, which discourages the use of unnecessary duplication. Casadevall [102] and Laine [103] suggest that, in order to make full use of each experiment, researchers "strive for reproducibility instead of simple replicability." For example, if the original experiment tests the efficacy of a drug on controlling glucose level in a certain time period, a second experiment could test for a dose-response relationship while simultaneously confirming the original conclusion of efficacy.

While external confirmation is the gold standard for establishing the reproducibility of a result, investigators can ensure substantial levels of reproducibility simply by choosing appropriate methods. The inferential properties of a statistical method, such as consistency and power (see Sect. 3.4), are directly connected to the reproducibility of that method's results. Large, well-designed studies, equipped with inferentially sound methods and powered appropriately for the questions of scientific interest, are inherently highly reliable. Investigators can save a great deal of time and effort by striving for such studies initially.

4 Clinical Interpretation

Ultimately, statistical validation and evaluation of a biomarker can only go so far. A biomarker's long-term worth depends heavily on its clinical interpretation and utility. Investigators should also take care not to dramatically depart from standardized, consistent definitions and properly executed methods. Inconsistencies across studies can hinder research producing robust conclusions.

In many instances, biomarkers are used as substitutes, or surrogates, for clinical endpoints. Surrogates may improve the feasibility of a trial through reduction in sample size or trial duration, and they are especially attractive when there are ethical concerns with the clinical endpoint, such as invasive procedures. Determining whether or not a biomarker is an appropriate surrogate endpoint relies on knowledge of the disease process and the causal pathways the biomarker lies in. However, disorders are often complex, clinically heterogeneous, and subject to large inter-individual variation, with the same disorder appearing drastically different at distinct points along the continuum between severe pathology and non-disease state. As such, investigators should be wary of jumping to belief of a causal link between a biomarker and a clinical outcome, even when they find a statistical and temporal association: the biomarker may affect a causal pathway present in only a small number of patients, or may not play a role in the relevant causal pathway at all [104]. Even when the biomarker is in the correct causal pathway, its effect on the biological process may be of insufficient size or duration to significantly affect the clinical outcome [104]. Hence, while surrogate biomarkers can be useful tools, relying solely on surrogates may lead to misleading conclusions and even harm to patients. For example, the use of ventricular premature depolarization (VPD) suppression as a primary outcome in clinical trials had led to the approval of antiarrhythmic drugs for patients with myocardial infarction [105]. Subsequent studies, however, found that, despite being effective in suppressing VPD, some antiarrhythmics actually increased mortality [106]. Prentice (1989) proposes an operational criterion to validate surrogate endpoints in clinical trials comparing two or more interventions: the surrogate endpoint should fully capture the net effect of the intervention on the clinical endpoint conditional on the surrogate endpoint [107]. The Prentice criterion is not universally accepted: some argue that it does not allow for valid inference on the effect of the intervention on the clinical endpoint [108].

Similarly, biomarkers that perform well in a narrow context may not be applicable to other settings [109]. Patient heterogeneity is among the most important factors to account for in biomarker studies; we recommend employing matching or stratification based on relevant characteristics, such as age, race/ethnicity, or body mass index, to help account for it. For example, there are consistent gender differences in patients with acute myocardial infarction: elevation of troponins are less common and lower in women than in men, while natriuretic peptides and C-reactive protein are more elevated in women than men [110]. The heterogeneity in symptoms between genders may contribute to the poorer prognosis of myocardial infarction in women, as well as demonstrating a facet of the "Yentl syndrome," the gender bias against women in the identification and management of coronary heart disease [111].

Finally, whether a biomarker is used for clinical prediction and screening should be based heavily on the benefits and risks involved. For example, the utility of prostate cancer screening that relies on the prostate-specific antigen (PSA) is controversial [112]. A substantial proportion of PSA-detected cancers are benign enough that they would not cause clinical problems during a man's lifetime. In these cases, the potential benefits of PSA testing may not outweigh the harms of the invasive diagnostic procedures and unnecessary treatment, including urinary, sexual, and bowel complications [112].

5 Limitations

This chapter was constructed to be general enough to appeal to a wide audience of researchers working in biomarker discovery. Moreover, we are attempting to provide an overall pipeline. As such, we could not address individual parts of this pipeline in sufficient detail, and we may have left out some important specifics. This section is meant primarily to serve as a safeguard for readers against known potential issues so they can avoid errors in advance.

While having a data management plan and protocol is necessary, it is rarely completely sufficient—issues always arise during data cleaning and management that were not anticipated by the plan. For this reason, we recommend having a research team member specializing in data management, whose expertise can help tackle unexpected issues.

While we have suggested several supervised and unsupervised methods for biomarker discovery, this review only scratches the surface of the sum total of methods available. Domain expertise is necessary to select appropriate methods; such knowledge is not provided in this chapter (due to limited space).

Power analysis was defined in very general terms, and we provided an all-purpose algorithm for its calculation through simulation. Our exposition may leave readers with the false impression that power calculations are an easy business with few complications. Nothing could be farther from the truth. Power calculations should, if possible, be left to a senior statistician well-versed in multiple comparisons, and adequate time should be allowed for them.

While sensitivity, specificity, and ROC curves are useful methods for biomarker evaluation, great care needs to be taken when specifying what magnitude of improvement is useful. Biomarkers might lead to an incremental improvement in the operating characteristics we have presented, but only at the expense of prohibitive cost. We recommend consulting with a physician with expertise on the particular application when considering this trade-off.

6 Conclusion

The rapid increase of available data—a process that is only accelerating—provides immense opportunities for improving the health of both individuals and populations. Fields that can harness "big data" to make health care decisions that take patient heterogeneity into account, as precision medicine seeks to do, have the potential to advance human health dramatically. However, the rise of "big data" presents not only opportunities, but a whole host of complications and challenges. The discovery and validation of biomarkers that can guide treatment decisions is more relevant than it has ever been, and methodologies that accomplish this task in a reliable, reproducible, and statistically rigorous way are of utmost necessity.

The discovery and validation of biomarkers is a complex process. Statistical issues are inherent to every step of the process, and they must be carefully considered as they are encountered. We propose the following master pipeline to help ensure the reproducibility of research related to biomarker discovery. For each step in the pipeline, we provide questions that researchers should answer affirmatively before proceeding.

1. Consider research goals to choose an appropriate design. *Does the study design reflect the role the biomarkers are expected to play in a clinical setting? Is the study design consistent with the current state of knowledge for the biomarkers being analyzed?*
2. Design data analysis plan. Consider both supervised and unsupervised statistical methods. *Are the method's assumptions consistent with the study design? Is the method suited to the research question at hand?*
3. Conduct power calculations to determine the appropriate sample size. *Did you correctly define a minimum clinical measure for calculating power? Is your minimum clinical measure chosen to represent a worst case scenario? Did the power calculation adjust for multiple comparisons? Did the power calculation incorporate prior knowledge effectively?*
4. Collect and curate data. *Are the data collected consistently and reliably? Are the data stored and linked in a way that respects patients' privacy?*
5. Conduct planned analyses and validate on an external data set. *Were the candidate biomarkers discovered in the initial analysis confirmed by the validation analysis?*
6. Evaluate the usefulness of the biomarker in practice. *Do the biomarkers offer a clinically relevant improvement in TPF and FPF?*
7. Consider clinical implications. *Is the cost of the proposed biomarker justified by its benefits?*

The above pipeline will need to be modified on a case-by-case basis, but it should provide a useful guide for any researcher and starting point for any study in the biomarker discovery field.

Many areas of research pertaining to the discovery of biomarkers are fervently active. The optimal approach for incorporating predictive biomarkers into modern, multi-stage study designs, such as SMARTs, is an area of open research, as is the proper use of electronic health record data and causal inference for observational and other non-randomized biomarker studies. The development of machine learning techniques with desirable inferential properties is an ongoing task, as is the use of these inferential properties to derive formal power calculations. Note that automatic approaches, such as many of the approaches described above, greatly benefit from "big data" with large training sets. However in some health informatics settings, we can be confronted with a small amount of data and/or rare events, where completely automatic approaches may suffer from insufficient training data. In these settings, interactive machine learning (iML) may be applicable, where a "doctor-in-the-loop" can help to refine the search space through heuristic selection of samples. Therefore, what would otherwise be an almost intractable problem, reduces greatly in complexity through the input and assistance of a human agent involved in the learning phase [113]. In all of these developments, proper attention to statistical considerations will enhance the ability of biomarker discovery studies to demonstrably improve clinical care through precision medicine.

References

1. Oldenhuis, C., Oosting, S., Gietema, J., De Vries, E.: Prognostic versus predictive value of biomarkers in oncology. Eur. J. Cancer **44**(7), 946–953 (2008)
2. National Institutes of Health: Precision Medicine Initiative Cohort Program (2016). Accessed 25 Feb 2016
3. Poste, G.: Bring on the biomarkers. Nature **469**(7329), 156–157 (2011)
4. Preedy, V.R., Patel, V.B.: General Methods in Biomarker Research and Their Applications. Springer, Netherlands (2015)
5. Novelli, G., Ciccacci, C., Borgiani, P., Papaluca Amati, M., Abadie, E.: Genetic tests and genomic biomarkers: regulation, qualification and validation. Clin. Cases Min. Bone Metab. **5**(2), 149–154 (2008)
6. Sun, Q., Van Dam, R.M., Spiegelman, D., Heymsfield, S.B., Willett, W.C., Hu, F.B.: Comparison of dual-energy x-ray absorptiometric and anthropometric measures of adiposity in relation to adiposity-related biologic factors. Am. J. Epidemiol. kwq306 (2010)
7. Flegal, K.M., Graubard, B.I.: Estimates of excess deaths associated with body mass index and other anthropometric variables. Am. J. Clin. Nutr. **89**(4), 1213–1219 (2009)
8. Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology: Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation **93**(5), 1043–1065 (1996)
9. Huikuri, H.V., Stein, P.K.: Heart rate variability in risk stratification of cardiac patients. Prog. Cardiovasc. Dis. **56**(2), 153–159 (2013)
10. Association, A.D., et al.: Standards of medical care in diabetes - 2015 abridged for primary care providers. Clin. Diab. **33**(2), 97–111 (2015)
11. Larsen, M.L., Hørder, M., Mogensen, E.F.: Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus. N. Engl. J. Med. **323**(15), 1021–1025 (1990)
12. Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., O'Callaghan, C.J., Tu, D., Tebbutt, N.C., Simes, R.J., Chalchal, H., Shapiro, J.D., Robitaille, S., et al.: K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N. Engl. J. Med. **359**(17), 1757–1765 (2008)
13. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., et al.: A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science **266**(5182), 66–71 (1994)
14. Lee, M., Shen, H., Huang, J.Z., Marron, J.: Biclustering via sparse singular value decomposition. Biometrics **66**(4), 1087–1095 (2010)
15. Pepe, M.S., Etzioni, R., Feng, Z., Potter, J.D., Thompson, M.L., Thornquist, M., Winget, M., Yasui, Y.: Phases of biomarker development for early detection of cancer. J. Natl Cancer Inst. **93**(14), 1054–1061 (2001)
16. Sargent, D.J., Conley, B.A., Allegra, C., Collette, L.: Clinical trial designs for predictive marker validation in cancer treatment trials. J. Clin. Oncol. **23**(9), 2020–2027 (2005)

17. Freidlin, B., McShane, L.M., Korn, E.L.: Randomized clinical trials with biomarkers: design issues. J. Natl Cancer Inst. **102**(3), 152–160 (2010)
18. Simon, R.: Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized Med. **7**(1), 33–47 (2010)
19. Mandrekar, S.J., Sargent, D.J.: Clinical trial designs for predictive biomarker validation: one size does not fit all. J. Biopharm. Stat. **19**(3), 530–542 (2009)
20. Jiang, W., Freidlin, B., Simon, R.: Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J. Natl Cancer Inst. **99**(13), 1036–1043 (2007)
21. Freidlin, B., Simon, R.: Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin. Cancer Res. **11**(21), 7872–7878 (2005)
22. Murphy, S.A.: An experimental design for the development of adaptive treatment strategies. Stat. Med. **24**(10), 1455–1481 (2005)
23. Denny, J.C.: Mining electronic health records in the genomics era. PLoS Comput. Biol. **8**(12), e1002823 (2012)
24. Society for Clinical Data Management, I: Good Clinical Data Management Practices (2005). Accessed 25 Feb 2016
25. Bruza, P.D., Van der Weide, T.P.: The semantics of data flow diagrams. University of Nijmegen, Department of Informatics, Faculty of Mathematics and Informatics (1989)
26. U.S. Department of Health & Human Services: HIPAA Administrative Simplification (2013). Accessed 25 Feb 2016
27. Wei, S., Kosorok, M.R.: Latent supervised learning. J. Am. Stat. Assoc. **108**(503), 957–970 (2013)
28. Chapelle, O., Schölkopf, B., Zien, A., et al.: Semi-supervised learning (2006)
29. Kosorok, M.R.: What's so special about semiparametric methods? Sankhya. Ser. B [Methodol.] **71**(2), 331–353 (2009)
30. Wei, L.: The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat. Med. **11**(14–15), 1871–1879 (1992)
31. Altstein, L., Li, G.: Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model. Biometrics **69**(1), 52–61 (2013)
32. Hastie, T., Tibshirani, F.: The Elements of Statistical Learning (2001)
33. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
34. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. Roy. Stat. Soc.: Ser. B (Methodol.) **70**(1), 53–71 (2008)
35. Tibshirani, R., et al.: The lasso method for variable selection in the Cox model. Stat. Med. **16**(4), 385–395 (1997)
36. Bien, J., Taylor, J., Tibshirani, R.: A lasso for hierarchical interactions. Ann. Stat. **41**(3), 1111 (2013)
37. Bondell, H.D., Krishna, A., Ghosh, S.K.: Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics **66**(4), 1069–1077 (2010)
38. Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R.: Fixed and random effects selection in mixed effects models. Biometrics **67**(2), 495–503 (2011)
39. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
40. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**(1), 266–298 (2010)
41. Zhu, R., Zeng, D., Kosorok, M.R.: Reinforcement learning trees. J. Am. Stat. Assoc. **110**(512), 1770–1784 (2015)

42. Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., Initiative, A.D.N., et al.: Random forest-based similarity measures for multi-modal classification of Alzheimer‘s disease. NeuroImage **65**, 167–175 (2013)
43. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)
44. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
45. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint (2012). arXiv:1207.0580
46. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al.: The human splicing code reveals new insights into the genetic determinants of disease. Science **347**(6218), 1254806 (2015)
47. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
48. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM (1992)
49. Smola, A., Vapnik, V.: Support vector regression machines. Advances in Neural Information Processing Systems, vol. 9, pp. 155–161 (1997)
50. Zhao, Y., Kosorok, M.R., Zeng, D.: Reinforcement learning design for cancer clinical trials. Stat. Med. **28**(26), 3294–3315 (2009)
51. Zhao, Y., Zeng, D., Socinski, M.A., Kosorok, M.R.: Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics **67**(4), 1422–1433 (2011)
52. Vansteelandt, S., Joffe, M., et al.: Structural nested models and G-estimation: The partially realized promise. Stat. Sci. **29**(4), 707–731 (2014)
53. Robins, J.: A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. Math. Model. **7**(9), 1393–1512 (1986)
54. Robins, J.M.: The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Health Service Res. Methodol.: A Focus on AIDS **113**, 159 (1989)
55. Witteman, J.C., D'Agostino, R.B., Stijnen, T., Kannel, W.B., Cobb, J.C., de Ridder, M.A., Hofman, A., Robins, J.M.: G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. Am. J. Epidemiol. **148**(4), 390–401 (1998)
56. Robins, J.M., Blevins, D., Ritter, G., Wulfsohn, M.: G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology **3**, 319–336 (1992)
57. Zhao, Y., Zeng, D., Rush, A.J., Kosorok, M.R.: Estimating individualized treatment rules using outcome weighted learning. J. Am. Stat. Assoc. **107**(449), 1106–1118 (2012)
58. Qian, M., Murphy, S.A.: Performance guarantees for individualized treatment rules. Ann. Stat. **39**(2), 1180–1210 (2011)

59. Zhou, X., Mayer-Hamblett, N., Khan, U., Kosorok, M.R.: Residual weighted learning for estimating individualized treatment rules. J. Am. Stat. Assoc., October 2015
60. Xu, Y., Yu, M., Zhao, Y.Q., Li, Q., Wang, S., Shao, J.: Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics **71**(3), 645–653 (2015)
61. Zhao, Y.Q., Zeng, D., Laber, E.B., Kosorok, M.R.: New statistical learning methods for estimating optimal dynamic treatment regimes. J. Am. Stat. Assoc. **110**(510), 583–598 (2015)
62. Su, X., Meneses, K., McNees, P., Johnson, W.O.: Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. J. Roy. Stat. Soc. C (Appl. Stat.) **60**(3), 457–474 (2011)
63. Zhang, B., Tsiatis, A.A., Laber, E.B., Davidian, M.: A robust method for estimating optimal treatment regimes. Biometrics **68**(4), 1010–1018 (2012)
64. Tian, L., Alizadeh, A.A., Gentles, A.J., Tibshirani, R.: A simple method for estimating interactions between a treatment and a large number of covariate. J. Am. Stat. Assoc. **109**(508), 1517–1532 (2014)
65. Cox, D.: Regression models and life tables (with discussion). J. Roy.Stat. Soc, B **34**, 187–220 (1972)
66. Kalantar-Zadeh, K., Kopple, J.D., Regidor, D.L., Jing, J., Shinaberger, C.S., Aronovitz, J., McAllister, C.J., Whellan, D., Sharma, K.: A1C and survival in maintenance hemodialysis patients. Diab. Care **30**(5), 1049–1055 (2007)
67. Kyan, M., Muneesawang, P., Jarrah, K., Guan, L.: Unsupervised Learning: A Dynamic Approach. IEEE Press Series on Computational Intelligence, pp. 275–276
68. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
69. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(2579–2605), 85 (2008)
70. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
71. Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. U.S.A. **102**(21), 7426–7431 (2005)
72. Shabalin, A.A., Weigman, V.J., Perou, C.M., Nobel, A.B.: Finding large average submatrices in high dimensional data. Ann. Appl. Stat. **3**(3), 985–1012 (2009)
73. Tan, K.M., Witten, D.M.: Sparse biclustering of transposable data. J. Comput. Graph. Stat. **23**(4), 985–1008 (2014)
74. Chen, G., Sullivan, P.F., Kosorok, M.R.: Biclustering with heterogeneous variance. Proc. Natl. Acad. Sci. **110**(30), 12253–12258 (2013)
75. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Inform. **2**, 59–78 (2006)
76. Swan, A.L., Mobasheri, A., Allaway, D., Liddell, S., Bacardit, J.: Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. OMICS **17**(12), 595–610 (2013)
77. Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**(6), 321–332 (2015)
78. Bender, R., Lange, S.: Adjusting for multiple testing? when and how? J. Clin. Epidemiol. **54**(4), 343–349 (2001)

79. Glickman, M.E., Rao, S.R., Schultz, M.R.: False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J. Clin. Epidemiol. **67**(8), 850–857 (2014)
80. Westfall, P.H., Young, S.S.: Resampling-based multiple testing: examples and methods for p-value adjustment, vol. 279. John Wiley & Sons, New York (1993)
81. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
82. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc.: Ser. B (Methodol.) **57**(1), 289–300 (1995)
83. Efron, B.: Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, vol. 1. Cambridge University Press, Cambridge (2012)
84. Van der Laan, M.J.: Multiple Testing Procedures with Applications to Genomics. Springer Series in Statistics. Springer, Heidelberg (2008)
85. Pepe, M.S.: The statistical evaluation of medical tests for classification and prediction. Oxford University Press, USA (2003)
86. Pepe, M.S.: A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. Biometrika **84**(3), 595–608 (1997)
87. Cai, T., Pepe, M.S.: Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. J. Am. Stat. Assoc. **97**(460), 1099–1107 (2002)
88. Chrzanowski, M.: Weighted empirical likelihood inference for the area under the ROC curve. J. Stat. Plan. Infer. **147**, 159–172 (2014)
89. Cai, T., Dodd, L.E.: Regression analysis for the partial area under the ROC curve. Statistica Sin. **18**, 817–836 (2008)
90. Cai, T., Moskowitz, C.S.: Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. Biostatistics **5**(4), 573–586 (2004)
91. Pepe, M.S.: An interpretation for the ROC curve and inference using GLM procedures. Biometrics **56**(2), 352–359 (2000)
92. Ware, J.H.: The limitations of risk factors as prognostic tools. N. Engl. J. Med. **355**(25), 2615–2617 (2006)
93. Pencina, M.J., D'Agostino, R.B., Vasan, R.S.: Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat. Med. **27**(2), 157–172 (2008)
94. Pencina, M.J., D'Agostino, R.B., Steyerberg, E.W.: Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat. Med. **30**(1), 11–21 (2011)
95. Gail, M., Simon, R.M.: Testing for qualitative interactions between treatmenteects and patient subsets. Biometrics **41**(2), 361–372 (1985)
96. Russek-Cohen, E., Simon, R.M.: Evaluating treatments when a gender by treatment interaction may exist. Stat. Med. **16**(4), 455–464 (1997)
97. Huang, Y., Gilbert, P.B., Janes, H.: Assessing treatment-selection markers using a potential outcomes framework. Biometrics **68**(3), 687–696 (2012)
98. Zhang, Z., Nie, L., Soon, G., Liu, A.: The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. Ann. Appl. Stat. **8**(4), 2336 (2014)
99. Polley, M.Y.C., Freidlin, B., Korn, E.L., Conley, B.A., Abrams, J.S., McShane, L.M.: Statistical and practical considerations for clinical evaluation of predictive biomarkers. J. Natl. Cancer Inst. **105**(22), 1677–1683 (2013)
100. Lawrence, I., Lin, K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**, 255–268 (1989)

101. Drummond, C.: Replicability is Not Reproducibility: Nor is it Good Science (2009)
102. Casadevall, A., Fang, F.C.: Reproducible science. Infect. Immun. **78**(12), 4972–4975 (2010)
103. Laine, C., Goodman, S.N., Griswold, M.E., Sox, H.C.: Reproducible research: moving toward research the public can really trust. Ann. Intern. Med. **146**(6), 450–453 (2007)
104. Fleming, T.R., DeMets, D.L.: Surrogate end points in clinical trials: are we being misled? Ann. Intern. Med. **125**(7), 605–613 (1996)
105. Connolly, S.J.: Use and misuse of surrogate outcomes in arrhythmia trials. Circulation **113**(6), 764–766 (2006)
106. Weir, M., Investigators, C.A.S.T., et al.: The cardiac arrhythmia suppression trial investigators: Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. Cardiopul. Phys. Ther. J. **1**(2), 12 (1990)
107. Prentice, R.L.: Surrogate endpoints in clinical trials: definition and operational criteria. Stat. Med. **8**(4), 431–440 (1989)
108. Berger, V.W.: Does the prentice criterion validate surrogate endpoints? Stat. Med. **23**(10), 1571–1578 (2004)
109. Strimbu, K., Tavel, J.A.: What are biomarkers? Curr. Opin. HIV AIDS **5**(6), 463 (2010)
110. Sbarouni, E., Georgiadou, P., Voudris, V.: Gender-specific differences in biomarkers responses to acute coronary syndromes and revascularization procedures. Biomarkers **16**(6), 457–465 (2011)
111. Healy, B.: The yentl syndrome. N. Engl. J. Med. **325**(4), 274–276 (1991)
112. Hoffman, R.M.: Screening for prostate cancer. N. Engl. J. Med. **365**(21), 2013–2019 (2011)
113. Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016)