
UTILISING HIGH-DIMENSIONAL DATA IN RANDOMISED CLINICAL TRIALS: A REVIEW OF METHODS AND PRACTICE

A PREPRINT

Svetlana Cherlin^{1,*} Theophile Bigirimurame¹ Michael J Grayling¹ Jérémie Nsengimana¹

Luke Ouma¹ Aida Santaolalla² Fang Wan³ S Faye Williamson¹ James M S Wason¹

May 18, 2023

¹Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

²Translational Oncology & Urology Research Group, Centre for Cancer, Society & Public Health, King's College London, UK

³Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

*svetlana.cherlin@newcastle.ac.uk

ABSTRACT

Introduction: Even in effectively conducted randomised trials, the probability of a successful study remains relatively low. With recent advances in the next-generation sequencing technologies, there is a rapidly growing number of high-dimensional data, including genetic, molecular and phenotypic information, that have improved our understanding of driver genes, drug targets, and drug mechanisms of action. The leveraging of high-dimensional data holds promise for increased success of clinical trials.

Methods: We provide an overview of methods for utilising high-dimensional data in clinical trials. We also investigate the use of these methods in practice through a review of recently published randomised clinical trials that utilise high-dimensional genetic data. The review includes articles that were published between 2019 and 2021, identified through the *PubMed* database.

Results: Out of 174 screened articles, 100 (57.5%) were randomised clinical trials that collected high-dimensional data. The most common clinical area was oncology (30%), followed by chronic diseases (28%), nutrition and ageing (18%) and cardiovascular diseases (7%). The most common types of data analysed were gene expression data (70%), followed by DNA data (21%). The most common method of analysis (36.3%) was univariable analysis. Articles that described multivariable analyses used standard statistical methods. Most of the clinical trials had two arms.

Discussion: New methodological approaches are required for more efficient analysis of the increasing amount of high-dimensional data collected in randomised clinical trials. We highlight the limitations and barriers to the current use of high-dimensional data in trials, and suggest potential avenues for improvement and future work.

Keywords Genetic data, High-dimensional information, Precision medicine, Randomised clinical trials, Statistical analysis

1 Introduction

Randomised controlled trials (RCTs) are the gold standard for assessing the safety and efficacy of an experimental treatment. However, despite the growing cost and time associated with developing and evaluating drugs, the probability of success of RCTs is relatively low.[1] One of the reasons is that there is rarely a "one size fits all" approach in most clinical areas because treatment typically has a heterogeneous effect on patients with different pathogenic mechanisms. For example, in a study that investigated predictors of response to Methotrexate in early rheumatoid arthritis, [2]

75% of the patients experienced a good response rate, according to the EULAR response criteria.[3] This study found that several demographic and clinical characteristics (including age, sex, smoking status and symptom duration) are associated with response to Methotrexate. Subsequently, a double-blind phase IV clinical trial in patients with rheumatoid arthritis identified genetic markers that could partly explain the heterogeneity of response to Methotrexate.[4]

With recent advances in the next-generation sequencing technologies, there is a rapidly growing number of human molecular biomarkers that could inform drug mechanisms and increase the success of clinical trials. [5] Molecular biomarkers are measurable molecular characteristics (small molecules) that could identify relatively homogeneous disease subsets in terms of clinical features, diagnosis, prognosis, or response to treatment. With the advent of personalised medicine, molecular biomarkers are gaining importance in clinical research.[6] The most common types of molecular biomarkers are genomic biomarkers such as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Single nucleotide polymorphisms (SNPs), which are the most abundant type of genetic variation, represent a difference in a single nucleotide. SNPs are often measured (genotyped) across the genome, and the associations between genome-wide SNPs and different human traits, i.e. genome-wide association studies (GWAS), are extensively used in genetics.[7] GWAS to date have analysed hundreds of thousands of genetic variants generated by next-generation sequencing technologies.

Proteomics and metabolomics also play an important role in many medical applications and are being increasingly used in drug research and development.[8] Proteomics is a study of molecules in proteins that allows characterisation of protein structure and function. Protein biomarkers are also increasingly used in clinical trials in patient stratification, disease diagnosis, and prognosis.[9] Another commonly used genetic biomarker is gene expression, which is a process that regulates the amount of protein or other molecules expressed by the cell, and thus is measured by the amount of the molecules or protein. The advantage of microarray technology is to allow for gene expression profiling which consists of measuring levels of thousands of genes. Changes in gene expression can reflect the change in a cell's environment, such as disease state,[10] response to treatment [11] or treatment side effect.[12] Metabolites, small molecules produced by the body when it breaks down food or drugs, are useful for biomarker discovery because they can be utilised to examine the underlying biochemical activity of cells. Modern technologies, such as mass spectrometry, allow for a large number of metabolites to be measured thus creating a metabolomic profile. [13] Metabolic changes are informative of the response to treatment and therefore have the potential to be useful in clinical trials.[14] For example, a randomised placebo-controlled clinical trial that examined the effect of sertraline on major depressive disorder patients found that baseline metabolic signatures could be predictive of response or non-response to sertraline.[15]

In clinical trials, biomarkers serve multiple purposes, such as prognosis of the likely progression of a disease, and prediction of the likely clinical outcome.[16] Prognostic biomarkers are those that are associated with disease prognosis in the absence of treatment or in the presence of a standard of care treatment. Predictive biomarkers are those that are associated with the effectiveness of a specific treatment. Predictive biomarkers could be used to identify subsets of patients who are likely to respond to treatment. For example, a pooled analysis of randomised trials found that women whose breast tumours have overexpressed the human epidermal growth factor receptor 2 (*HER2*) protein or amplified *HER2* gene (*HER2*-positive) benefited from adjuvant treatment with anthracyclines, while women with *HER2*-negative breast tumours derived no added benefits from adjuvant chemotherapy with anthracyclines.[17] Thus, the *HER2* status of a breast tumour is a predictive biomarker for response to adjuvant treatment with anthracyclines. Prognostic and predictive biomarkers are usually measured once, before the start of treatment. Biomarkers that are measured repeatedly during the trial could be used as a surrogate endpoint, i.e. as a proxy for a clinical endpoint. Biomarkers based on a continuous single gene measurement can be used as classifiers by considering a threshold, or a series of thresholds, to specify a biomarker-positive and biomarker-negative group [18, 19]. However, identifying single-gene biomarkers requires knowledge and biological interpretation of the disease pathway, which may not always be available.

Recent advances in whole genome biotechnology allow for measuring multiple genetic variants during clinical trials.[20, 21, 22, 23] This allows biomarkers across multiple genes to be developed, i.e. biomarkers based on high-dimensional data. A variety of predictive and prognostic biomarkers based on high-dimensional molecular profiling have been proposed in oncology.[24, 25, 26] These biomarkers are especially relevant for finding potential responders to a treatment in settings where an assay for identifying biomarker-positive patients is not yet available.[27] While prognostic biomarkers based on high-dimensional data are becoming increasingly available, predictive biomarkers based on high-dimensional data are rare due to the challenge of understanding a treatment's mechanism of action.[28] Additional challenges of using high-dimensional data are identifying which biomarkers to include in the model, and how to effectively/appropriately combine the individual biomarkers.[29]

In this paper, we provide an overview of several statistical methods for utilising high-dimensional data in the analysis of RCTs. We also present a review of recently published clinical trials that utilised high-dimensional data to investigate how often various methods have been used in practice.

2 Overview of methods for utilising high-dimensional data in clinical trials

In this section, we describe statistical methods used for analysing high-dimensional data in RCTs; many of which have been implemented in standard statistical software such as R. [30] A summary of these methods is provided in Table 1. When considering suitability of the methods, it is important to distinguish between testing for association and prediction. Association tests, such as the Chi-Square test, can shed light on the biological processes by providing better understanding of the phenomenon in question. Association tests are useful for testing hypotheses about the differences between the groups of observations, such as the difference between the treatment arms, or for finding biomarkers that are associated with response to treatment.

In prediction analysis, statistical models such as regression are applied to data in order to build predictors that could be applied to future studies. The quality of prediction should be assessed on an independent dataset using some measure of the discrepancy between the observed and predicted outcomes.

Some of the methods we review in this manuscript focus on either testing for association or on prediction, while others focus on both. However, it is important to note that models that have high power to detect associations do not necessarily have high predictive power.[31]

2.1 Notation

In this section we describe a two-arm RCT where participant i ($i = 1, \dots, n$) is randomised to either an intervention arm ($t_i = 1$) or control arm ($t_i = 0$). For each participant i , a set of $j = 1, \dots, m$ biomarkers, x_{ij} , are collected, and an outcome y_i is measured. Regression modelling is often used to model the outcome y_i as a function of the covariates x_{ij} , which are measurable quantities related to the outcome. For different types of outcome, different types of regression are used. The most common types are linear regression (for continuous outcomes), logistic regression (for binary outcomes) and Cox regression (for time-to-event outcomes). Linear regression models the mean of the continuous outcome, assuming that the outcome is normally distributed. Logistic regression models the log odds, $\text{logit}(p_i) = \log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right)$, where $P(Y_i = 1)$ denotes the probability of a successful outcome. Cox regression models the hazard ratio of an event at time t , $\log\left(\frac{h_i(t)}{h_{0i}(t)}\right)$, where $h_{0i}(t)$ is the baseline hazard at time t . In these regression models, the link function of the response variable connects the covariates with the expected value of the outcome variable in a linear way, while the covariates are being weighted by their coefficients. The null hypothesis of a specific coefficient being zero represents testing for an effect of the corresponding covariate.

2.2 Univariable approach

A univariable approach consists of testing a single biomarker's relationship to a response variable. In linear regression, the outcome y_i for patient i takes the form

$$y_i = \beta_{j0} + \beta_{j1}t_i + \beta_{j2}x_{ij} + \beta_{j3}t_ix_{ij} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ is the error term. In logistic regression, the probability of the outcome y_i for patient i takes the form:

$$\text{logit}(p_i) = \beta_{j0} + \beta_{j1}t_i + \beta_{j2}x_{ij} + \beta_{j3}t_ix_{ij}.$$

In Cox regression, the hazard ratio of an event at time t for patient i takes the form:

$$\log\left(\frac{h_i(t)}{h_{0i}(t)}\right) = \beta_{j1}t_i + \beta_{j2}x_{ij} + \beta_{j3}t_ix_{ij}.$$

The null hypothesis $H_{j2} : \beta_{j2} = 0$ represents testing for a prognostic effect of biomarker j , while the null hypothesis $H_{j3} : \beta_{j3} = 0$ represents testing for a predictive effect of biomarker j . These hypotheses could then be tested using a Wald test, for example.

Applying statistical tests to one biomarker at a time could result in an inflated number of false positives, due to multiple independent comparisons.[32] To prevent this, the Bonferroni correction [33] is often applied, which adjusts the significance level of individual tests to level α/m , where m is the number of tests and α is the desired family-wise

Table 1. Summary, advantages and disadvantages of methods utilising high-dimensional data.

Method	Summary	Advantages	Disadvantages
Univariable approach	Testing one biomarker at a time	Simplicity.	Multiple testing issue
Multivariable approach	Testing a number of biomarkers simultaneously	Fitting a single model for a several biomarkers	Overfitting
Penalised approach	Penalises regression coefficients, causing them to shrink, maybe to zero	Prevention of overfitting	Tuning of parameters
Random forests	Collection of regression or classification trees	Allows modelling non-linear interactions	Lack of intuitive interpretation
Support vector machines	Building a classifier by fitting a hyperplane between different groups of observations	Allows modelling non-linear interactions	Computational complexity
Cluster analysis	Grouping data based on a measure of similarity	Allows modelling non-linear interactions	Sensitivity to outliers
Gene sets and networks	Undirected graphs representing associations between the genes	Dimensionality reduction	Computational complexity
Principal component analysis	Transforming high-dimensional data into low-dimensional variables that account for most of the original data's variation	Allows modelling non-linear interactions	Lack of intuitive interpretation of the principal components
Adaptive signature design	Constructing a low-dimensional score from high-dimensional data	Finding group of patients benefiting from treatment	Multiple testing issue

error rate. To reduce multiple testing burden, a two-step procedure has been proposed [34] that accounts for correlation between the biomarkers via penalised regression. In the first stage of the procedure, a screening test selects a subset of biomarkers, and in the second stage, only the selected biomarkers are tested for interaction.

An additional challenge in detecting interactions is due to the large sample size required to obtain high power.[35, 36] In the case of a binary biomarker, in which the trial population can be divided into biomarker-positive and biomarker-negative subgroups, the sample size for testing a null hypothesis of no interaction is at least four times higher than the sample size needed to test the main effect (see Appendix).

Univariable analysis models are straightforward to fit and produce intuitive results. However, in the real world there is often more than just one biomarker involved. Analysing one biomarker at a time ignores the correlation between the biomarkers, which could lead to incorrectly concluding that some biomarkers are predictive.

2.3 Multivariable approach

A multivariable regression takes into account two or more biomarkers. Similarly to the univariable regression, there are three commonly used regression types: linear (for continuous outcomes), logistic (for binary outcomes) and Cox regression (for time-to-event outcomes), which take the following form when m biomarkers are simultaneously adjusted for:

$$y_i = \beta_{j0} + \beta_1 t_i + \sum_{j=1}^m \beta_{j2} x_{ij} + \sum_{j=1}^m \beta_{j3} t_i x_{ij} + \epsilon_i,$$

$$\text{logit}(p_i) = \beta_{j0} + \beta_1 t_i + \sum_{j=1}^m \beta_{j2} x_{ij} + \sum_{j=1}^m \beta_{j3} t_i x_{ij}, \text{ and}$$

$$\log \left(\frac{h_i(t)}{h_{0i}(t)} \right) = \beta_1 t_i + \sum_{j=1}^m \beta_{j2} x_{ij} + \sum_{j=1}^m \beta_{j3} t_i x_{ij},$$

respectively. Multivariable analysis estimates the contribution of each biomarker x_{ij} while adjusting for the effect of other biomarkers or covariates. Therefore, unlike univariable analyses, it takes into account correlation between biomarkers.

The main drawback of the multivariable approach is the large number of parameters that may be included. With high-dimensional data, this approach can lead to a model with more parameters than observations (i.e. the ‘‘curse of dimensionality’’). In this case, multivariable linear regression cannot be used because the unique ordinary least squares estimators of the regression coefficients are not defined. To reduce the complexity of the model, several variable

selection approaches have been proposed, including machine learning approaches (discussed below). However, a large number of parameters in the model could still lead to overfitting, which is the phenomenon of modelling the observed data too precisely so that it captures the noise in the data. In this case, the model shows an inferior performance when applied to a new dataset. To reduce the potential effects of overfitting, a rule-of-thumb is that at least ten events are required per variable in logistic and Cox regression models, though this rule is often debated.[37] For linear regression estimated using ordinary least squares, the number of covariates that can be included in the model is generally higher; it has been shown that two subjects per value would be sufficient for adequate estimation of regression coefficients.[38]

2.4 Regularised (penalised) regression

Regularised, or penalised, approaches penalise models by shrinking the estimates of the regression coefficients. Suppose a regression model with a $(m + 1)$ -dimensional vector of covariates $\beta = (\beta_0, \beta_1, \dots, \beta_m)^T$ is fitted by maximising the log-likelihood function $\ell(\beta)$. In penalised regression, $\ell(\beta)$ is maximised subject to a penalty function $P(\beta)$ and a regularisation parameter λ , that is, $\hat{\beta} = \operatorname{argmax}[\ell(\beta) - \lambda P(\beta)]$. As a result, the regression coefficient estimate $\hat{\beta}$ is shrunk towards zero in comparison to the maximum likelihood estimate, with λ controlling the amount of shrinkage.

The method induces different degrees of sparsity, depending of the type of penalty used. For example, the Least Absolute Shrinkage and Selection Operator (LASSO) regression [39] allows shrinkage of the coefficients to zero by penalising the model with $P(\beta) = \|\beta\|_{\ell_1} = \sum_{j=1}^m |\beta_j|$ and is therefore a sparse method which allows for variable selection. Another type of penalised regression is ridge regression [40] in which the penalty function has the form $P(\beta) = \|\beta\|_{\ell_2} = \sum_{j=1}^m \beta_j^2$. Ridge regression shrinks the coefficients *towards* zero, however it does not shrink them to zero. Elastic net [41] is a type of penalised regression in which both penalties are used, i.e.

$$\hat{\beta} = \operatorname{argmax} \left[\ell(\beta) - \lambda \left(\eta \sum_{j=1}^m |\beta_j| + \frac{1 - \eta}{2} \sum_{j=1}^m \beta_j^2 \right) \right].$$

The combination of the penalties is controlled by a penalty weight parameter η . When $\eta = 1$, the elastic net is identical to LASSO, whereas when $\eta = 0$ it is identical to ridge. Elastic net combines setting of the coefficients to zero using LASSO and shrinking of the coefficients using ridge, to improve the model's performance.

A penalised logistic regression model, which included ten genes, was used to predict the overall complete pathologic response rate in a phase II genomic study of ixabepilone as neoadjuvant treatment for breast cancer.[42] A pharmacogenetic study used ridge regression to predict a response to treatment.[11] It has been found that using LASSO regression improved the accuracy of the treatment effect estimator in a RCT.[43] A review of neoadjuvant clinical trials in breast cancer that analysed gene expression data [44] found that penalised methods outperform competing methods when applied to estrogen receptor-positive (ER+) early breast cancer patients treated with neoadjuvant aromatase inhibitor letrozol. However, an application of a penalised high-dimensional Cox model to an early breast cancer RCT of chemotherapy with or without adjuvant trastuzumab resulted in highly variable expected survival probabilities with very large confidence intervals.[45]

Group-lasso [46] is a special case of LASSO that performs selection of important groups of variables. For example, the groups could represent specific biological pathways of the biomarkers, or variables that reflect a specific aspect of a treatment. Extending the group-lasso by considering interactions,[47] however, can result in many false positive interactions for high-dimensional problems.

Penalised regression requires optimisation of the penalty parameter, which could be done using cross-validation. In the cross-validation procedure, a model is fitted to a subset of the data and its accuracy is assessed on a different subset of the data. The process is repeated multiple times with different partitions of the data for fitting (training subset) and assessing (testing subset). Parameters that lead to the best accuracy are chosen. However, when cross-validation is used to examine model performance, tuning of the parameters requires nested cross-validation, in which the inner cross-validation (for tuning of parameters) is encapsulated inside the outer cross-validation (for assessing model performance). This procedure requires large sample sizes. It is also necessary to ensure homogeneous partitioning of the data with respect to important features, in order to achieve a valid cross-validation procedure [48].

2.5 Machine learning approaches

Machine learning is a class of algorithms that analyse data based on existing (training) data.[49] Machine learning algorithms can either be supervised or unsupervised, with the difference being the labelling of the input data. In supervised machine learning algorithms such as classification, the training data is labelled, while in unsupervised methods such as clustering, the training data is not labelled. Supervised approaches are used for predictive modelling

when the classification of the training data is known in advance, and the trained algorithm is used to predict or classify new data with unknown classification, such as response or non-response to treatment in clinical trials. Unsupervised methods are used for feature selection problems, such as identifying a predictive biomarker in the context of biomarker analysis, and dimensionality reduction [50].

2.5.1 Random forests

Random forests are a type of high-dimensional nonparametric model aimed at prediction,[51] and therefore belong to the class of supervised machine learning algorithms. They are represented as a collection of regression trees (for a continuous outcome) or classification trees (for a binary outcome). Each tree is a decision model that consists of a recursive partitioning of a dataset into subsets that are determined by a randomly selected group of input variables. The subsets are homogeneous with respect to the group of variables. At each node of a tree, different groups of variables might be used. Random forests are formed by trees constructed from training datasets sampled with replacement from the original dataset. The remaining samples form the testing datasets and are used for assessing prediction accuracy. For example, the probability of misclassifying an observation could be used as a measure of prediction accuracy. Random forests are flexible in that regression and classification trees can incorporate non-linear interactions between the variables.[52]

Traditional random forests are designed for one treatment group and are therefore suitable for prognostic, rather than predictive, purposes. A few adaptations of the method for more than one treatment group have been developed that facilitate identification of a subset of patients who benefit from the treatment. For example, the “Virtual Twins” method [53] is a random forest-based method of identifying a subgroup of enhanced treatment effect by incorporating treatment-covariate interactions.

A variation of the random forest has been developed, that uses a measure based on a difference in survival times as an alternative to the accuracy prediction, for deciding on a best possible split.[54] When applied to a phase III RCT with high-dimensional SNP data, this approach has been shown to outperform a univariable analysis. The challenges of this method include specifying model parameters, such as the number of trees in the forest.

2.5.2 Support vector machines

Support vector machines (SVM) are a supervised machine learning method for building a classifier that can be used to account for non-linear relationships between variables.[55] SVM assign an observation to a specific category, or class, by fitting a hyperplane between the samples from different classes so that the distance between the hyperplane to the nearest sample is maximised. This distance is maximised using support vectors, i.e. data points that are closer to the hyperplane. SVM involve transforming the data using a kernel function to allow linear separation of the data. An advantage of SVM is that it can effectively incorporate high-dimensional data that can be noisy and/or correlated. It has been widely applied to classification problems using high-dimensional biomarkers.[56, 57, 58, 59, 60] SVM could be used in RCTs if treatment-covariate interaction effects are introduced into the feature space of SVM. Using SVM constructed from the combination of brain imaging and demographic and clinical biomarkers, a group of Mild Cognitive Impairment patients who were most likely to cognitively decline has been identified.[61] Limitations of SVM include their computational complexity, especially the need to optimise their parameters.

2.6 Cluster analysis

Clustering methods are unsupervised methods of grouping data based on some measure of similarity, so that the observations in each group are similar (but dissimilar to those in other groups). The most common measure of similarity between the observations is correlation. Traditional clustering methods include hierarchical clustering and partitioning.[62] In hierarchical clustering, the data is organised into a tree-shape structure (a dendrogram) constructed from hierarchical series of nested clusters, while partitioning does not assume hierarchical relationships between clusters. An example of partitioning is k -means clustering, which partitions the data into a pre-specified number k of mutually exclusive groups so that the the sum of the squared distances between the members of the group and the means of the clusters is minimised.[63] Another example is Partitioning Around Medoids clustering, which is similar to the k -means but is more robust to outliers.[64]

Hierarchical clustering employs agglomerative and divisive strategies. Hierarchical agglomerative clustering starts by treating each sample as a separate cluster and then merges the most similar clusters together. This process is repeated iteratively until all samples are clustered. Hierarchical divisive clustering starts by treating all the observation as one cluster, and then recursively splits the cluster into two, until the desired number of clusters is obtained. Hierarchical clustering could be used to analyse genes that are differentially expressed between different experimental conditions, such as the different treatment groups in clinical trials. To estimate the number of clusters in the dataset, consensus

clustering could be used which utilises bootstrapping to classify each observation multiple times. Finally, observations are assigned to the cluster with the highest consensus score and the number of clusters is derived from objective metrics.[65] Other methods, called model-based clustering, exploit the same idea of making clustering robust to model misspecification and estimation of the number of clusters. They assume that observations follow a mixture of distributions rather than belonging to discrete classes.

Clustering is often used in gene expression analysis because it simplifies visualisation and allows one to trace specific biological pathways.[66, 67, 68, 69] Moreover, it can be used to identify specific disease subtypes. For example, hierarchical clustering was able to identify pre- and post-vaccine samples in a study of the effect of an influenza vaccination on gene expression. [70]

2.7 Gene sets and networks

Gene networks belong to the class of the unsupervised machine learning algorithms. They are undirected graphs with nodes representing genes and edges representing gene-gene associations. Genes with similar co-expression patterns are then grouped into modules using clustering techniques. Different types of co-expression networks are discussed elsewhere.[71, 72]

Weighted gene co-expression network analysis (WGCNA) [73] is a common co-expression network method that is used for finding clusters of highly correlated genes. It summarises the clusters using the representative gene (the eigengene), thus performing dimensionality reduction. The eigengene is a vector that represents the expression of all the genes in the model. WGCNA has been used to analyse metabolites in an ancillary study of vitamin D supplementation for the prevention of asthma.[74] The eigenvalues for the modules of metabolites were used to find association with asthma.

Gene set enrichment (GSE) is another subtype of gene networks that clusters genes into pre-defined sets that share common biological functions, and summarises the gene expressions into a single score for each set. Scores represent the extent of the differences in gene expression between the phenotypic classes of interest, for example tumours that are responsive or non-responsive to treatment. Testing the statistical significance of the scores allows detection of an enrichment signal.[75] GSE analysis has been used to compare advanced colorectal cancer subtypes in a RCT of first-line treatment of metastatic colorectal cancer.[76] Gene networks are useful for dimensionality reduction of a large number of correlated genes. To our knowledge, this method has not been used for comparing treatment arms or finding predictive biomarkers in clinical trials.

2.8 Principal component analysis

Principal component analysis (PCA) is a statistical technique that provides information on directions of variability in data. PCA consists of transforming high-dimensional data into a lower-dimensional set of variables (principal components) such that the first principal component (PC) is associated with the largest source of variation, the second PC with the largest remaining source of variation and so on. The procedure of computing the PCs involves computing the eigenvalues and eigenvectors for the covariance matrix of standardised data. PCs are formed by transforming the original data using a matrix constructed from the eigenvectors.[77]

Each PC is constructed as a linear combination of the original high-dimensional data in such a way that the PCs are mutually uncorrelated. Thus, PCs could prevent multicollinearity issues in regression models and be very useful for correlated biomarkers. Once computed, the PCs can be used as covariates in linear regression models, as well as a dimensionality reduction technique for clustering. PCA also makes high-dimensional data more suitable for visualisation. For example, PCs are widely used to identify genetic variation associated with geographic region, [78] with most geographic variation explained by the first two PCs. However, it has been shown that in the analysis of gene expression data, many more PCs might be needed to detect relevant variability, depending on the sample sizes and effect sizes. [79]

A challenge of PCA is the interpretation of the PCs, as well as identifying the most informative PCs. In the field of clinical trials, the use of PCA is limited to finding prognostic rather than predictive biomarkers.

2.9 Adaptive signature design and risk scores

Adaptive signature design methods utilise high-dimensional data to construct a low-dimensional (or scalar) signature. They combine information from multiple genetic markers to create a signature that could be used for diagnostic, prognostic or predictive purposes. Adaptive signatures are motivated by the fact that genetics play an important role in the heterogeneity of disease progression and response to treatment, and could therefore be used to facilitate personalised medicine. The original adaptive signature design constructed a low-dimensional signature based on the interaction

between the treatment and the high-dimensional baseline biomarker data. [27, 80] It was developed for situations with no pre-defined predictive biomarker and utilised a threshold on the number of biomarkers included in the signature. Initially, two non-overlapping groups of trial participants have been used to develop and validate the signature.[27] while later a cross-validation has been implemented, which uses patient information more efficiently. [80]

A few studies [81, 82, 83, 84] construct a signature as a sum of the effects of the interactions between the treatment and each of the covariates separately. In these methods, the adaptive signature is represented by a single score for each patient. Specifically, for a binary outcome, a single covariate logistic model is fitted for each biomarker $j = 1, \dots, m$ as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1 t_i + \beta_{j2} x_{ij} + \beta_{j3} t_i x_{ij},$$

where p_i is the probability of the outcome of interest, β_{j2} represents a prognostic effect of biomarker j , and β_{j3} represents a predictive effect of biomarker j .

A risk score for patient i (RS_i) is computed as the sum of the maximum likelihood estimate of the treatment-covariate interaction coefficients $\hat{\beta}_{j3}$ weighted by the value of the biomarker x_{ij} , i.e.

$$RS_i = \sum_{j=1}^m \hat{\beta}_{j3} x_{ij}.$$

The collection of risk scores RS_i for all i could be subdivided in different ways to represent different strata of patients in terms of the predicted treatment benefit. [83, 84] At the end of the trial, a test is performed for the overall comparison between the arms, as well as for the comparison between the arms in the subgroup, using an α -splitting approach to control the type I error rate. Alternatively, they could be used as covariates to test for an association with the outcome.[82]

Adaptive signature designs often utilise a combination of the previously described approaches. For example, an adaptive signature which is predictive of response to MAGE-A3 immunotherapeutic in patients with metastatic melanoma has been developed and validated in a randomised phase II trial [85] using a variation of PCA and hierarchical clustering. Another phase II trial [86] used a combination of a scoring system and the penalised approach. For each patient i , the following risk score was constructed that represented the hazard ratio under the two treatments on the logarithmic scale:

$$RS_i = \log[h_0(t|\mathbf{x}_i)] - \log[h_1(t|\mathbf{x}_i)],$$

where $h_j(t|\mathbf{x}_i)$ is the hazard rate for treatment $j = \{0, 1\}$ for patient i , and \mathbf{x}_i is the vector of gene expressions for patient i . The hazard functions were estimated with penalised Cox regression.

The adaptive signature designs are applied in a post-hoc manner, i.e. they identify the subgroup of patients at the end of the trial and therefore do not fit the classical definition of an adaptive design. Rather, they are adaptive in the sense that they allow adaptive selection of patient subgroups. For example, an adaptive signature design has been proposed that finds the optimal subgroup in terms of maximising the power for identifying treatment benefit. [87] To address the issue of adaptive changes in trials, the risk scores-based adaptive signature has been utilised in the adaptive enrichment framework, where the trial population is adaptively enriched with patients who are predicted to benefit from the treatment.[88]

In summary, adaptive signature designs have the advantage of improving the efficiency of clinical trials by identifying enhanced benefit subgroups. More reliably identifying patient subgroups who benefit from the treatment would prevent the situation in which a potentially effective treatment is disregarded because the treatment effect in the overall population is overlooked. Moreover, adaptive signature designs have the potential to avoid patients who receive no benefit from receiving the treatment, thus preventing unnecessary exposure to possible side effects. However, adaptive signature designs come with a statistical challenge of a multiple comparisons issue. Additionally, there may be a need for dimensionality reduction in situations with a large number of baseline biomarkers. [89]

3 Current use of methods for utilising high-dimensional data in RCTs

3.1 Review methods

We performed the following literature search of RCTs using the *PubMed* database:

```
("gene expression" OR "nucleotide*"
OR "*omic*" OR "genetic signature"
OR "SNPs") AND (trial[Title/Abstract])
AND ((ffrft[Filter]) AND
```



```
(randomizedcontrolledtrial[Filter])
AND (2019/5/1:2021/5/1[pdat]))
```

This search, performed on June 2021, covers publication of RCTs between May 1st 2019 and May 1st 2021, with at least one of the terms: “gene expression”, “nucleotide”, “omics”, “genetic signature” or “SNPs”, appearing in the title or abstract. We included full-text articles published in English. The search identified 174 papers which were screened for eligibility.

After preliminary screening of titles and abstracts, eight reviewers (SC, TB, MJG, JN, LO, FW, SFW, JMSW) independently assessed the full text of relevant publications for final inclusion.

Papers were deemed eligible if they described RCTs that collected high-dimensional data. Here, data variables refer to biological variables collected at randomisation that could be used for comparing between the treatment arms or stratifying patients. For the purpose of this review, we adopted a flexible definition of high-dimensionality with respect to the number of variables. Specifically, we included studies containing at least 10 variables as they could benefit from methods suitable for high-dimensional data. An additional study that analysed 7 SNPs was included in this review as it used a multivariable approach.

We analysed the type of high-dimensional data (e.g. DNA, gene expression, etc.), the number of covariates used, purpose of collecting high-dimensional data, method of analysis of high-dimensional data, clinical area, and number of treatment arms. See the Supplementary Materials for the full summary of extracted data.

3.2 Results

Out of the 174 papers returned, 100 (57.5%) met the inclusion criteria. A summary of the data extracted from included articles is given in Table 2.

Most of the articles were for clinical trials in oncology (30%) and various chronic diseases (28%), including liver, kidney, rheumatic and respiratory diseases. Other clinical areas included nutrition and ageing (18%) and cardiovascular diseases (7%).

The majority of articles (70%) analysed gene expression data. The second most common type of data analysed was DNA data (21%), including genome-wide SNP data. Five percent of the articles analysed metabolomic data, protein data and data from questionnaires. Four percent of the articles analysed multiple types of data.

We divided the number of analysed covariates into four categories: “<10”, “10–100”, “101–1000”, and “>1000”. A large proportion of the analysed articles (41%) had 10–100 covariates available for analysis. A similar proportion of articles (38%) used >1000 covariates in their analysis. Fewer studies (20%) had 101–1000 covariates for the analysis, and one study had seven covariates, thus falling into the “<10” category.

The methods used in the analyses and their advantages are summarised in Table 1. 42% of the studies used multiple methods of analysis. The most common analysis technique was a univariable analysis (36.3%), followed by a multivariable analysis (17.5%), cluster analysis (11.25%) and PCA (10.6%). Other methods that were reported included: gene networks, multiple correspondence analysis, [90] penalised approaches and risk scores, Shannon entropy and Simpson index, [91, 92] significance analysis of microarrays, [93] single sample predictor classifier, [94] SVM.

Most trials had two arms (79%), followed by three-arm (17%) and four-arm trials (4%). In this review, we only analysed RCTs and therefore single-arm studies have been excluded.

The purpose of collecting high-dimensional data varied substantially between trials and was often not reported clearly. For those trials where it was reported, categorisation of the reasoning proved challenging. Some trials used high-dimensional data as the (primary or secondary) outcome by analysing the effect of the intervention on gene expression, for example. In some cases, high-dimensional data was used to explore predictive biomarkers or to compare treatment arms. In other cases, the prognostic properties of the high-dimensional data were investigated, i.e. they did not compare the treatment arms but analysed the data as if it were observational.

Figures 1-3 show the distribution of different types of data, methods of analysis and clinical areas, respectively, stratified by the number of covariates. With regards to data types, most studies used gene expression or DNA and had 10-100 covariates, 100-1000 covariates or >1000 (Figure 1). Regarding analysis methods, most studies using univariable and multivariable approaches utilised 101-1000 covariates, while gene sets and networks, clustering, and PCA most commonly used 10-100 covariates (Figure 2). In oncology, chronic diseases, and nutrition and ageing, the most common number of variables was 10-100; a substantial number of studies across all clinical areas analysed a larger number of covariates (101-1000 and >1000, Figure 3).

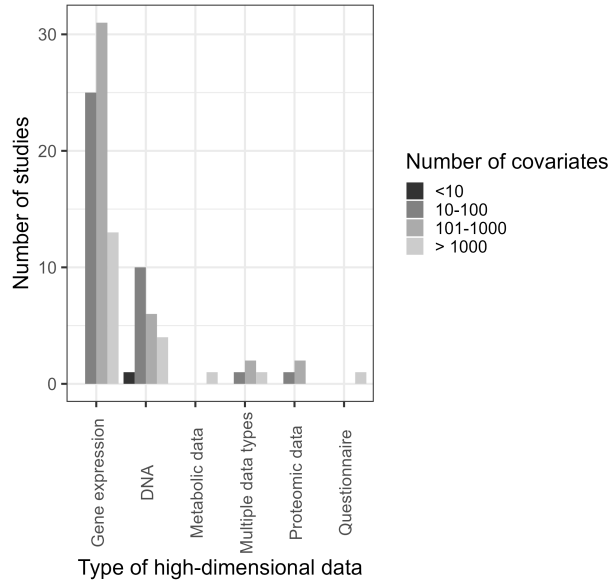


Figure 1. Number of covariates per type of high-dimensional data.

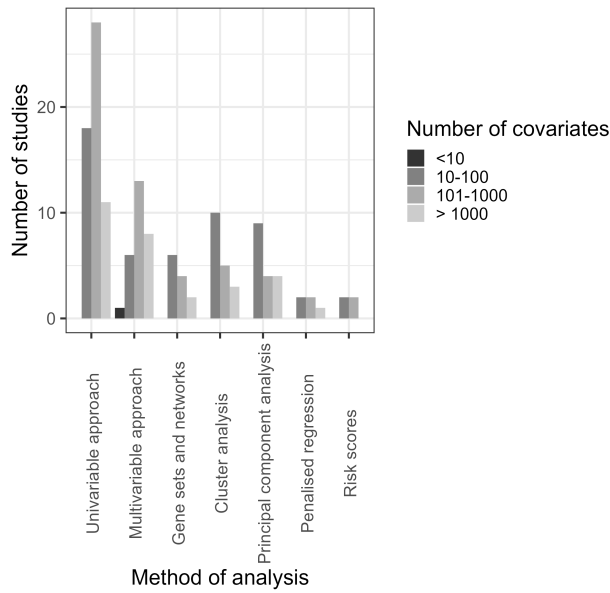


Figure 2. Number of covariates per method of analysis.

Table 2. Summary of extracted data. The denominator used to compute the percentages is 100 (number of eligible papers) unless specified. The most common answers appear in bold.

Question	Answer	<i>n</i> (%)
Type of high-dimensional data	Gene expression	70 (70%)
	DNA	21 (21%)
	Metabolomic data	1 (1%)
	Multiple data types ¹	4 (4%)
	Proteomic data	3 (3%)
Number of covariates used	Questionnaire	1 (1%)
	<10	1 (1%)
	10–100	41 (41%)
	101–1000	20 (20%)
Method of analysis ²	>1000	38 (38%)
	Univariable approach	58 (36.3%)
	Multivariable approach	28 (17.5%)
	Gene sets and networks	12 (7.5%)
	Cluster analysis	18 (11.25%)
	Principal component analysis	17 (10.6%)
	Penalised regression	5 (3.1%)
	Risk scores	4 (2.5%)
	Not stated	2 (1.25%)
	Other ³	16 (10%)
Clinical area	Oncology	30 (30%)
	Chronic diseases	28 (28%)
	Nutrition and ageing	18 (18%)
	Cardiovascular diseases	7 (7%)
	Other ⁴	17 (17%)
Number of treatment arms	2	79 (79%)
	3	17 (17%)
	4	4 (4%)

¹ Questionnaires, omics data, biochemical characteristics and laboratory parameters.

² The denominator used to compute these percentages is 160, because 42 (42%) studies used multiple methods of analysis.

³ Functional analysis, Shannon entropy and Simpson index, significance analysis of microarrays, single sample predictor classifier, SVM.

⁴ HIV, malaria, mental health, neuropathy, ophthalmology.

4 Discussion

In this paper, we provided an overview of methods for analysing high-dimensional data collected in clinical trials. We also reviewed 100 recently published articles reporting RCTs that utilised high-dimensional data to identify which methods are typically used in practice. Although we focused on high-dimensional genetic data, the methods described could be applied to other types of high-dimensional data, such as questionnaires, imaging data or data from wearable technologies.

In our search, gene expression and DNA data were the most common data analysed, covering a combined total of 91% of the high-dimensional data types included. A majority of the articles collected a large number of genetic data (>1000 variables), which reflects the progress in high-throughput technologies and highlights the need for increased uptake of more sophisticated methods to utilise the high-dimensional data efficiently.

Although most of the trials we reviewed had two-arms, over 20% had three or four arms. This reflects the additional complexity and challenges of utilising high-dimensional data in conjunction with multi-arm trials. Most clinical trials in this review were in the areas of oncology (30%) and several chronic diseases (28%). One of the challenges of trials for chronic diseases is learning how best to treat patients in the long-term. In particular, different treatments might be used for patients at different disease stages. Therefore, efficient methods that utilise changes in high-dimensional data over time are needed, for example methods that utilise longitudinal modelling.

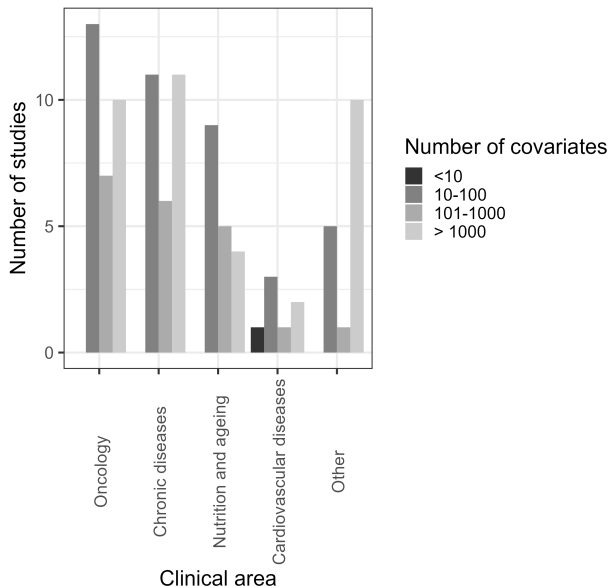


Figure 3. Number of covariates per clinical area.

Although we found some examples of more sophisticated methods being used to analyse high-dimensional data, the majority implemented straightforward approaches to examine interactions, such as univariable analysis. Methods such as machine learning, penalised approaches and risk scores appeared rarely in the analysis. For example, LASSO was seldom used despite being widely studied and having advantages. Adaptive signature design was not used. In some studies, high-dimensional data were measured, but only a small proportion of it was analysed. Therefore, there is strong potential for much more efficient use of high-dimensional data.

We investigated the distribution of the number of covariates across data types, methods of analysis and clinical areas. The number of covariates varied widely in each of these settings, highlighting the need for developing methods that would be applicable to data of different orders of magnitude.

High-dimensional data was collected for a variety of reasons, from being the primary outcome to identifying prognostic biomarkers in exploratory analysis to investigating biological pathways. However, few studies used high-dimensional data to compare treatments or to identify predictive biomarkers, which highlights a gap and presents an opportunity to use the data more effectively.

The limited use of sophisticated methods could be explained by perceived complexities and limitations of using high-dimensional data in clinical trials. Firstly, high-dimensionality of the data still requires *a priori* knowledge of the disease mechanism, in the form of existing disease classification, to efficiently reduce the dimensionality of the data.[24] Secondly, there may be a discrepancy between the signature constructed from genetic data and its biological meaning, which obscures the intuitive interpretation of high-dimensional data. For example, it has been found that a large number of breast cancer signatures constructed from a variety of gene sets do not explain the biological mechanism of the disease.[95] In oncology, the most common field that collected high-dimensional data according to this review, this leads to genetic signatures being rarely used in clinical trials. It has been suggested that incorporating different types of omics data and using standardised methodology has the potential to make more effective use of high-dimensional data in clinical trials in order to improve patient outcomes. [96] In this review, we have only described the methods that were used in the analysed studies. Alternative methods, such as Bayesian classifiers, [97] also have the potential to analyse high-dimensional data in clinical trials.

In conclusion, although we only used a single database and limited timelines, we show that an increasing number of clinical trials are collecting high-dimensional data. Many of them could benefit from implementing more sophisticated analysis methods, such as those outlined in this manuscript. Further research is needed to make full use of the high-dimensional data collected in RCTs.

A Appendix

Consider a hypothetical randomised placebo-controlled clinical trial of n participants with a normally distributed outcome $N(\mu_0, \sigma^2)$ for the control arm, and $N(\mu_1, \sigma^2)$ for the experimental arm. The number of participants in each group is the same ($n/2$). We would like to test $H_0 : \delta = 0$ where $\delta = \mu_1 - \mu_0$. A Wald statistic to test H_0 would be

$$W = \frac{\hat{\delta}}{\sqrt{\frac{4\sigma^2}{n}}}.$$

For a two-sided α significance level, the sample size n required for power $1 - \beta$ is

$$n = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{\delta^2}.$$

Now suppose we have a binary biomarker that divides the population into biomarker-positive and biomarker-negative patients, with r being the proportion of biomarker-positive patients. We assume that the treatment effect is $\delta_+ = \mu_{1+} - \mu_{0+}$ in biomarker-positive patients, and $\delta_- = \mu_{1-} - \mu_{0-}$ in biomarker-negative patients. The treatment-biomarker interaction effect, $\delta_+ - \delta_-$, could be estimated by

$$\begin{aligned} \hat{\delta}_+ - \hat{\delta}_- &= (\hat{\mu}_{1+} - \hat{\mu}_{0+}) - (\hat{\mu}_{1-} - \hat{\mu}_{0-}) \\ &\sim N\left(\delta_+ - \delta_-, \frac{\sigma^2}{rn} + \frac{\sigma^2}{rn} + \frac{\sigma^2}{(1-r)n} + \frac{\sigma^2}{(1-r)n}\right). \end{aligned}$$

A Wald statistic to test $H_0 : \delta_+ - \delta_- = 0$ would be

$$W_{int} = \frac{\hat{\delta}_+ - \hat{\delta}_-}{\sqrt{\frac{\sigma^2}{rn} + \frac{\sigma^2}{rn} + \frac{\sigma^2}{(1-r)n} + \frac{\sigma^2}{(1-r)n}}} = \frac{\hat{\delta}_+ - \hat{\delta}_-}{\sqrt{\frac{4\sigma^2}{nr(1-r)}}}.$$

For a two-sided α significance level, the sample size n_{int} required for power $1 - \beta$ is

$$n_{int} = \frac{4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{(\delta_+ - \delta_-)^2 r(1-r)}.$$

Thus, $n_{int} = \frac{n}{r(1-r)}$, i.e the sample size required to detect treatment-biomarker interaction increases by factor $\frac{1}{r(1-r)}$, with $\min\left\{\frac{r}{1(1-r)}\right\} = 4$ for $r = 0.5$. Therefore, the sample size for detecting the treatment-biomarker interaction is at least four times higher than the sample size needed to detect the main treatment effect.

B Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

C Funding

This research was supported by the Medical Research Council (MR/S014357/1). JMSW, LO and SC are funded by the National Institute for Health and Care Research (NIHR301614).

References

- [1] Wong CH and Siah KW. Estimation of clinical trial success rates and related parameters. *Biostatistics* 2019; 20: 273–286. DOI:10.1093/biostatistics/kxx069.
- [2] Saevarsdottir S et al. Predictors of response to methotrexate in early DMARD naïve rheumatoid arthritis: results from the initial open-label phase of the SWEFOT trial. *Ann Rheum Dis* 2011; 70: 469–475. DOI: 10.1136/ard.2010.139212.

- [3] van Gestel S et al. Development and validation of the european league against rheumatism response criteria for rheumatoid arthritis: Comparison with the preliminary american college of rheumatology and the world health organization/international league against rheumatism criteria. *Arthritis Rheumatol* 1995; 39: 39–40. DOI:10.1002/art.1780390105.
- [4] Aslibekyan S et al. Genetic variants associated with methotrexate efficacy and toxicity in early rheumatoid arthritis: results from the treatment of early aggressive rheumatoid arthritis trial. *Pharmacogenomics J* 2014; 14: 48–53. DOI:10.1038/tpj.2013.11.
- [5] Nelson MR et al. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015; 47: 856–862. DOI:10.1038/ng.3314.
- [6] Buyse M et al. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn* 2011; 11: 171–182. DOI: 10.1586/ERM.10.120.
- [7] Visscher PM et al. 10 years of GWAS discovery: Biology, function, and translation. *Am J Hum Genet* 2017; 101: 5–22. DOI:10.1016/j.ajhg.2017.06.005.
- [8] Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* 2016; 15: 473–484. DOI:10.1038/nrd.2016.32.
- [9] He T. Implementation of proteomics in clinical trials. *Proteomics Clin Appl* 2019; 13: 1800198. DOI: 10.1002/prca.201800198.
- [10] Emilson V et al. Genetics of gene expression and its effect on disease. *Nature* 2008; 452: 423–428. DOI: 10.1038/nature06758.
- [11] Geeleher P et al. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 2014; 15R47. DOI:10.1186/gb-2014-15-3-r47.
- [12] Duffy A et al. Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Sci Adv* 2020; 6: eabb6242. DOI:10.1126/sciadv.abb6242.
- [13] Shaham-Niv S et al. Metabolite medicine offers a path beyond lists of metabolites. *Commun Chem* 2021; 4: 225. DOI:10.1038/s42004-021-00551-w.
- [14] Kaddurah-Dauok R et al. Metabolomic signatures for drug response phenotypes: pharmacometabolomics enables precision medicine. *Clin Pharm Therap* 2015; 98: 71–75. DOI:10.1002/cpt.134.
- [15] Kaddurah-Dauok R et al. Pretreatment metabotype as a predictor of response to sertraline or placebo in depressed outpatients: a proof of concept. *Transl Psychiatry* 2011; 1: e26–e26. DOI:10.1038/tp.2011.22.
- [16] Antoniou M et al. Biomarker-guided adaptive trial designs in phase II and phase III: A methodological review. *PLoS One* 2016; 11(2): p.e0149803. DOI:10.1371/journal.pone.0149803.
- [17] Gennari A et al. HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials. *J Natl Cancer Inst* 2008; 100: 14–20. DOI:10.1093/jnci/djm252.
- [18] Jiang W et al. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst* 2007; 99: 1036–1043. DOI:10.1093/jnci/djm022.
- [19] Simon R. Development and validation of biomarker classifiers for treatment selection. *Stat Plan Inference* 2014; 138:39: 308–320. DOI:10.1016/j.jspi.2007.06.010.
- [20] Hu C and Dignam JJ. Biomarker-driven oncology clinical trials: Key design elements, types, features, and practical considerations. *JCO Precis Oncol* 2019; 1: 1–12. DOI:10.1200/PO.19.00086.
- [21] Johnson DH et al. Genome-wide association study of atazanavir pharmacokinetics and hyperbilirubinemia in AIDS clinical trials group protocol A5202. *Pharmacogenet Genomics* 2014; 24: 195–203. DOI:10.1097/FPC.000000000000034.
- [22] Wanda V et al. Genome-wide association study of tenofovir pharmacokinetics and creatinine clearance in AIDS clinical trials group protocol A5202. *Pharmacogenet Genomics* 2015; 25: 450–461. DOI:10.1097/FPC.000000000000156.
- [23] Wei Y et al. Confident identification of subgroups from SNP testing in RCTs with binary outcomes. *Biome J* 2022; 64: 256–271. DOI:10.1002/bimj.202000170.
- [24] Theilhaber J et al. Construction and optimization of gene expression signatures for prediction of survival in two-arm clinical trials. *BMC Bioinform* 2020; 21: 333. DOI:10.1186/s12859-020-03655-7.
- [25] Ye Y et al. Identification of a multidimensional transcriptome prognostic signature for lung adenocarcinoma. *J Clin Lab Anal* 2020; 33: e22990. DOI:10.1002/jcla.22990.

- [26] Li H et al. Development of a novel transcription factors-related prognostic signature for serous ovarian cancer. *Sci Rep* 2021; 11: 7207. DOI:10.1038/s41598-021-86294-z.
- [27] Freidlin B and 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* 2005; 11: 7872–7878. DOI: 10.1158/1078-0432.CCR-05-0605.
- [28] Simon R. Biomarker based clinical trial design. *Chin Clin Oncol* 2014; 3(3):39. DOI:10.3978/j.issn.2304-3865.2014.02.03.
- [29] Johnstone I and Titterton D. Statistical challenges of high-dimensional data. *Phil Trans R Soc A* 2009; 367: 4237–4253. DOI:10.1098/rsta.2009.0159.
- [30] R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing. Vienna, Austria 2021; DOI:https://www.R-project.org/.
- [31] Shmueli G. To explain or to predict? *Stat Sci* 2010; 25: 289–310. DOI:10.1214/10-STS330.
- [32] Herzog MH et al. *The Multiple Testing Problem*. Cham: Springer International Publishing. ISBN 978-3-030-03499-3, 2019. pp. 63–66. DOI:10.1007/978-3-030-03499-3_5.
- [33] Bland JM and Altman DG. Multiple significance tests: the Bonferroni method. *Br Med J* 1995; 310: 170. DOI:10.1136/bmj.310.6973.170.
- [34] Wang J et al. Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. *Biometrics* 2021; 15: 1–10. DOI:10.1111/biom.13424.
- [35] Brankovic M et al. Understanding of interaction (subgroup) analysis in clinical trials. *Eur J Clin Invest* 2019; 49: e13145. DOI:10.1111/eci.13145.
- [36] Brookes ST et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004; 57: 229–236. DOI:10.1016/j.jclinepi.2003.08.009.
- [37] Grant SW et al. Statistical primer: multivariable regression considerations and pitfalls. *Eur J Cardiothorac Surg* 2019; 55: 179–185. DOI:10.1093/ejcts/ezy403.
- [38] Austin PC and Steyerberg EW. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol* 2015; 68: 627–636. DOI:10.1016/j.jclinepi.2014.12.014.
- [39] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996; 56: 267–288.
- [40] Cessie SL and Houwelingen JCV. Ridge estimator in logistic regression. *J R Stat Soc C* 1992; 41: 191–201. DOI:10.2307/2347628.
- [41] Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; 67: 301–320. DOI:10.1111/j.1467-9868.2005.00503.x.
- [42] Baselga J et al. American society of clinical oncology: Phase II genomics study of Ixabepilone as neoadjuvant treatment for breast cancer. *J Clin Oncol* 2009; 27(4): 526–534. DOI:10.1200/JCO.2007.14.2646.
- [43] Bloniarz A et al. Lasso adjustments of treatment effect estimates in randomized experiments. *PNAS* 2016; 113: 7383–7390.
- [44] Ternès N et al. Statistical methods applied to omics data: predicting response to neoadjuvant therapy in breast cancers. *Curr Opin Oncol* 2014; 26: 576–583. DOI:10.1097/CCO.000000000000134.
- [45] Ternès N et al. Robust estimation of the expected survival probabilities from high-dimensional Cox models with biomarker-by-treatment interactions in randomized clinical trials. *BMC Med Res Methodol* 2017; 17: 576–583. DOI:10.1186/s12874-017-0354-0.
- [46] Yuan M and Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc B* 2006; 68: 49–67. DOI:10.1111/j.1467-9868.2005.00532.x.
- [47] Lim M and Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat* 2015; 24: 627–652. DOI:10.1080/10618600.2014.938812.
- [48] Krstajic D et al. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* 2014; 6, 10. DOI:10.1186/1758-2946-6-10.
- [49] Jordan MI and Mitchell TM. Machine learning: trends, perspectives, and prospects. *J Biopharm Stat* 2015; 349: 255–260. DOI:10.1126/science.aaa8415.
- [50] Wei Y et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials* 2021; 22:537. DOI:10.1186/s13063-021-05489-x.

- [51] Brieman L. Random forests. *Mach Learn* 2001; 45: 5–32. DOI:10.1023/A:1010933404324.
- [52] Reif DM et al. Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination. *Genes Immun* 2009; 10: 112–119.
- [53] Foster JC et al. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; 30: 2867–2880. DOI:10.1002/sim.4322.
- [54] Ubels J et al. RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics* 2020; 36(26): i601–i609. DOI:10.1093/bioinformatics/btaa799.
- [55] Vapnik V. *The nature of statistical learning theory*. New York: Springer, 1995. ISBN 978-1-4757-3264-1.
- [56] Hua S and Sung Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 2001; 17: 721–728. DOI:10.1093/bioinformatics/17.8.721.
- [57] Dror G et al. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics* 2005; 21: 879–901. DOI:10.1093/bioinformatics/bti132.
- [58] Liu J et al. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet* 2006; 2: e29. DOI:10.1371/journal.pgen.0020029.
- [59] Ng KLS and Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics* 2007; 23: 1321–1330. DOI:10.1093/bioinformatics/btm026.
- [60] Huang S et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* 2018; 15: 41–51. DOI:10.21873/cgp.20063.
- [61] Kohannim O et al. Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiol Aging* 2011; 31: 1429–1442. DOI:10.1016/j.neurobiolaging.2010.04.022.
- [62] Reynolds AP et al. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algorithms* 2006; 5: 475–504. DOI:10.1007/s10852-005-9022-1.
- [63] Hartigan JA. *Clustering Algorithms*. New York: John Wiley & Sons, 1975. ISBN 0-471-35645-X.
- [64] Kaufman L and Rousseeuw P. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 2009.
- [65] Monti S et al. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 2003; 52: 91–118. DOI:10.1023/A:1023949509487.
- [66] Jiang D et al. Cluster analysis for gene expression data: A survey. *IEEE Trans Knowl Data Eng* 2004; 16: 1370–1386. DOI:10.1109/TKDE.2004.68.
- [67] D’haeseleer P. How does gene expression clustering work? *Nat Biotechnol* 2005; 23: 1499–1501. DOI: 10.1038/nbt1205-1499.
- [68] Vavoulis D et al. DGEclust: differential expression analysis of clustered count data. *Genome Biol* 2015; 16: 39. DOI:10.1186/s13059-015-0604-6.
- [69] Oyelade J et al. Optimized adaptive enrichment designs. *Bioinform Biol Insights* 2016; 10: 237–253. DOI: 10.4137/BBI.S38316.
- [70] Drury R et al. The effect of H1N1 vaccination on serum miRNA expression in children: A tale of caution for microRNA microarray studies. *PLoS One* 2019; 14: e0221143. DOI:10.1371/journal.pone.0221143.
- [71] Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; 3: Article17. DOI:10.2202/1544-6115.1128.
- [72] van Dam S et al. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* 2018; 19(4): 575–592. DOI:10.1093/bib/bbw139.
- [73] Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* 2008; 9: 559. DOI:10.1186/1471-2105-9-559.
- [74] Lee-Sarwar KA et al. Integrative analysis of the intestinal metabolome of childhood asthma. *AAAAI* 2019; 144: 442–454. DOI:10.1016/j.jaci.2019.02.032.
- [75] Subramanian A et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 2005; 102: 15545–15550. DOI:10.1073/pnas.0506580102.
- [76] Takahashi S et al. Advanced colorectal cancer subtypes (aCRCS) help select oxaliplatin-based or irinotecan-based therapy for colorectal cancer. *Cancer Sci* 2021; 112: 1567–1578. DOI:10.1111/cas.14841.

- [77] Jolliffe IT and Cadima J. Principal component analysis: a review and recent developments. *Phil Trans R Soc A* 2016; 374: 20150202. DOI:10.1098/rsta.2015.0202.
- [78] Abegaz F et al. Principals about principal components in statistical genetics. *Brief Bioinform* 2019; 20(6): 2200–2216. DOI:10.1093/bib/bby081.
- [79] Lenz M et al. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Sci Rep* 2019; 6: 25696. DOI:10.1038/srep25696.
- [80] Freidlin B et al. The cross-validated adaptive signature design. *Clin Cancer Res* 2010; 16: 691–698. DOI: 10.1158/1078-0432.CCR-09-1357.
- [81] Radmacher MD et al. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002; 9: 404–511. DOI:10.1089/106652702760138592.
- [82] Matsui S et al. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clin Cancer Res* 2012; 18: 6065–6073. DOI:10.1158/1078-0432.CCR-12-1206.
- [83] Cherlin S and Wason JMS. Developing and testing high-efficacy patient subgroups within a clinical trial using risk scores. *Stat Med* 2020; 39: 3285–3298. DOI:10.1002/sim.8665.
- [84] Cherlin S and Wason JMS. Developing a predictive signature for two trial endpoints using the cross-validated risk scores method. *Biostatistics* 2021; DOI:10.1093/biostatistics/kxaa055.
- [85] Ulloa-Montoya F et al. Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J Clin Oncol* 2013; 31: 2388–2395. DOI:10.1200/JCO.2012.44.3762.
- [86] Dreno B. MAGE-A3 immunotherapeutic as adjuvant therapy for patients with resected, MAGE-A3-positive, stage III melanoma (DERMA): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol* 2018; 19: 916–929. DOI:10.1016/S1470-2045(18)30254-7.
- [87] Zhang Z et al. Subgroup selection in adaptive signature designs of confirmatory clinical trial. *J R Stat Soc C* 2017; 66: 345–361.
- [88] Cherlin S and Wason JMS. Cross-validated risk scores adaptive enrichment (CADEN) design. <https://arxiv.org/abs/211102299> 2021; .
- [89] Bhattacharyya A and Rai SN. Adaptive signature design- review of the biomarker guided adaptive phase-III controlled design. *Contemp Clin Trials Commun* 2019; 15: 100378. DOI:10.1016/j.conctc.2019.100378.
- [90] Greenacre M and Blasius J. *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, 2006. ISBN 9780429141966. DOI:10.1201/9781420011319.
- [91] Shannon C. A mathematical theory of communication. *Bell Syst Tech J* 1948; 27: 379–423.
- [92] Simpson E. Measurement of diversity. *Nature* 1949; 163: 688. DOI:10.1038/163688a0.
- [93] Tusher VG et al. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 2001; 98: 5116–5121. DOI:10.1073/pnas.091062498.
- [94] Guinney J et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; 21: 1350–1356. DOI:10.1038/nm.3967.
- [95] Manjang K et al. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Sci Rep* 2021; 11: 156. DOI:10.1038/s41598-020-79375-y.
- [96] Qian Y et al. Prognostic cancer gene expression signatures: current status and challenges. *Cells* 2021; 10: 648. DOI:10.3390/cells10030648.
- [97] Lampinen J and Vehtari A. Bayesian approach for neural networks—review and case studies. *Neural Netw* 2001; 14: 257–274. DOI:10.1016/S0893-6080(00)00098-8.