Chapter 14
Biomarkers

Chris Harbron

Abstract Biomarkers are playing an increasingly important role throughout many aspects of the pharmaceutical discovery and development pipeline. They have many differing roles and applications and statistics plays a critical role in their discovery, validation or qualification and how they are applied and utilised. In this chapter we shall discuss what biomarkers are, the types of data that they generate and the impact on the subsequent statistical analysis, paying particular attention to the avoidance of false positives in biomarker discovery and confirming the technical performance of assays measuring biomarkers.

Keywords Biomarkers • Multivariate analysis • Variability • Cross-validation • Concordance

14.1 What Is a Biomarker?

At its simplest the word "biomarker" can be decomposed into a measure or marker of a biological process. Various different definitions have been proposed to expand upon this, for example one of the most frequently quoted is the National Institutes of Health's (Atkinson et al. 2001, p. 91) definition of a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention".

The word biomarker covers a huge range of different measurements, substrates being measured, technologies to perform the measurement and ways in which they can be applied and utilised. One key feature of biomarkers is that they are measured to make a decision or to provide additional information to assist and inform in the making of a decision. This decision may be at many different levels for many different stakeholders, for example:

- a physician determining the treatment with the most appropriate risk-benefit profile for an individual patient based upon a personalised healthcare biomarker

C. Harbron (✉)
Roche Products Ltd, Welwyn Garden City, UK
e-mail: chris.harbron@roche.com

L. Zhang (ed.), *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*, Statistics for Biology and Health, DOI 10.1007/978-3-319-23558-5_14

- a regulatory body deciding if a development compound can be progressed into man based upon a pre-clinical safety biomarker
- or a pharmaceutical company deciding whether to invest in moving a compound into a later stage of clinical testing by looking at the response to treatment of a Proof Of Mechanism marker

14.1.1 Types of Biomarkers: Uses and Applications

Biomarkers can be categorised in many different ways. Note that these are not mutually exclusive: the same biomarker may fulfil many roles in different scenarios or sometimes within the same scenario.

Various different classifications of biomarkers have been proposed and discussed, for example by Frank and Hargreaves (2003) and Jenkins et al. (2011). Here we will not focus on these definitions, but just provide an overview of some of the most common applications:

Longitudinal or Pharmacodynamic Biomarkers measure dynamic changes in response to a treatment intervention, these can be used to show evidence of the activity of a compound.

The following types of biomarkers provide evidence, in increasing levels, of a compound's interaction with the disease and its potential for clinical efficacy: *Proof Of Mechanism Biomarkers* demonstrate modulation of the drug target; *Proof Of Principal Biomarkers* demonstrate modulation of the cellular phenotype and a *Proof Of Concept Biomarker* demonstrates some therapeutic benefit to a patient (for example an imaging biomarker showing an improvement in the patients' condition) whilst not being a regulatory clinical endpoint. Surrogate endpoints are the next increment in these levels of evidence supporting a compounds' efficacy. For an endpoint to have demonstrated surrogacy it has to have reached strict criteria (Prentice 1989) requiring a large body of evidence.

Dynamic biomarkers can also apply to the safety of a compound at any stage of development. At all stages of pre-clinical and clinical development, the number of individual animals or patients treated will be small compared to the size of the population which will ultimately be treated meaning there is a sizeable risk that a rare adverse event, or evidence that an adverse event is treatment related, will not be observed even in large phase 3 programmes. Safety biomarkers can either alert to potential risks or assist in confirming or otherwise the relationship of the action of the drug to any observed or postulated adverse events.

Another main application, underpinning the growth area of interest of what is variously termed personalised or stratified or precision medicine, are biomarkers explaining and being able to predict the variation in clinical outcome between individuals. These fall into two classes: *Prognostic Biomarkers* which predict the prognosis or likely outcome of disease irrespective of treatment and *Predictive*

Biomarkers which predict the outcome of disease following a specific treatment. These correspond to biomarker main effects and treatment-by-biomarker interactions within a statistical model.

Frequently biomarkers are both prognostic and predictive, as if there are multiple forms of a disease which can be distinguished by a marker, with only one of the forms being amenable to treatment by a certain compound, the underlying prognosis of these differing forms of the disease is also liable to be different.

Predictive biomarkers are frequently based upon measurements taken prior to treatment to determine the most appropriate therapy. However, an alternative "trial-and-error" strategy which may be appropriate in non-acute scenarios with low-risk therapies is to use early response, either in a recognised endpoint, a surrogate endpoint or a pharmacodynamic endpoint, as a predictor of long term response.

14.1.2 Types of Biomarker Data: Binary/Categorical/Continuous

Reported biomarkers will generally either be continuous (e.g. the expression of a gene or concentration of an analyte), binary or categorical (e.g. the presence or absence of a specific mutation) or ordered categorical (e.g. 0, 1+, 2+, 3+ immunohistochemical staining). Statistically, these will require to be treated differently in models, for example adopting either a regression or an ANOVA approach when the biomarker is a response variable. However frequently these different classes of data aren't as distinct as they may initially be presented. For example in contrast with inherited host genetics where mutations will generally either be expressed or not, a mutation test in a heterogeneous tumour will return a percentage of cells exhibiting the mutation—a continuous measure. Similarly whilst IHC markers are typically categorised into 3 or 4 levels they are measuring a continuous underlying level of protein expression, which is sometimes captured by an H-score. There is frequently a tendency to categorise or dichotomise biomarker data and indeed before a biomarker can be utilised as a patient selection tool it will have to be dichotomised into biomarker negative and biomarker positive populations, but this will always lead to a loss of information and power. As a general principal the underlying continuous measure should be sought out and the decision of where to place a cut-off to dichotomise this measure should be delayed for as long as possible, both to maintain power in the meantime and to allow for the final choice of threshold to be based upon as much information as possible.

14.1.3 Types of Biomarker Data: Univariate/Multivariate

In some situations there may be a single biomarker, typically with strong scientific rationale, to be measured and analysed, whilst in other situations there may be

multiple biomarkers, either measured independently or generated simultaneously from a multiplex technology, extending up to many thousands of markers with gene expression microarrays or even millions of genetic markers. When the data is multivariate this may be analysed adopting either a univariate or a multivariate approach or often most fruitfully a combination of the two, this is discussed further in the biomarker discovery section below.

It should also be borne in mind that many univariate biomarkers are measured alongside other variables such as positive and negative controls and the presented result is the outcome of a normalisation procedure. This means that even these apparently univariate variables are the result of a combination of multiple variables and the details of how the normalisation is performed will impact the final result. This is discussed in more detail within the pre-processing section below.

14.1.4 Pre-processing of Biomarker Data

One hidden feature of much biomarker data is that the presented values are frequently not directly measured, but are the result of a number of pre-processing steps. As a general rule the more complex the technology, the greater the degree of data pre-processing that will need to be applied before commencing analysis. These are frequently complex procedures and can develop in an evolutionary manner as issues are discovered and addressed. As an example the Affymetrix gene expression microarray platform has generated in excess of 30 different methods. Clearly this creates the potential for confusion as the application of different pre-processing algorithms will generate different results with the potential to generate different conclusions.

The methods for data pre-processing are bespoke to each technology, but typically will include steps such as background correction, positive control normalisation and housekeeping normalisation. These steps will reduce the effects of both unavoidable technical variability, for example small variations in the quantity of analyte being studied, as well as nuisance biological variability for example related to the quantity or the amount of degradation that has occurred within the biological sample. The objective of a pre-processing algorithm is to generate a measure reflecting the true underlying biological variability that we are interested in rather than any artefact induced throughout the sampling, storage and processing stages. Some simple diagnostic plots can help to assess the effectiveness of the pre-processing algorithm, one of the most powerful is plotting the final normalised biomarker expression against a measure of the sample quality, for example a housekeeping gene. This should be uncorrelated as we would expect that the true biological variability is unrelated to any processing artefacts. This may also highlight outlying poor quality samples which should be treated with caution in any downstream analysis.

14.2 Biomarker Discovery

14.2.1 Data and Biological Support

For a biomarker to have maximal credibility it should be supported both by experimental data and by a wider scientific body of understanding. Differing levels of confidence in the prior knowledge, may lead to different strategies in the identification of biomarkers. With a strong initial focussed hypothesis, there may only be a need to measure the known biomarker as a specific hypothesis test. Alternatively in the absence of much or any strong prior knowledge, a hypothesis free approach looking at a large number of potential biomarkers may be more appropriate. With highly dimensional data such as from microarrays the wealth of data could make finding a few relevant biomarkers a needle in a haystack activity, where any true biomarker would have to be highly significant in order to stand out amongst the false positives. Between these two extremes are alternative strategies, for example a candidate gene approach analysing just a selected subset of markers with enhanced biological rationale or a mixed analysis looking at a single or limited set of candidate markers as a primary analysis, but also a wider more comprehensive set of markers as a secondary or exploratory analysis. This could potentially be a beneficial area for a Bayesian approach, using prior knowledge to place prior likelihoods for each gene.

14.2.2 Multivariate Techniques and Algorithms

There are a large number of variously called multivariate, data mining or machine learning techniques or algorithms. These algorithms generate mathematical models which combine the results from a number of variables to generate a prediction of another variable, typically an outcome. They have the properties that they can cope with situations with more variables than observations, although this power must be used with proper caution to avoid overfitting. That is fitting an overly complex model to the data, modelling noise, which then does not apply more generally to other data sets, and may result in false claims of accuracy. These can be applied in two different ways: to generate predictive models which based upon the measured values predict an outcome or property of a new sample, or variable selection—by observing which variables contribute most to the model and taking these forward for further investigation without being concerned about how the algorithm was combining the individual variables.

A huge number of different data mining algorithms have been developed, each with their own sets of properties. A non-exhaustive selection of these are listed below:

14.2.2.1 Regression Based Methods

Standard regression based methods become unstable and subsequently fall over when the number of variables approaches or exceeds the number of observations and can also become unstable when there are high levels of correlation between the predictor variables.

The *Elastic Net* (Zou and Hastie 2005), including the *LASSO* and *ridge regression* as special cases, minimises the sum of squares of the residuals as in regression with an additional term penalising the size of the coefficients in the fitted model. The balance between the fit to the training set and the size of the coefficients is determined by a tuning parameter, λ, which ranges from an overfitted model perfectly fitting all data points to a null model with all coefficients set to zero when the term for the size of the model coefficients dominates the goodness of fit component. The level of this tuning parameter is set by cross-validation to achieve an appropriate balance between goodness of fit to the training set whilst avoiding overfitting. An elastic net model typically sets the regression coefficients of most variables equal to zero, generating a parsimonious model and acting as an efficient variable selector.

Partial Least Squares (PLS) identifies the linear combination of X-variables which has the greatest covariance with the response variable(s), lying part way between Principal Components Analysis (PCA) which identifies the linear combination explaining the greatest proportion of total variation of the variables independent of outcome and regression selecting the linear combination with the greatest correlation with outcome. In contrast to elastic nets, most variables within a PLS model will have a non-zero coefficient.

Tree Based Methods form the basis of several other data mining algorithms. A standard decision tree, has the advantage of generating a simple readily visibly communicable model, but is a purely heuristic approach with no guarantee of optimality and the final model is frequently one of many different but similarly performing models where there is no reason to believe any one will be more correct or give improved predictions on subsequent data than any other.

PartDSA (Molinaro et al. 2010) attempts to get closer to an optimal tree by using a forward, backward and swapping selection algorithm to generate trees rather than the normal iterative growth.

One of the issues with decision trees is that there are typically a large number of potential trees which are roughly as good as each other within the training data, and none of which are definitively correct. The one that gets selected may be optimal for this particular dataset, but would not necessarily translate to giving

optimal predictions in future datasets. To address this methods have developed which generate a large collection or ensemble of trees which are averaged over to give the final prediction, giving a much more robust prediction.

Random Forests (Breiman 2001) introduce two sources of variability into generating each individual tree. The datasets are perturbed by generating a bootstrap sample of the same size of the original data for each individual tree, and only a random sample of variables are considered for each individual split. These also provide computational advantages, reducing the computational requirements by only examining a subset of variables at each step and the bootstrapping of the data allows out-of-bag estimation of error rates by considering those observations which did not contribute to each individual tree without the much more computationally intensive requirement of cross-validation.

Gradient Boosting Machines (Friedman 2002) also generate an ensemble of trees, but by a different iterative mechanism. Each tree is grown optimally, but then only a shrunken version of the tree predictions is used, by dividing the prediction by a shrinkage parameter. Subsequent trees are fitted to the residuals from the sum of all previous shrunken trees. Eventually this process would converge to a model perfectly fitting the training dataset, so cross-validation is applied to determine the optimal number of trees used within the final model.

These methods can be sensitive to some of the tuning parameters within the algorithm, in particular the parameter which restricts the minimum size of the leaves at the end of each individual tree and in random forests the *mtry* parameter determining the number of variables to consider for each split. The GBM shrinkage parameter has to be balanced between being as small as possible to generate the best models against the practical requirements for computational time, especially if employing a cross-validation strategy.

Support Vector Machines (Cortes and Vapnik 1995) takes a different approach by identifying the widest multidimensional hyperplane that gives the maximum separation between members of different classes.

Nearest Neighbours looks at the closest observations in the training set to a new observation and predicts a class for the new observation based upon a voting scheme. This is most effective in lower dimensional scenarios and in very low dimensional scenarios can behave similarly to random forests.

Genetic Function Approximators mimic an evolutionary process with mathematical equations. These have a tendency to generate complex functions which may not be robust when applied to subsequent data sets.

Whenever using a multivariate analysis, performing a univariate gene-by-gene analysis, i.e. analysing each variable in turn as if it was the only biomarker collected, alongside is frequently highly informative. Although counterexamples can artificially be constructed, in practice it is unlikely that there will be strong signals within a multivariate analysis without some of the signal being visible within

the univariate analyses. Similarly strong signals in a set of univariate analyses will naturally translate to a multivariate algorithm. The *False Discovery Rate* (Storey 2002) provides a valuable quantification of the results of a set of univariate analyses by estimating the proportion of the variables giving an unadjusted significant result which are false positives and with a larger number of variables provides a more pragmatic summary of results than adopting a stricter family-wise error rate control.

Initial data visualisation plays a key role prior to any multivariate modelling. Principal Components Analysis (PCA) is a valuable tool, identifying the key overall sources of variability within the data as well as multivariate outliers. Observations can be coloured both by nuisance parameters such as data source, processing batch or sample quality to identify if there are likely to be issues with the data, or by the parameters being modelled to give an early indication of the likely degree of success of any modelling activities.

Clustering is also sometimes used as an exploratory tool. However, these can be sensitive to the choice of clustering algorithm and distance metrics employed. Underlying this is the fact that the majority of data sets don't fall into neat clusters, but form a continuous distribution and so any attempt to forcibly separate them into distinct clusters is artificial.

The FDA set up the MAQC-II project (Shi et al. 2010) as a collaborative effort to try and establish best practice in the generation of predictive models, particularly to data generated from microarrays although it is reasonable to assume the conclusions are more widely applicable. In MAQC-II 13 microarray data sets had predictive models fitted by 36 separate analysis teams where they concluded that some data sets contained more information that could be modelled than others: "Some endpoints are highly predictive ... provided that sound modeling procedures are used. Other endpoints are inherently difficult to predict regardless of the model development protocol" (p. 834) and that the skill and approach of the analysts had more impact than the choice of one modelling technique above another and there is no universal best modelling technique which is uniformly better than all others : "There are clear differences in proficiency between data analysis teams correlated with the level of experience of the team.", "Many models with similar performance can be developed from a given data set.", "Simple data analysis methods often perform as well as more complicated approaches", "Applying good modelling practices appeared to be more important than the actual choice of a particular algorithm" (p. 834).

14.2.3 Cross-Validation

The concept of validation or replication is to demonstrate a wider applicability of the finding from analysis of any individual dataset by demonstrating that a finding is also seen within another independent data set. The more complex the finding

is, particularly if it is involving multivariate data invoking a complex algorithm or based upon a complex technology requiring many processing steps, the more critical this is.

Whilst validation in an independent dataset will and should remain the gold standard approach, other strategies can be employed to provide an early view of the performance of a biomarker within an independent data set which may be valuable in the frequent situation where there isn't an abundance of appropriate datasets. The simplest approach is to split the data set into training and test sets, where the model is fitted to the training set and tested on the test set. This will understate the difference to a truly independent data set as two parts of the same dataset will be more similar than two separate datasets, but provides a useful initial indication. However splitting the dataset in this manner will lower the power of detecting an optimal biomarker in the training set as it is a smaller subset of the complete data, and also requires retaining a large enough test set to have the power to validate the findings from the training set. There are many possible ways which a data set could be split into training and test sets, and each of these will potentially generate different results, both in the discovery and the validation phases. To maintain credibility for any findings it is therefore critical to specify up front before starting any analysis exactly how the data will be split.

Cross-validation provides a method for efficiently performing independent testing of a model within the same dataset used to generate the model, which can be viewed as repeatedly splitting the data into training and test sets. A proportion $(1/k)$ of the data is selected as a test set and the remainder of the data analysed as a training set to generate a model which is tested on the left out test set. This process is repeated k times until all observations have been excluded from the training set and predicted as members of the test set, and the results aggregated across all k test sets. k can vary from 2 up to the number of observations n (leave-one-out cross validation). Typically values in the range of 5–10-fold cross validation is considered optimal. However, as with the test-training set approach, different divisions of the data into k-folds will give different results. This suggests an approach where the cross-validation is repeated many times each time with different splits in order to reduce this variability.

Cross-validation is used in two different ways. One as discussed above is to estimate the error associated with a model derived from a particular dataset. The other is to optimise a modelling parameter, for example the number of trees in a Gradient Boosting Model or the lambda parameter balancing model fit against coefficient size in an elastic net, where the model is chosen to minimise the cross-validation error. The error from this optimisation will be optimistically biased as the particular parameter or model chosen is that which is optimal for the training dataset. To also obtain an unbiased error estimate, a two-stage cross-validation process must be deployed where the cross-validation to select the optimal parameter is considered part of the modelling process and an additional level of cross-validation is wrapped around the whole process including the first level of cross-validation.

A subtle but important distinction between cross-validation and the training-test set approaches is that the later generates a specific outcome model, and it is

this specific model which is examined in the test set. In contrast cross-validation examines the process for generating the outcome and how well this process typically performs and then extrapolates this to the performance of applying this process to generate a model from the complete dataset. The specific model which is generated by the full-dataset analysis isn't directly tested.

The training-test set and cross-validation approaches are unlikely to capture the variability that will be seen moving to a different data set or from a study to a real-world situation with many more sources of variability. Although a cross-validated estimate of error will be more representative than the model fit error which will be optimistically biased, this error is likely to increase when tested in a completely independent dataset.

14.2.4 Publicly Available Datasets: Invaluable Assets with Hidden Dangers

There is a very welcome movement of making experimental data publicly available, either as supplementary data linked to publications or in public repositories such as the Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo). Indeed some journals are now making this a pre-requisite for publication. This wide sharing and availability of data can only be good for advancing science and this will facilitate both biomarker discovery and testing of biomarker discoveries across a range of independent data sources. However accessing external data sources without fully understanding how the data were collected and any hidden features within the data should be treated with caution. Baggerly and Coombes (2008) provide an example where they demonstrated that an apparent multivariate prediction reported in the literature which had subsequently been deployed into clinical trials could be explained by batch processing effects within the data set.

There is also a similar danger from combining data from several different sources into a single large dataset. Frequently the largest differences in both biomarkers and clinical phenotypes will be between-dataset variation. Failure to be aware of and to subsequently account for this confounding may lead to false results which may exist between datasets but are not visible within datasets and do not extrapolate more widely.

14.3 Technical Performance of Biomarkers

Whatever their intended application, understanding the operating characteristics of a biomarker, its intrinsic level of variability and any biases that it might be subject to is critical. It can be shown that the benefits of a personalised medicine approach enriching by a biomarker rapidly diminish with increasing levels of variability in

the biomarker. Similarly for biomarkers being used as analysis endpoints, increasing levels of variability reduces the power of detecting responses or differences between treatment groups.

The ultimate aim is that a biomarker will give a result reflecting the true current biology of the individual, independent of the particular sample taken and the processing route taken. This is a necessary but not sufficient prerequisite for clinical utility, i.e. that the biomarker measures something useful that has been demonstrated to aid clinical decision making.

14.3.1 Sources of Variability

Biomarkers are subject to many sources of variability, broadly these can be split into two categories: technical variability and biological variability. Once identified there is, theoretically at least, generally the potential to reduce technical variability by tightening up processes, whilst biological variability is frequently something we have to live with and where the most frequent mitigating action is to address with a greater number or density of samples.

Whilst individual sources of variability may be examined separately, the total variability that will be observed in practice is the sum of all the contributing variabilities. This means that if there are a number of moderately sized sources of variation, whilst none of these on their own may be a show stopper, in combination their total variability may limit the information available from the biomarker assay.

14.3.2 Technical Variability

Factors potentially contributing to technical variability cover all steps that occur in the entire process between the patient and the final biomarker result. These include sampling processes, sample thickness, storage and transport conditions, and inter-machine, reagent batch, reader and operator variability.

In the not unusual situation of limited numbers or quantities of samples, approaches such as fractional factorial experimental design may provide more efficient ways to investigate the variation associated with a range of potential sources of variability.

14.3.3 Biological Variability

Biological variability can appear in several different ways each potentially requiring different mitigation strategies.

14.3.3.1 Short Term Temporal Variability

Many biological processes are known to vary on a daily cycle, reflecting waking and sleeping patterns as well as food intake. For this reason, wherever possible, samples should be taken from subjects at equivalent times of day throughout the study.

14.3.3.2 Long Term Temporal Variability

Longer term biological fluctuations can also occur, for example responses to temperature changes throughout the course of a year. In oncology tumours will evolve over time either naturally or developing resistance to treatment so that archival tissue samples, especially those seen in patients who have undergone multiple lines of therapy since the sample was taken, may not be representative of the current disease within a patient. The concordance of archival and fresh samples should be checked and if there is a high discordance consideration should be made of the validity of using archival samples for treatment decisions.

14.3.3.3 Spatial Variability

In oncology different samples taken from the same individual may display heterogeneity. This variability may occur within a tumour, some tissue types e.g. gastric frequently display a large degree of heterogeneity, or may occur between the primary tumour and metastases.

14.3.4 Impact of Variability on Design

With all of these types of variability, statistics has a key role in identifying and describing the levels and types of variability present, and the impact of this variability on the design of studies utilising the biomarker and how the biomarker will be deployed in practice.

Where the agreement between historical and current samples is low this may require obtaining fresh samples from a patient instead of relying on archival samples. Similarly heterogeneity between types of tumour tissue may suggest the tissue type that should be sampled.

Where there is considerable spatial tumour heterogeneity this may require using several samples, for example when using small biopsy samples, the HER2 test for gastric cancer recommends examining 7–8 evaluable specimens from different regions of the tumour, calling a patient HER2 positive and eligible to receive Herceptin™ if any one of these samples shows a positive result.

14.3.5 Acceptance Criteria

Before embarking on a study of variability it is useful to establish acceptance criteria, that is levels of performance which the biomarker assay or diagnostic should achieve. Ideally these should be linked to the clinical impact of any incorrect outcome in either direction, however in advance of knowing the clinical data this can be hard to establish in advance as the relationship of the biomarker to clinical response is likely not to have been established at this stage. An alternative is to observe what levels of variability have been observed previously with this or similar technologies in comparable scenarios.

14.3.6 Comparison to Gold Standard

Where an assay for a biomarker is well-established and can be considered a gold standard, then any new assay for this biomarker should be compared to the current assay and be expected to demonstrate high levels of concordance.

In the absence of a gold standard, for example a new biomarker, then this comparison is not possible, although comparing with other markers to understand the relationships between different markers can be illuminating and help to build confidence if the findings correspond with biological understanding.

The development of an improved biomarker assay, for example with greater sensitivity than previous assays provides an interesting challenge. In this situation some differences to the previous assay are to be expected and hoped for as they represent the improvement from using the new assay. In this scenario it is just changes in the wrong direction, that is a loss of sensitivity which should be identified for concern.

14.3.7 Evaluating Biomarker Performance

Within personalised medicine, biomarkers are frequently dichotomised to a binary measure corresponding to treat or don't treat with a particular drug. This may be a natural binary split e.g. a single nucleotide polymorphism mutation is present or not, may be a split based on a clear biomodality of a continuous measure for example (oestrogen receptor) ER positive or negative in breast cancer or may be an optimal split of a continuous variable derived from an observed relationship to clinical response for example the categories in the Genomic Health Mammaprint tool. Understanding the variability is critical in these situations, it is clearly desirable that if a patient would be prescribed a certain treatment in one centre on Monday, will they be prescribed the same treatment in a different centre on Tuesday.

Table 14.1 Summary statistics for a binary classification test

		Gold standard		
		Positive	Negative	
New biomarker	Positive	TP	FP	PPV = TP/(TP + FP)
	Negative	FN	TN	NPV = TN/(TN + FN)
		Sensitivity = TP/(TP + FN)	Specificity = TN/(TN + FP)	

TP true positives, *FP* false positives, *FN* false negatives, *TN* true negatives, *PPV* positive predictive value, *NPV* negative predictive value

In these cases the relationship to a gold standard is frequently summarised by sensitivity and specificity, respectively the proportion of true positives and true negatives which are correctly identified by the new test. Positive and negative predictive values are complementary concepts, respectively the proportion of positive and negative results from the new test which are correct predictions. Confusingly clinical validity is also often summarised by the same terms where the gold standard test is replaced by clinical outcome, and a positive or negative test correspond to responding and non-responding patients respectively. Table 14.1 below shows these concepts.

The Receiver Operating Characteristic (ROC) curve provides a summary method for examining the performance of a biomarker in terms of its sensitivity and specificity over a range of potential cutoffs that could be used to dichotomise the population into biomarker positives and negatives. These can then be summarised by an area under the ROC curve (AUC), which ends up being a scaled version of the Mann–Whitney statistic comparing the biomarker between the two groups. Recently the theoretical basis of the AUC has been criticised with Hand (2009) proposing the H-measure as an alternative with improved properties.

14.3.8 Evaluating Concordance

A typical concordance study will assess a number of samples a number of times, depending upon the source of variability being studied. A simple overall agreement can be calculated by counting the proportion of times the ratings from the same sample agrees. However, this is highly dependent upon the overall prevalence, for example if predicting a rare disease the majority of samples will be negative so a new diagnostics could attain a high overall agreement by scoring every single sample as negative, which is not very informative. The Kappa statistic provides a chance corrected measure of agreement. The original Cohen's kappa was calculated for two raters and allowed for the two raters to have different marginal distributions. Fleiss's Kappa assumes a common marginal distribution for all raters, allowing extension to any number of assessors. Krippendorf's Alpha provides a method for calculating Fleiss's Kappa for incomplete data with missing values.

One issue can be that as concordance studies will always be of limited size, only a limited range of the total variability that may be observed in practical use will be observed, both in terms of the range of samples being observed and the range of conditions, e.g. pathologists being observed. In fact because of the attention given to aspects such as training in these studies, there is a risk that these studies will understate the true level of variability that will be observed when the test is more widely exposed to the greater range of samples and conditions that will be encountered in real scenarios.

14.4 Summary

Biomarkers are playing an increasingly important role throughout many aspects of the pharmaceutical discovery and development pipeline, being applied in many different ways. Statistics have multiple inputs which are critical to the discovery, validation or qualification and application of biomarkers and how they are applied and utilised.

Fundamentally biomarkers generate data which can be analysed using the same statistical best practice and professional judgement as any other data or endpoint. But as with any analysis understanding the context, how the results will be interpreted and applied, how the data was generated and any features that will influence the data, is critical.

References

Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, Woodcock J, Zeger SL (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95

Baggerly K, Coombes K (2008) Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. J Clin Oncol 26:1186–1187

Breiman L (2001) Random forests. Mach Learn 45:5–32

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Frank R, Hargreaves R (2003) Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov 2:566–580

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378

Hand DJ (2009) Measuring classier performance: a coherent alternative to the area under the ROC curve. Mach Learn 123:77–103

Jenkins M, Flynn A, Smart T, Harbron C, Sabin T, Ratnayake J, Delmar P, Herath A, Jarvis P, Matcham J (2011) A statistician's perspective on biomarkers in drug development. Pharm Stat 6:494–507

Molinaro AM, Lostritto K, van der Laan M (2010) partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. Bioinformatics 26:1357–1363

Prentice R (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431–440

Shi L et al (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827–838

Storey J (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64(3):479–498

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301–320