Part II

Methodology

Chapter 3

EVALUATING DIAGNOSTIC CLASSIFIERS

> "*The Classification problem is: Given a finite set of classified examples from a population, described by their values for some set of attributes, infer a mechanism for predicting the class of any member of the population given only its values for the attributes.*" [Martin & Hirschberg. 1996].

A classifier is any method (such as LD or MLP), which can be used to attempt to solve "The Classification Problem". This chapter is an exposition of methodological issues relevant to evaluating a classifier and to comparing two or more classifiers with one another and drawing conclusions about their relative benefits in clinical settings.

3.1 Measuring the Accuracy of a Diagnostic Classifier

3.1.1 The "Gold Standard"

The accuracy of a classifier can only be assessed against another classifier. This other classifier is usually referred to as the "Gold Standard". This infers that this other classifier is the best possible (or at least a relatively good) method for making that particular classification for cases. This begs the question of why not use the Gold Standard all the time. There may be several reasons:

Firstly the gold standard is sometimes an outcome that only becomes fully known at some later time. We may be interested in predicting this outcome an earlier stage. In this case we can use our knowledge of the eventual outcome to retrospectively validate earlier predictions and conclude about the usefulness of those predictions at that time. An example of this is the later (in this thesis) use of an evaluation of response to treatment with stimulant medication of children with Attention Deficit Hyperactivity Disorder (ADHD) greater than 6 weeks after treatment has commenced to validate predictions about this response made before treatment commences.

Secondly, the gold standard may not be widely or practically available. It might only be possible in a small number of locations; it might require equipment that is not widely in use, it might require expertise, which is relatively rare. An example of this is the later (in this thesis) use of a diagnosis of Autistic Disorder made at specialist diagnostic clinics to validate a diagnosis made with a parent-completed checklist in community settings.

Thirdly, the gold standard may be the currently accepted method for making the classification or clinical decision. When a new method for making the same decision arrives, it is natural to make a comparison. An example of this is the later (in this thesis) use of current diagnostic practices for the diagnosis of Melancholia as the gold standard against which to compare the newer procedure.

Finally as Faraone & Tsuang [1994] point out, in Psychiatry there is often a lack of an absolute gold standard. In many other branches of medicine, there are definitive outcomes such as death, or clear physical evidence of the presence or absence of a disease, which can be used as outcomes. However in Psychiatry, outcomes are usually evaluated by the

judgments of professionals rather than a physical process such as laboratory test. In such a situation, Faraone & Tsuang [1994] recommend the use of multiple convergent criteria as gold standards. An example of this is the later (in this thesis) use of three different diagnostic criteria for Melancholia (Clinical, DSM and Newcastle) to evaluate diagnoses of Melancholia made with a neural network.

3.1.2 Measures of Accuracy

Once a gold standard is chosen, the next question is how can information about the gold standard status of cases be used to evaluate a classifier and to compare different classifiers. There are several indices, which can be used to evaluate classifiers relative to a gold standard. Refer to Table 3.1 below for definitions of the various indices

	Classifier		
Gold Standard	**Positive**	**Negative**	**Total**
Positive	True Positives (TP)	False Negatives (FN)	Positive Population (PP)
Negative	False Positives (FP)	True Negatives (TN)	Negative Population (NP)
Total	Classed Positive (CP)	Classed Negative (CN)	Total Population (P)

Table 3.1 Classifier Outcomes compared to a Gold Standard. (adapted from Ley [1972] and Rey et al [1992])

Proportion Correctly Classified

The simplest measure of a classifier's performance is the *proportion of cases that it correctly classifies*, relative to the gold standard.

$$\textit{Proportion Correctly Classified} = (TP + TN)/P$$

This is the probability of a correct classification relative to the "Gold Standard". It is sometimes also referred to simply as "Accuracy". The inverse of Proportion Correctly Classified (1- Proportion Correctly Classified), the Misclassification Error Rate (also known as the "Error Rate") is also sometimes used, particularly in engineering applications, as a measure of classifier accuracy.

Sensitivity

The sensitivity of a classifier is a measure of how well it identifies cases with a positive diagnosis. Formally it is the *proportion of cases with a positive diagnosis classified as having the diagnosis*. It can range from 0 (worst) to 1 (best). The term True Positive Rate is synonymous with Sensitivity.

$$\textit{Sensitivity} = TP/PP$$

Which is the probability that a positive case will be classified as a positive case.

Specificity

It is desirable for a classifier to be sensitive, but it should also be specific. The Specificity of a classifier is *the proportion of cases with a negative diagnosis classified as not having the diagnosis.* It can range from 0 (worst) to 1 (best). The term False Positive Rate (also known as the "False Alarm Rate") is equal to 1 minus Specificity.

$$Specificity = TN/NP$$

This is the probability that a negative case will be classified as negative case.

Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

Whilst Sensitivity and Specificity are good overall indicators of a classifiers performance, information about the base rates of classes needs to be considered in individual clinical seetings. The Positive Predictive Value (PPV) is *the proportion of the cases classified as positive who actually have a positive diagnosis.* This tells an individual patient (or a clinician about an individual patient) classified as having a positive diagnosis what their probability of actually having that diagnosis is. It can range from 0 (worst) to 1 (best).

$$PPV = TP/(TP + FP)$$

Which is the probability that a case classified as positive, actually is positive.

Similarly, the Negative Predictive Value (NPV) is *the proportion of the cases that are classified as negative that actually do not have the diagnosis.* This tells an individual patient classified as not having the diagnosis what is their probability of actually not having that diagnosis. It can range from 0 (worst) to 1 (best).

$$NPV = TN/(TN + FN)$$

Which is the probability that a case classified as negative, actually is negative

A_z - Area Under the Receiver Operating Characteristics (ROC) Curve

A classifier usually classifies by first calculating a score on scale and then using a cut-off point or threshold on that scale to classify into categories. Sensitivity, Specificity, PPV and NPV will all vary as function of the classification threshold (cut-off point) of a classifier. At varying cut-off values, trade-offs occur between sensitivity and specificity and between PPV and NPV.

The ROC (Receiver Operating Characteristics) curve is a plot of how the classifier performs over the entire range of possible choices of cut-off values. Each point on the curve represents the True-Positive Rate (Sensitivity) plotted on the y-axis and the False-Positive Rate (1 – Specificity) plotted on the x-axis that results from a particular cut-off value.

The curve is anchored at two points, 0,0 and 1,1. At 0,0 there is no sensitivity (i.e. no cases are classified as positive and therefore none of the positive cases are identified), but since all cases are classed as not having the diagnosis, Specificity is perfect (1) in that all negative cases are classified correctly. The cut-off value at this end is high. At the other end (1,1) the cut-off value is low and all cases are classified as having the diagnosis. Therefore Sensitivity is perfect (1) and all the positive cases are classified correctly. However Specificity is zero because no cases are classified as not having the diagnosis.

A ROC curve, which is a straight line between the two anchor points, represents a classifier, which does not classify above chance. That is, there is no relationship between the classifier's output score and the gold standard. This is the worst possible classifier. The more the ROC curve bulges away from the "worst possible classifier" line, the better the

classifier is at classifying. This is because improvements in Sensitivity or Specificity, that come about as result of changes to the cut-off, do not come at great reciprocal cost. That is a given gain in Sensitivity will not cost an equal loss in Specificity. The best way to measure the size of this bulge away from the midline is to calculate the Area under the ROC curve (AUC), which is also known as the statistic A_z. This index gives an overall measure of the classifier as whole, across all possible cut-off point.

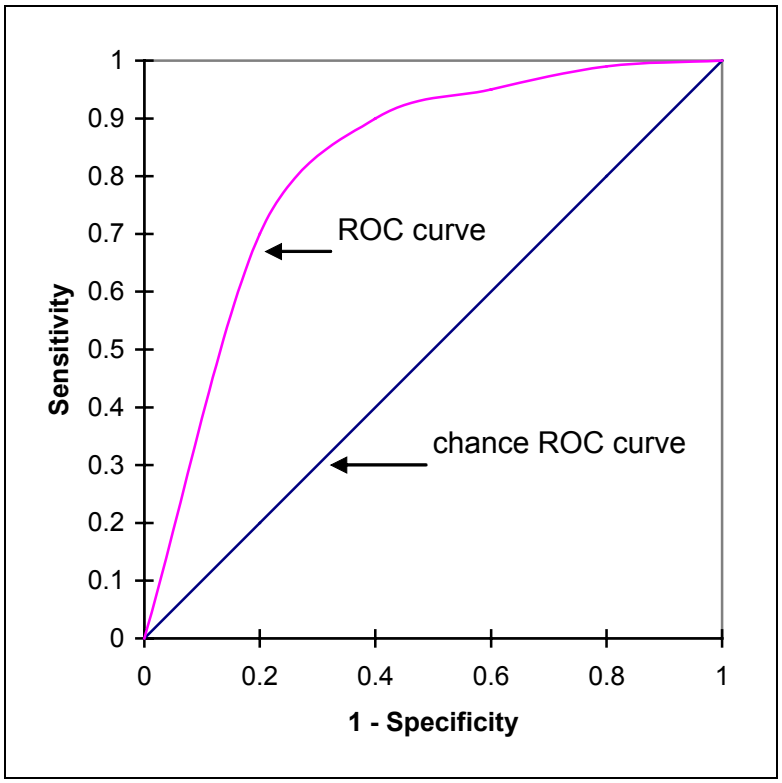

Figure 3.1 Typical Receiver Operating Characteristics (ROC) curve. The 45-degree line is the chance ROC Curve (no agreement with the "Gold Standard" which is above chance). The Area under the ROC curve (A_z) is an index of how well the classifier, on which the curve is based, classifies overall.

A_z	Classifier Accuracy
.90 - 1.00	Excellent
.80 - .89	Good
.70 - .79	Fair
.60 - .69	Poor
.50 - .59	Fail

Table 3.2 **C**lassification of Areas Under the ROC Curve [Tape, 2002]

Table 3.2 above presents a classification scheme for values of A_z (the Area under the ROC Curve), which can be used for evaluating the results of analyses using ROC Curves.

Which Measure(s) should be used?

The answer is different measures are better in different situations. Different measures give different information about a diagnostic system(s) under evaluation. It is important to know which measures contribute to which evaluations.

For the overall comparison of classifiers, the Area Under the ROC Curve is the best measure. It measures compares classifiers across their whole range of possible cutoff values. In abstract terms this measure can be used to judge and compare classifiers as a whole. However there is one proviso. Two ROC curves can have equal areas but still be different. They can bulge (or flatten) at different parts of the curve. Thus their areas can be the same but their shapes different. In clinical applications this can have important consequences. We always use a classifier with a specific cut-off point, which translates into

specific values for Sensitivity and Specificity. In such a case, where two ROC curves have the same or similar values for A_z, careful comparisons of Sensitivity and Specificity in ranges of interest for these, should guide the choice of classifier amongst candidates [McNeil & Hanley 1984].

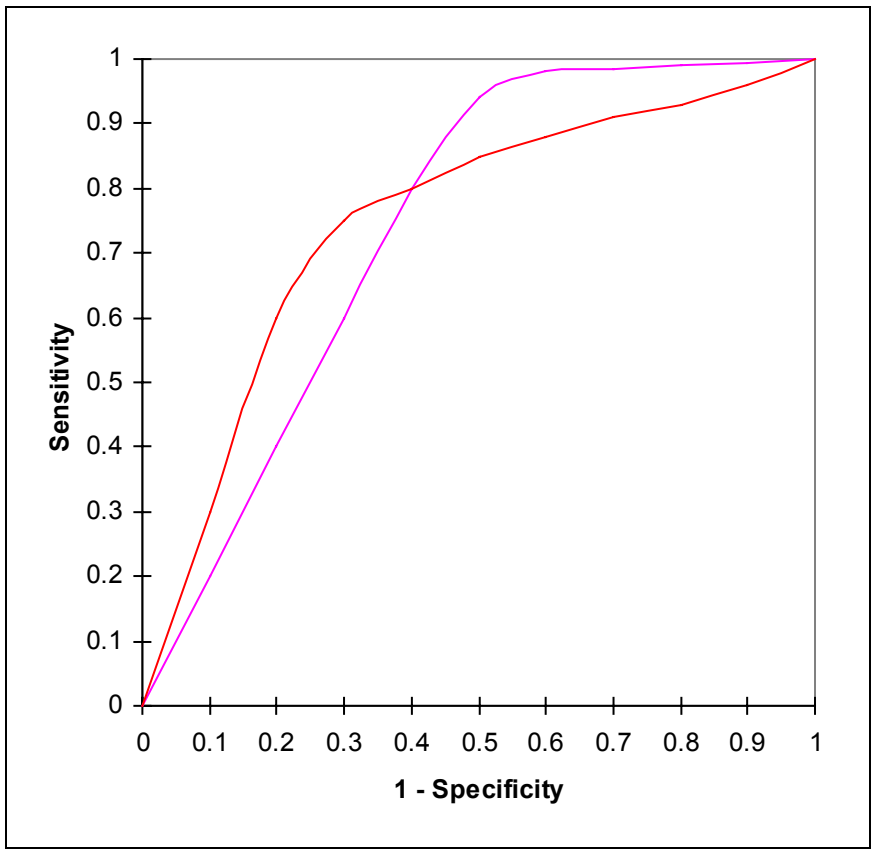

Figure 3.2 Two ROC Curves with the same or similar values for A_z (Area under the ROC Curve), but different shapes and therefore different accuracies in specific Sensitivity or Specificity ranges of interest.

At the time of application of a classifier as a diagnostic practice, we are more interested in the performance of the classifier at a specific cut-off value, than we are in performance over the entire range of possible cut-offs. Once a cut-off value is chosen Sensitivity and

Specificity are the appropriate measures by which to judge and compare classifiers and diagnostic practices. These measures inform clinicians about the performance of the classifier as implemented, not in abstract terms. Conversely the choice of cut-off can be made to so as to achieve a given level of sensitivity and/or specificity [McNeil & Hanley 1984]. Therefore before making a choice amongst candidate classifiers, where A_z, values are similar, visual inspection of the ROC curves is recommended.

Finally, from the point of view of individual patients, or of clinicians in a particular clinical setting, the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) of a diagnostic practice need to be considered. These measures take account of the "Base Rate" [Ley, 1972], which is proportion of patients, amongst all patients to whom the practice is applied, that actually have the diagnosis (PP/P in Table 3.1).

3.2 Generalisation Assessment

Regardless of the measure used, the apparent accuracy of a classifier on a training dataset is an optimistic overestimate of the classifier's true accuracy [McLachlan, 1990]. It is optimistically biased as a result of capitalisation on chance relationships between the input variables and the output variable(s) that occur in samples drawn from a population. The classification accuracy on new cases will be less than that obtained with the training set [Hand 1985, 1997, McLachlan, 1990, Ripley 1994, 1996, Bishop 1995, Reed & Marks 1999, Hastie et al 2001]. Therefore it is not a valid practice to compare classifiers using accuracy measures derived from a training set.

Optimistic bias occurs because classifiers can *overfit,* as previously discussed in Chapter 2 under the topic of the bias–variance trade off. Classifiers fit (approximate) the target function between the inputs and outputs (Bias component), but in addition they also fit chance relationships, which arise due to sampling noise (Variance component). The location in the input space of each case in the training set is determined by two factors:

1. The underlying target function (which is a deterministic relationship between inputs and outputs), referred to as the bias component and;
2. Variations in input values, which are due to all other factors (which are not being measured or used as inputs), referred to as the variance component, or as *noise*.

Across a population variance has a zero mean effect upon the values of input variables. That is it averages out across all cases. When we draw a sample from a population, it is a subset (in some instances a small subset) of the population and the impact of variance across all the cases is unlikely to average out in that sample. Thus each sample has unique biases, due to variance or noise, which a classifier cannot discriminate from the target function, and therefore the classifier mistakes these biases as being a part of the target function.

This problem, which was discussed more extensively in chapter 2 as the bias-variance trade off, is widely recognised and number of strategies, have been developed to overcome it. These are discussed below.

3.2.1 Strategies for the Measurement of Generalisation

There have been a large number of methods developed which attempt to measure generalisation accuracy. That is how well the classifier will be able to classify future cases. In this section we discuss those which seems to be the most commonly used in the medical clinical decision making literature

Validation using a large independent test dataset

The best way to assess generalization (for any classifier) is to use a large independent test data set, which has been randomly drawn from the population to which it is intended that it will be applied. This method is sometimes incorrectly referred to as cross-validation, but strictly speaking it is not (see next section), it is validation.

It is important that the test dataset (as well as the training data set) is randomly sampled from the population, because this reduces sampling biases. Without random sampling there is an increased probability that sampling could result in a poor representation of the target function with respect to at least some region(s) of the input space [1]

For similar reasons it is also important that the size of the test data set is large. A small or medium sized test set may not have a good representation of all the important features of the function being approximated. If this is the case then the prediction of generalization made from this test data set may not be accurate.

[1] a space defined using the input variables as dimensions and delimited to a finite size by the range of values of the input variables.

Both the training data set and the test dataset need to contain a large number of cases. In many clinical decision-making studies, the number of cases available to an investigator(s) for use in training and in generalisation assessment may not be large. How should the available cases be apportioned to training and test datasets? Reed & Marks (1999) reason that since both are important and it is not possible to weight one as more important than the other, then the apportionment should be 50-50.

When the number of cases available for an investigation is small, use of a large test dataset for generalisation assessment may not be a viable strategy.

Cross-validation

Cross-validation, makes use of all the available cases for both training and generalisation assessment. In simple cross-validation the available cases are randomly divided into two datasets. Both datasets are independently used as training datasets and then generalisation is measured, in both cases by applying the resultant model on the 'other' set. The average of the two generalisation measurement results is used as the estimate of generalisation error of this kind of model.

"Leave k out" cross-validation extends this concept further. In this method the classification rule is derived several times, each time leaving out a fixed number (k) of cases and each time cross tested on this small set of k cases. The process is repeated, each time with a different set of k cases left out and used for validation, until all cases have participated in cross-validation. Thus all cases are used for derivation and for cross-validation, but no individual case participates in both simultaneously. The number of correctly classified cases in the cross-validation samples, across all N/k trials (where N is

the total number of subjects in the study), determines the estimated accuracy of the classifier. Monte Carlo studies have found that "leave k out" cross-validation, whilst better than no cross-validation, has an optimistic bias (McLachlan, 1990; Efron & Tibshirani, 1993). That is, it overestimates the true accuracy of a classifier.

Leave one out cross-validation

Leave one our cross-validation is an extreme version of the "leave k out" cross-validation in which, one case is left out of the training data set, the classification rule is derived and then the left out case is classified with it. This procedure is repeated for all cases, that is N times. The obvious problem with this is that it is computationally expensive, though with small data sets this may not matter.

The Jacknife

The Jacknife is also based on leaving one case out at a time and rotating through the entire available dataset. However rather than directly calculating error or accuracy, the Jacknife is used to derive an estimate of the optimistic bias due to overfitting. This estimated bias due to overfitting is then used (by addition or subtraction) to adjust the training dataset derived value of error or accuracy.

Bootstrapping

A more recently developed technique for estimation of classifier accuracy, that makes full use of the dataset to derive the classification rule, is *bootstrapping* [2](Efron & Tibshirani, 1993). This method is similar to "leave k out" cross-validation but has some important

[2] The term *bootstrapping* is engineering jargon for a process that starts and/or fuels itself and does not rely upon external influences. The www.dictionary.com definition is: "Being or relating to a process that is self-initiating or self-sustaining"

differences. First the classification rule is derived on the whole data set to give an optimistic estimate of the classification accuracy (known as the apparent accuracy). Next a number of bootstrap data sets are developed each with the same N as the original data set, using random sampling with replacement (i.e. a case is selected and placed in the bootstrap set and then it is replaced in the pool of N cases so that it can be possibly re-selected). Thus each bootstrap data set has N subjects some of which may be duplicates. Each bootstrap sample is then used to derive a classification rule, which is applied to the original full dataset and classification accuracy is measured. The mean difference, across all bootstrap samples, between classification accuracy on the bootstrap sample itself and on the original data set, is used to estimate the magnitude of the optimistic bias between a derivation sample and the population from which it is randomly drawn. Also, because the optimistic bias estimate is a mean derived from many bootstrap samples, it has a standard deviation, which can be used as an estimate of the accuracy of the optimistic bias estimate. The generalisation accuracy of the classifier (how it should perform on any random sample of the population) can then be calculated by subtracting the estimated optimistic bias from the apparent accuracy of the classifier trained on the whole original data set. Efron & Tibshirani (1993) and McLachlan (1990) both summarise the theoretical work and simulation studies, which indicate that use of bootstrap estimation of classifier accuracy is only slightly more optimistic than the use of a large test data set.

3.3 Subgroup or Spectrum Effects

Another important consideration for the evaluation of classifiers and/or diagnostic procedures is to have knowledge about the accuracy of the classification in reference to subgroups, which are likely to be in the population(s) in which it will be implemented. The presence, absence and mix of subgroups may vary considerably from one place of application to another, and this will naturally lead to variations in diagnostic accuracy, possibly from very poor to very good, if accuracy is high in some groups and low in others. This phenomena an extension of the "base rate problem" discussed by Ley [1972], was first raised by Ransohoff and Feinstein [1979] and has recently been termed "the Spectrum Effect" by Mulherin and Miller [2002].

Many classifiers and new diagnostic procedures are developed and validated in artificially constructed populations, which are composed by amalgamating an identified clinical group with a control group. "Real" clinical populations, that is those seen in the clinical settings in which the new practice is likely to be applied, are likely to contain a different mix of these subgroups, and may also contain subgroups not present in the developer's population.

Using clinical populations for development and validation is generally a good strategy for countering the spectrum effect. A better strategy is to identify important subgroups amongst the cases and to report classification accuracy by subgroup [Mulherin & Miller, 2002]. This allows intending users who have a different, but known, composition of subgroups, in their local clinical population to calculate the local accuracy and efficacy of the new putative diagnostic practice.

There is a variation of the subgroup effect which has been named ‘population drift’ by Hand [1997]. He points out that over time clinical populations can change, sometimes relatively rapidly for various reasons such as change in causative factors, changing demographics, successful prevention or intervention, or other factors. Underlying this change it is likely that the mix of subgroups, which make up the population has changed. By understanding the behaviour of classifier in reference to important subgroups and having awareness of the changes in subgroup composition which are occurring, then it may be possible to tune a classifier to a changed population without wholly re-developing it. A related concept ‘case-mix’, has been found to be very useful in looking at resource utilisation in health services, and provides a good explanation of differences in hospital service utilisation such as length of stay, and costs.

3.4 Conclusions

A classifiers accuracy can only be measured by comparison with another classifier, “The Gold Standard”, which should be the either best possible or one of the best possible classifiers for a particular classification problem.

There are several widely used measures (or statistics) of a classifier’s accuracy, and some of these have more than one name. These include the Proportion Correctly Classified (also known as Accuracy), the Misclassification Error Rate (Error Rate), Sensitivity (True Positive Rate), Specificity (1 – False Positive Rate), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and the Area Under the ROC Curve ($\mathbf{A_z}$). All these measures are highly interdependent, but each provides some form of unique information about a classifier and each is an appropriate measure for consideration in an appropriate context. The most general for comparing classifiers is $\mathbf{A_z}$ (the Area under the ROC Curve). However when it comes to the application of a classifier in a specific clinical context,

Sensitivity and Specificity are more important as considerations. And when it comes to interpreting the result of a classifier for individual patient, PPV and NPV are the important considerations.

Regardless of the classification accuracy measure used, it is almost certain to be optimistically biased (that is make the classifier look better than it really is), if it is derived from the same training dataset, which was used to develop the classifier. This is because all classifiers "Overfit": that is they capture chance-variation (which is unique only to that training dataset sample) and use it to classify cases in the training dataset. The "over-fitted" unique-to-training dataset sample chance-variation will not generalise to any other samples drawn from the same population as the training dataset sample. As a result the accuracy of classification of new cases from this population will in actuality be less than estimated from the training dataset. (see discussion on 'variance' in the Bias-Variance Trade off section of Chapter 2 for a better elucidation of this).

The obvious solution to this problem is to derive measures of classification accuracy by applying the classifier to a second independent very large dataset of cases, which has been randomly drawn from the population of interest. This is known as a large test dataset. If the supply of case data is large and inexpensive, then this may be an appropriate strategy (bearing in mind an equally strong need for cases for model development).

However in psychiatry (and in medicine in general), the supply of cases is limited and/or the cost of data collection per case is expensive. A good alternate strategy is to use Bootstrapping: a re-sampling with replacement technique used to create a large number of slightly different datasets (bootstrap datasets with N equal to the original dataset), derive the classifier many times by using these datasets, validate each of these classifiers on the original dataset, calculate the optimistic bias for each bootstrap derived classifier, find the average optimistic bias, and subtract this figure from that of a classifier derived using the whole training dataset, to obtain a corrected estimate for the classifiers accuracy.

Finally it is important to be cognisant of “Spectrum Effects”. This is the differential accuracy of a classifier amongst different population subgroups to which it is applied. Though the same number and kinds of subgroups may exist in similar clinical settings, the mix (the relative proportions of individual subgroups) can differ from one setting to another. These differences can lead to unexpected differences in classification accuracies, in different applied settings, or in the same population if the subgroup mix has changed over time (“population drift”). If we know the relative accuracies of a classifier amongst subgroups, and we have knowledge of the subgroup composition of a site being considered for application of the classifier, then we can calculate both: the overall accuracy of the classifier at this new site and; the efficacy of application of the classifier to individuals from different subgroups.