Statistical Methods in Medical Research 1998; **7**: 324–336

Analysis of clustered data in receiver operating characteristic studies

Craig A Beam Medical College of Wisconsin, Milwaukee, Wisconsin, USA

Clustered data is not simply correlated data, but has its own unique aspects. In this paper, various methods for correlated receiver operating characteristic (ROC) curve data that have been extended specifically to clustered data are reviewed. For those methods that have not yet been extended, suggestions for their application to clustered ROC studies are provided. Various methods with respect to their ability to meet either of two objectives of the analysis of clustered ROC data are compared to consider a variety of ROC indices and their accessibility to researchers.

The available statistical methods for clustered data vary in the range of indices that can be considered and in their accessibility to researchers. Parametric models permit all indices to be considered but, owing to computational complexity, are the least accessible of available methods. Nonparametric methods are much more accessible, but only permit estimation and inference about ROC curve area. The jackknife method is the most accessible and permits any index to be considered.

Future development of methods for clustered ROC studies should consider the continuation ratio model, which will permit the application of widely available software for the analysis of mixed generalized linear models. Another area of development should be in the adoption of bootstrapping methods to clustered ROC data.

1 Introduction

In this paper, the methodology for the analysis of clustered data in receiver operating characteristic (ROC) studies is reviewed. Various methods are compared with respect to their ability to meet either of two objectives of the analysis of clustered ROC data, to consider a variety of ROC indices and their accessibility to researchers.

Clustering presents a special case of correlated data. To understand its uniqueness, the concept of a 'Diagnostic Unit Of Study' or DUOS will be introduced. A DUOS is the smallest entity receiving a diagnosis. Examples of DUOS are: a woman being screened for breast cancer and receiving a single recommendation for follow-up; a hemisphere of the brain that is assessed pathologically for the presence of 'tangles'; a region on a liver scan that is located by an observer and then graded for the likelihood of the presence of a mass.

Clustering occurs in a diagnostic study if and only if there exists correlation among DUOS after conditioning on experimental or other artificial sources of correlation (such as induced by multiple modalities, multiple readers, and so on). In other words, clustering refers to correlation among DUOS that is induced by natural, biological processes rather than experimental or artificial processes.

Typically, clustering in diagnostic studies comes about by the nesting of DUOS within biological entities. Two examples from radiology serve to illustrate nesting:

Address for correspondence: CA Beam, Medical College of Wisconsin, 8901 Watertown Plank Road, PO Box 26509, Milwaukee, WI 53226-0509, USA.

 0962-2802(98)SM167RA

1) In a study of the accuracy of ultrasound (US) in screening for carotid stenosis, measurements are taken from the left and the right carotid arteries.[1] These DUOS are nested within subjects and are correlated because of the biological characteristics of the nesting unit. There is no reason, however, to suspect any difference in the performance of US to detect stenosis between the left and right carotids, so the analytical objective of this study is to estimate a single ROC curve for US using correlated data. However, what makes the clustered data in this example a unique type of correlated data is that carotids from the same subject can have different disease states ('stenotic' or 'not stenotic').
2) In a study of the diagnostic accuracy of magnetic resonance imaging (MRI) in lung cancer, different regions of the lung were considered (chest wall and mediastinal) for the presence of cancerous invasion.[2] Each subject in the study has, therefore, two DUOS and distinct ROC curves for each region were constructed. Statistical methods to provide appropriate contrasts between these two correlated ROC curves were then required.

From the prior two examples, it is clear that clustered data are not simply correlated data, but have their own unique aspect. Of course, correlation is an important component in the analysis of clustered data. In Sections 2–7, I review various methods for correlated ROC curve data that have been developed. Some of these methods have been extended specifically to clustered data. For those that have not, suggestions for their application to clustered ROC studies are supplied.

2 Principles of ROC curve analysis

2.1 Basic notation and definitions

Let d_{ij} denote the ith DUOS within nesting unit j, z_{ij} be the value of a diagnostic marker from d_{ij} and $x_{ij} = (x_{ij}1, x_{ij}2, ...)$ be the vector of covariates associated with d_{ij} with $x_{ij}1 = 0$ if d_{ij} is an 'unafflicted' unit and $x_{ij}2 = 1$ if d_{ij} is an 'afflicted' unit.

It is further assumed that

$$(z_{ij} \mid x_{ij}) \sim dF_{x_{ij}} \text{ when } x_{ij1} = 0$$

and

$$(z_{ij} \mid x_{ij}) \sim dG_{x_{ij}} \text{ when } x_{ij1} = 1$$

where the location and/or scale of the distribution might depend on the remaining components of the covariate vector.

By the definition of clustering used in this paper,

$$\text{cov}(z_{ij}, z_{ij'} \neq 0) \text{ for some } j \neq j'$$

and

$$\text{cov}(z_{ij}, z_{i'j} \neq 0) \text{ for all } i \neq i'$$

Suppose it is the case that the DUOS within each nesting unit must have the same disease status. Some researchers have adopted the assumption that

$$\text{cov}(z_{ij}, z_{i'j} \mid x_{ij1} = x) = 0, \text{ for } x = 0 \text{ or } x = 1$$

Although such assumptions have been applied in multimodality studies, they do not seem particularly relevant to clustered data, due to the fact that in the later case DUOS within nesting units can have different disease statuses. Conditioning on disease status would not, therefore, 'condition out' the nesting factor that generates correlation among clustered DUOS.

The ROC curve is usually defined to be the plot of the two upper quantiles

$$1 - G_{x_{ij}}(\cdot) \text{ vs } 1 - F_{x_{ij}}(\cdot)$$

$1 - G_{x_{ij}}(\cdot)$ is often referred to as the 'sensitivity' of the test, modality or reader while $1 - F_{x_{ij}}(\cdot)$ is often referred to as the 'false positive probability'.

We immediately note a complication: the analysis of clustered data will yield several ROC curves for each nesting unit (e.g. 'patient') whenever covariate values differ among DUOS and these differing values determine differing CDFs in the associated diagnostic marker. Studies involving multiple organ systems are not uncommon in diagnostic radiology. Bilateral organs often behave differentially to contrast media and different anatomical regions are often distinctly different so that multiple ROC curves are not easily assumed away. The analyst is therefore often faced with the challenging problem of how to make sense of multiple ROC curves. Of course, the complexity is easily treated by assuming the covariate vector (except for the 'disease covariate') is the same for all DUOS from the same nesting unit. This has been done, for example, with the analysis of US in carotid stenosis screening.[3]

This paper will focus on ROC curve data arising from ordinal, or 'ratings', data. Such data are very common in medicine, particularly due to the ubiquitous involvement of human observers in the process of diagnosis. Parametric approaches to estimating ROC curves from clustered continuous data are fairly straightforward and will not be addressed directly here. However, many of the conceptual considerations in this paper are also applicable to continuous or nominal data. Nonparametric approaches for clustered continuous data follow the same methodological strategies described below.

2.2 ROC curve indices of diagnostic accuracy

Several indices of the diagnostic accuracy of tests, modalities or observers based on the ROC curve have been developed. Perhaps the most commonly used index is that of the area under the ROC curve[4]

$$A_{[0,1]} = \int_0^1 \{1 - G[F^{-1}(p)]\} \mathrm{d}p$$

Better modalities have higher sensitivity at each false positive probability and so the ROC curves of better tests dominate those of lesser tests and arch more closely to the ideal (0,1) point. Therefore, better tests should have greater area under their ROC curve.

However, examination of the definition of $A_{[0,1]}$ reveals that it is the sensitivity of the modality averaged over uniform selections from the entire range of false positive

probabilities. Since large false positive probabilities are often not clinically relevant or acceptable, the area index can assign a favourable value to tests that dominate only over these regions. Conversely, tests whose ROC curves dominate over a restricted range of small false positive probabilities, but then 'cross under' another ROC curve, may have a smaller total ROC curve area.

The partial ROC curve area[5,6] restricts attention to clinically relevant regions of false positive probability. It is defined to be

$$A_{[0,l]} = \int_0^l \{1 - G[F^{-1}(p)]\}\,\mathrm{d}p$$

where l is a chosen upper limit to clinically relevant or acceptable false positive probability.

As an extreme version of partial area, the index

$$TP_{fp} = 1 - G[F^{-1}(p)]$$

has been considered.[4] This index, however, requires specification of a single false positive probability at which to evaluate or compare tests. However, the definition is determined automatically whenever one test is based on a discrete diagnostic marker and therefore possesses only a finite number of false positive probabilities.[7]

2.3 Goals of the analysis of clustered ROC curve data

In ROC studies involving clustered data, the use of the previous indices depends on whether on not distinct DUOS give rise to distinct ROC curves. In the carotid example, we might consider that the two carotids contribute information to the diagnostic accuracy of ultrasound in screening for carotid stenosis. There is a single ultrasound ROC curve and data from the two carotids contribute to its estimation. However, when we consider the problem of the detection of lung masses, we might treat the chest wall and the mediastinal region as distinct diagnostic entities giving rise to distinct ROC curves. In the latter case, our goal might be to compare these two distinct ROC curves arising from the use of one modality.

The goals of the analysis of clustered ROC curve data can, therefore, either be to appropriately estimate ROC indices of a single modality, or to estimate indices of several distinct ROC curves generated from the same modality imaging distinct, but nested, DUOS. Statistical methods for correlated data are required for the first goal. They are required for the second goal only when comparisons of distinct ROC curves (for example, between chest wall and mediastinal regions) is desired. Such considerations are very important as they significantly impact study design and analysis. Specifically, the analysis of uncorrelated ROC curve with contemporary software is much more accessible to typical researchers than is the analysis of correlated ROC curves, except in the very simplest settings. Unfortunately, as will be seen in the subsequent discussion, the simplest settings often do not hold with clustered diagnostic data.

3 Classical approaches and limitations

Much of the statistical work in diagnostic medicine in the past decade has been concerned with the analysis of ordinal data from human observers. Two major approaches have evolved: (1) the nonparametric analysis of ROC curve area and (2) parametric and semi-parametric analysis of ROC curve indices based on latent class models. First modelling approaches from the latent class tradition and their relevance to clustered data are described. Nonparametric approaches are covered in Section 5.

The typical diagnostic study asks an observer to assign a rating to each of a set of DUOS. Typically, the rating indicates the degree of suspicion the observer has about the presence or absence of a disease or condition in the DUOS.

The classical approach to analysing these data is to then assume the ratings are a discretized version of a latent (i.e. 'unobservable') continuous marker as follows:[8]

$$P[Y_{ij} \leq c] = P(\theta_{c-1} < z_{ij} \leq \theta_c)$$

where Y_{ij} represents the rating given to the ith DUOS within nesting unit j, $c = 1, 2, \ldots,$ C, z_{ij} is a continuous but unobservable diagnostic marker and $\theta_1, \theta_2, \ldots, \theta_{C-1}$ are 'cutoffs' that partition the support of z_{ij}.

An important feature of this latent marker model is that it accommodates the empirically based fact that human observers can intentionally alter their diagnostic performance from experiment to experiment, even when interpreting exactly the same diagnostic material.

Dorfman and Alf[9] introduced a maximum likelihood procedure for estimating parameters of the ROC curve from nonclustered ordinal data coming from a single reader and single modality using the latent marker model, and when one can assume that

$$\mathrm{d}F = N(0,1) \text{ and } \mathrm{d}G = N(a,b)$$

Using this assumption, the ROC curve for that single reader-modality combination is completely specified by the parameter set $\{a,b\}$ and the ROC curve area is the functional

$$\Phi\left(\frac{a}{\sqrt{(1+b^2)}}\right)$$

Metz *et al.*[10] extended the maximum likelihood procedure to correlated ROC curve data arising from experiments in which two modalities are applied to the same set of subjects or two readers read the same set of diagnostic material. The basis for this extension is the assumption of a bivariate normal distribution for the two latent values coming from the same subject. The method, however, is restricted to accounting for correlation within subjects when subjects are DUOS. This type of correlation does not meet the definition of clustering used in this paper and so the method is not applicable to the situation of within subject DUOS that can have different disease statuses. This limitation in the classical, maximum likelihood approach can be overcome by the adoption of the more general ordinal regression model described below.

4 Ordinal probit regression models for clustered ROC data

4.1 The ordinal probit regression model

Tosteson[11] introduced an ordinal regression approach for the analysis of ROC curves from ratings data. This general model includes the classical bivariate normal model as a special case when the probit link function is used and only a single covariate, associated with disease status, is included.

The general form of the ordinal probit regression model for diagnostic ratings data assumes that

$$\Phi^{-1}\left[P(Y_{ij} \leq c)\right] = \frac{\theta_c - \alpha' x_{1ij}}{\exp(\beta' x_{2ij})}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\alpha = [\alpha_1, \alpha_2, ...]$ is a set of 'location' regression parameters, $\beta = [\beta_1, \beta_2, ...]$ is a set of 'scale' parameters and $\theta_1, \theta_2, ..., \theta_{C-1}$ are cutoffs. The vectors x_{1ij} and x_{2ij} are referred to as location and scale covariates, respectively.

4.2 Generalized estimating equations

Toledano and Gatsonis[12] extended the probit regression model to include the situation of correlated data arising from a combination of multiple modalities and/or multiple readers and developed a 'generalized estimating equations' (GEE) approach[13] for this situation. This development included, as necessary, the regression modelling of correlation. To accomplish this latter modelling, the authors assume that

$$P(Y_{ij} \leq c, Y_{ij'} \leq c') = \Phi_2(\Phi^{-1}(\mu_{ijc}),\ \Phi^{-1}(\mu_{ij'c'} \mid \rho_{i,jj'}(\zeta x_{3i}))$$

where $\Phi_2(\cdot,\cdot)$ is the bivariate normal cumulative distribution function, $\mu_{\mathrm{ijc}} = P(Y_{ij} \leq c)$ and where it is further assumed that correlations can be modelled by

$$\rho_{i,jj'} = \frac{(1 - \exp(\zeta x_{3i}))}{(1 + \exp(\zeta x_{3i}))}$$

with $x_{3i} = (x_{3i1}, x_{3i2}, \ldots)$ and $\zeta = (\zeta_1, \zeta_2, \ldots)$ being vectors of covariates and parameters associated with the correlation between suspicion ratings.

Estimating equations for both the ordinal regression model and the correlation regression model are developed and each parameter vector iteratively solved until convergence. A consistent estimate of the asymptotic variance of each vector is then obtained.

The GEE method provides consistent and asymptotically normal estimates whether or not the initial, or 'working', assumptions about the joint distribution of pairs of suspicion ratings is correct.[13] This result permits the analyst to elect to treat the correlations as nuisance parameters that can, essentially, be ignored. On the other hand, the Toledano model does permit the estimation of correlations and the modelling of the dependence of correlations on covariates as well. Therefore, the Toledano method is suitable to either goal of the analysis of clustered ROC curve data.

The carotid study is an example of a situation of clustered ROC data in which correlation between suspicion ratings would probably not be of interest since the goal

is to estimate a single ROC curve using data from both carotids. In this case, the working assumption of independence would be an entirely satisfactory, although not statistically most efficient, specification.

On the other hand, the lung study is an example for which correlation might be clinically important. Using the Toledano method, one could model the correlation between the two regions with the covariate vector

$$x_{3i} = (x_{3i1}, x_{3i2})$$

where x_{3i1} is the disease covariate and $x_{3i2} = 1 \Leftrightarrow$ the region is the bilary region.

In either case, estimation and inference on ROC curve indices, and their contrasts, can be based on Taylor series expansion using the estimated variance–covariance matrix of parameter estimates. The primary advantage of this method lies in its great flexibility. Disadvantages arise from its complexity. The examples described above considered only the single-reader, single-modality situation. With multiple readers and/or modalities, the number of pairwise correlations grows quickly. Another disadvantage is that, until software is available for general use, the method is inaccessible to most clinical researchers.

4.3 Bayesian approaches

Gatsonis[14] considered the problem of multiple ratings from studies that involve several readers and/or institutions from a hierarchical, 'random-effects', perspective. The natural hierarchy of that data (cases–readers–institutions) is reflected by the naturally nested structure of clustered ROC data. Similarly, DUOS are typically nested in patients that have been randomly sampled, so that the contribution of DUOS to total variation can be considered to be a random-effect. Another experimental setting in which DUOS could naturally be treated as random-effects are location and detection studies in which observers must identify regions on an image they suspect might have a condition and then report the degree of their suspicion. Both the number of regions and location of such regions can be thought of as occurring at random.

If there are two DUOS per subject, and the goal is to estimate a single ROC curve, the 'uniform-shift' model described by Gatsonis might be applicable. For example, in the Carotid study this model would assume that the ROC curve from the left and right carotids are identical but that diagnostic performance might differ between these two DUOS because observers select different decision thresholds, or cutoffs, for left versus right carotids.

Continuing this example, the uniform-shift random-effects model would have the form

$$\Phi^{-1}[P(Y_{ij} \leq c)] = \theta_c - \alpha x_i - d_i - \delta_j$$

where $\{\theta_c\}$, $c = 1...C-1$, is a set of common cutoffs and d_i, $i = 1, 2$, are deviations from the common cutoffs that occur when the left ($i = 1$) and right ($i = 2$) carotid is interpreted, and where it is further assumed that

$$\delta_j \sim N(0, \sigma^2)$$

and the δ_j are independent.

The random-effects component, δ_j, is required when using the random-effects approach to account for the correlation among DUOS from the same nesting unit, j. It is immediately observed that use of this model then implies a shift in cutoffs from subject to subject. This sort of assumption is not typically found in ROC studies and needs to be applied with caution as it implies the reader reacts uniquely to each subject. In some cases this may be a plausible assumption, however, since it might be seen to reflect differences between subjects in diagnostic information they present to the observer. Nonetheless, the implications of this model need to be fully addressed by the analyst.

Following Gatsonis' fully Bayesian approach, distributions on all model parameters are required. Extending Gatsonis' uniform-shift model further, and restricting covariates to disease status only *and* $C = 5$, possible priors that might be considered are (see Gatsonis[14])

$$\begin{aligned}
\theta &\sim \text{uniform on } -3 \leq \theta_1 < \ldots < \theta_4 \leq 3 \\
\alpha &\sim \text{uniform on } (0, U_1) \\
d &\sim N(0,1) \text{ and} \\
\sigma^2 &\sim \text{inverse gamma } (L_1, U_2)
\end{aligned}$$

where limits U_1, U_2 and L_1 are based on subject matter considerations.

In another model developed by Gatsonis, the 'accuracy-shift' model, allows distinct ROC curves from distinct DUOS and would be applicable to the lung study example. This would be accomplished by including anatomical region as a covariate. The model would then imply that ROC curves from different regions are 'parallel' in ROC 'space' – i.e. after transforming the ROC axes to normal deviates.[8]

Note that neither of the models so developed include a scale term, so that the assumption made is of equal variance between afflicted and unafflicted populations, regardless of DUOS. Inclusion of the scale term, of course, introduces another level of complexity and difficulty in interpretation of the model.

The posterior distribution of the parameters does not have a closed analytic form and estimation and inference must proceed by either approximation or simulation. Gatsonis developed the latter approach, using Gibb sampling[15,16] to generate samples of the joint posterior distribution of the parameters. These observations form the basis for estimating marginal densities for each parameter and for functions of the parameters.

A similar simulation approach to Bayesian analysis of correlated ROC curve data was developed by Peng and Hall.[17] This method uses 'data-augmentation'[18] to impute the unobserved continuous measurements from the latent distribution. This enables derivation of posterior distributions based on the likelihood of the imputed data and with 'flat' priors. Gibbs sampling is then used to obtain posterior sampling of the regression parameters.

Advantages of either Bayesian method to the analysis of clustered ROC data are that they allow small samples to be used and that the random-effects approach does not require complicated specification of correlations. Disadvantages are: (1) the imposed

condition that cutoffs change with each new nesting unit; (2) the required specification of many priors, even in the simplest model; and (3) computational complexity which makes this approach inaccessible to most researchers.

5 Nonparametric approaches

Nonparametric approaches for clustered ROC data have recently been developed by Obuchowski[3] and are briefly reviewed below.

The nonparametric analysis of ordinal ROC curve data is largely restricted to area estimation and inference. However, partial area analysis is possible under the general framework developed by Wieand *et al*.[5] Nonparametric analysis of ROC curve data rests on the recognition by Bamber[19] of the equivalence of the trapezoidal area estimate and the Mann–Whitney statistic. Approaches to the nonparametric analysis of ROC curve area from correlated data were developed by Hanley and McNeil,[20] who gave a semi-parametric approach, and Delong *et al*.[21] who applied the theory of multivariate U-statistics.

Following the work of Delong *et al*., Obuchowski replaced the ratings data with the following proportions.

Let X denote an afflicted DUOS and Y an unaffected DUOS. Define

$$V_{10}(X_{ij}) = \frac{1}{N}\sum_{i'=1}^{I_{01}}\sum_{k=1}^{n_{i'}}\psi(X_{ij}, Y_{i'k}) \text{ for all } X_{ij}$$

and

$$V_{01}(Y_{i'k}) = \frac{1}{M}\sum_{i'=1}^{I_{10}}\sum_{j=1}^{m_{i'}}\psi(X_{ij}, Y_{i'k}) \text{ for all } Y_{ik}$$

where I_{01} is the numbers of nesting units with at least one unafflicted DUOS, I_{10} is the number of nesting units with at least one afflicted DUOS, n_i is the number of afflicted DUOS in nesting unit i, m_i is the number of afflicted DUOS in nesting unit i and

$$\begin{aligned}\psi(X_{ij}, Y_{i'k}) &= 1.0 \text{ if } Y < X\\ &= 0.5 \text{ if } Y = X\\ &= 0.0 \text{ if } Y > X\end{aligned}$$

Then, a nonparametric estimate of the ROC curve area, A, is given by

$$\hat{A} = \sum_i\sum_j V_{10}(X_{ij})/M = \sum_i\sum_j V_{01}(Y_{i'k})/N$$

where

$$M = \sum_i m_i \text{ and } N = \sum_i n_i$$

Letting S_{10} and S_{01} represent the mean square of the V_{10} and V_{01}, and S_{11} be the mean cross product, an estimate of the variance of a common ROC curve area from clustered data (as applicable to the carotid data) is given by

$$\hat{V} = \frac{1}{M}S_{10} + \frac{1}{N}S_{01} + \frac{2}{MN}S_{11}$$

$\hat{A}$ suitably normalized, has an asymptotic $N(0,1)$ distribution provided

$$\lim_{I\to\infty} I_{10}/I_{01} \text{ is bounded and nonzero}$$

The extension required for the nonparametric analysis of distinct ROC curves from distinct DUOS is direct, by considering cross products of the transformed data. This leads to an estimate of the variance–covariance matrix of multiple ROC curve area estimates. Suitably normalized contrasts are asymptotically normally distributed.

The method developed by Obuchowski has the advantage that it does not require the specification of a complex model nor does it require specification of correlation between nested DUOS. It is also more easily executed using standard statistical software than the methods described earlier. Yet, it is still somewhat computationally challenging and so is not accessible to all researchers. Another disadvantage is that, at this time, the method is restricted to the estimation and comparison of ROC curve areas.

6 The jackknife

The jackknife has gained popularity recently for the analysis of ROC data. Much of its popularity comes from the fact that it can be used to estimate the standard errors for any ROC curve index and that it is implementable by any standard statistical software with only a very modest amount of programming required.

Similar to the nonparametric methods of Delong and Obuchowski, the jackknife replaces the original data with computed values. These computed values, called 'pseudovalues', are then treated as a set of observations to yield estimates of the variance and covariance of ROC curve indices.

In some sense, pseudovalues represent the contribution of individuals to the variability of an ROC curve index. Specifically, if the index A is to be computed from N subjects, the pseudovalue for the ith subject is computed as

$$A^* = NA - (N-1)A_{-i}$$

where A_{-i} is the index computed on the sample with subject i deleted. It is important to note that

$$\frac{1}{N}\sum_i A_i^* = A$$

The sample variance of the A_i^* provides a consistent estimate of the standard error of A.

Hajian-Tilaki *et al.*[22] developed the use of jackknifing for the analysis of clustered ROC curve data. Following their method, covariances of indices between nested DUOS are estimated from the sample covariances of pseudovalues. As in the carotid example, when distinct DUOS are thought to have a common ROC curve, a combined ROC curve index can be formed by taking a weighted average across the DUOS. Optimal weights are determined by the inverse of the variance–covariance matrix of the pseudovalues. When distinct DUOS give rise to distinct ROC curves, comparisons of indices are based on suitable functions of the variance–covariance matrix of pseudovalues. Song[23] explored the use of various forms of the jackknife when analysing correlated ROC data and compares nonparametric methods with the jackknife.

The main advantage to using the jackknife is its computational simplicity, making it the most accessible analytic option. Another advantage is that it can be used for any ROC curve index. The primary disadvantage of the jackknife method is that it is wholly limited to computing the variance of statistics and does not provide information about individual variability, whether from randomly sampled DUOS, nesting units or observers.[24]

7 Future directions

Obviously, the jackknife will continue to play an important role in the analysis of clustered ROC data for the immediate future. This method is suited for studies that are solely concerned with the average properties of diagnostic modalities. The average might be across DUOS within nesting units, average of distinct DUOS, or the average across readers in a population of readers.[25] However, the method does not estimate the variability among individuals, an aspect of diagnostic medicine that is particularly of interest to so-called 'effectiveness studies'.[24]

The advantage that the more complex, less accessible parametric methods have over the jackknife or nonparametric methods is that they can provide estimates of population variability. The price to be paid is in complex model specification and computational difficulty. Much of the reason for these limitations comes from the use of the ordinal probit regression model. This model is not a member of the family of generalized linear models, now a part of standard software packages. An alternative specification has recently been introduced in the form of the 'continuation-ratio model'.[26] This model provides the same sort of approach as does the ordinal probit regression model, but is a generalized linear model. Therefore, packages that provide mixed-effects analysis of generalized linear models will permit the fully parametric analysis of clustered ROC data without resorting to Bayesian methods or to specialized software. This method, however, needs to be developed specifically for clustered data.

Another option yet to be fully developed is bootstrapping.[27] This method reflects the contributions coming from subject variability when subjects are randomly sampled.[28] To date, however, published applications of bootstrapping in diagnostic medicine is limited. Beam *et al.*[29] used it to generate confidence intervals when sampling a

population of readers. Dorfman *et al.*[30] used bootstrapping to compare jackknifing with the Toledano method. Examples of the application of bootstrapping to the analysis of clustered ROC data need to be pursued.

8 Summary

Clustered data arise frequently in ROC studies. Clustered data are not simply a case of correlated data since nested DUOS can have different disease states. Statistical methods for clustered data must permit multiple diagnostic units per subject, each of which might have a unique disease status. If distinct DUOS contribute to a common ROC curve, statistical methods must account for the correlation between nested DUOS when aggregating data. If distinct DUOS give rise to distinct ROC curves, statistical methods for correlated data need only be considered when contrasts between the curves are desired.

The available statistical methods for clustered data vary in the range of indices that can be considered and in their accessibility to researchers. Parametric models permit all indices to be considered but, owing to computational complexity, are the least accessible of available methods. Nonparametric methods are much more accessible, but only permit estimation and inference about ROC curve area. The jackknife is the most accessible method and permits any index to be considered.

Future development of methods for clustered ROC studies should consider the continuation ratio model, which will permit the application of widely available software for the analysis of mixed generalized linear models. Another area of development should be in the adoption of bootstrapping methods to clustered ROC data.

References

1 Langlois Y, Roederer G, Chan A *et al.* Evaluating carotid artery disease. The concordance between pulsed Doppler/spectrum analysis and angiography. *Ultrasound Medical Biology* 1989; **9**: 51–63.

2 Webb WR, Gatsonis C, Zerhouni EA *et al.* CT and MR imaging in staging non-small cell bronchogenic carcinoma: report of the radiologic diagnostic oncology group. *Radiology* 1991; **178**: 705–13.

3 Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997; **53**: 567–78.

4 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.

5 Wieand S, Gail MH, James BR *et al.* A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**: 585–92.

6 McClish DK. Combining and comparing area estimates across studies or strata. *Medical Decision Making* 1992; **12**: 274–79.

7 Beam CA, Wieand HS. A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* 1991; **130**: 1065–1066.

8 Egan JP. *Signal detection theory and ROC analysis*. New York: Academic Press, 1975.

9 Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology* 1969; **6**: 487–96.

10 Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F ed.

Information processing in medical imaging. The Hague: Nijhoff, 1984: 432–45.
11 Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1987; 1–30.
12 Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996; **15**: 1807–26
13 Liang KA, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
14 Gatsonis CA. Random-effects models for diagnostic accuracy data. *Academic Radiology* 1995; **2**: S14–S21.
15 Gelfand A, Smith A. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 398–409.
16 Gelfand A, Smith A, Lee T. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* 1992; **87**: 523–32.
17 Peng F, Hall WJ. Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods. *Medical Decision Making* 1996; **16**: 404–11.
18 Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**: 528–50.
19 Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**: 387–415.
20 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; **148**, 839–43.
21 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–45.
22 Hajian-Tilaki KO, Hanley JA, Lawrence J, *et al*. Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks. Academic *Radiology* 1997; **4**: 222–29.
23 Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997; **53**: 370–82.
24 Beam CA. Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. *Academic Radiology* 1995; **2**: S4–S13.
25 Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* 1992; **27**: 723–31.
26 Smith PJ, Thompson TJ, Engelgau MM *et al*. A generalized linear model for analyzing receiver operating characteristic curves. *Statistics in Medicine* 1996; **15**: 323–33.
27 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; **1**: 54–77.
28 Beam CA. A two-stage ROC curve regression model when sampling a population of diagnosticians. Technical Report #14 Division of Biostatistics, Medical College of WI, 1996.
29 Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US Radiologists. *Archives of Internal Medicine* 1996; **156**: 209–13.
30 Dorfman DD, Berbaum KS, Lenth RV. Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. *Academic Radiology* 1995; **2**: 626–33.