Receiver Operating Characteristic (ROC) Curves

Receiver operating characteristic (ROC) analysis was developed to summarize data from signal detection experiments in psychophysics [21]. Today, the term refers to

> a method of quantifying how accurately experimental subjects, professional diagnosticians and prognosticators (and their various tools: tests or instruments yielding numerical results, combinations of data-collection and data-display devices, different amounts and types of information ...) perform when they are required to make a series of fine discriminations or to say which of two conditions or states of nature, confusable at the moment of decision, exists or will exist [50].

In biomedical applications, the two states are often referred to as diseased and nondiseased, or D+ and D− for short. Central to this analysis is the ROC curve, which displays diagnostic accuracy as *a series of pairs* of performance measures. Each pair consists of a true positive fraction (TPF) and the corresponding **false positive** fraction (FPF) for a given definition of "test" (t) positivity, t+. These fractions are calculated from the D+ and D− groups respectively, TPF as the proportion of (t+, D+) among those D+, and FPF as the proportion of (t+, D−) among those D−. In the medical literature, the term TPF is called **sensitivity** and the complement of the FPF is called **specificity**. For the performance of statistical tests, the term **power**, rather than sensitivity, tends to be used. Equivalently, one can use its complement, the **false negative** fraction (FNF, the complement of TPF) or the frequency (β) of a so-called type II error. There is no direct statistical term for specificity. Instead, statisticians again focus on the complement, using the false positive fraction (FPF, the complement of the true negative fraction, TNF) to denote the frequency (α) of what they call type I error.

Diagnostic performance is sometimes naively characterized using a *single* overall index of "accuracy", calculated as the sum of two proportions, i.e. the proportion of (t+, D+) instances plus the proportion of (t−, D−) instances, where the proportions are based on all patients undergoing the test. This index is a weighted average of sensitivity and specificity, using as weights the (particularistic) relative frequencies of the D+ and D− states. Using two measures, namely a (TPF, FPF) *pair*, avoids this arbitrariness.

Although a (TPF, FPF) pair is a big improvement over an overall accuracy index, it is often not sufficient. A single (TPF, FPF) pair still does not allow meaningful comparison of the performance of one diagnostic test with another, or even with the same test performed in another setting or by another observer, when different criteria for test positivity are used in the two instances compared. The ROC curve, in the form of a *series* of (TPF, FPF) pairs (see Figure 1), isolates a test's capacity to discriminate between a given disease and its absence, from the **confounding** influence of the decision criterion (confidence level or cutting score) that is adopted for test positivity [37, 52, 58]. A more accurate test will be located on an ROC curve closer to the top left corner than a less accurate one. A noninformative test will have an ROC curve that lies along the diagonal.

Statistical techniques to handle the full range of ROC study designs continue to be developed [3, 5, 9, 26]. Analyses can vary in complexity from deriving an ROC curve for a single diagnostic test involving numerical values derived from patients at a single institution, to complex multi-institution studies to compare two or more imaging modalities. The complexity also depends on the purpose of the discrimination test, the setting and context to which it refers, whether in the study interpretations are performed individually in real time [18] or later in "batch" mode [52], and whether the tests under study and the procedures for independent definitive determination of the true state of nature (the **gold standard**) are costly, invasive, uncomfortable or dangerous.

This latter issue can create special problems since ROC curves are strongly influenced by the source of the test material used [6]. Distortions occur when the result of the test being studied affects the subsequent work-up needed to establish a definitive diagnosis. Information available on the distribution of test results and clinical indicants in the source population can be used to remove quantitatively this "verification bias" from ROC curves [20, 30]. Other **biases** in the assessment of diagnostic tests and guidelines for circumventing the problems in prospective studies have been described [4, 7].

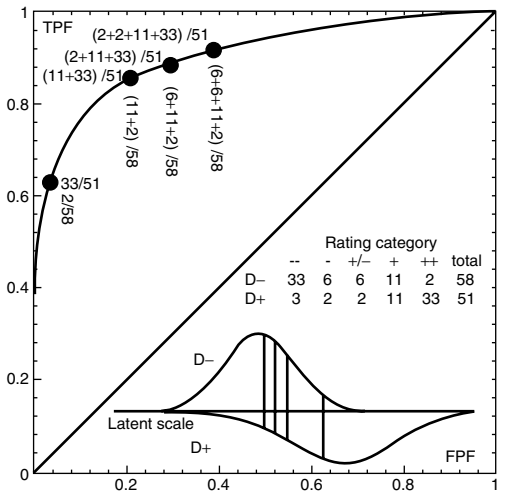

Figure 1 Example of empirical ROC points and smooth curve fitted to them. The empirical points are calculated from successively more liberal definitions of test positivity applied to the 2×5 table (inset) of disease status (D+ or D−) and rating category (−− to ++). The smooth ROC curve is derived from the fitted binormal model (inset, lower right, with parameters $a = 1.657$ and $b = 0.713$ on a continuous latent scale) by using all possible scale values for test positivity. The fitted parameters a and b, together with the four estimated cutpoints, produce fitted frequencies of $\{32.9, 6.4, 5.9, 10.7, 2.1\}$ and $\{3.2, 1.5, 2.1, 11.2, 32.9\}$ for the D− and D+ rows of the 2×5 table. Note that a monotonic transformation of the latent axis may produce overlapping distributions with nonbinormal shapes, but will yield the same multinomial distributions and the same fitted ROC curve

The complexity of the test material can have an important bearing on the ability of a study to compare tests. Cases resulting in an ROC curve that is midway between the diagonal (subtle or completely obscure ones) and the upper left corner (all 'obvious') allow for sizable differences in performance; however, the closer the curve is to the upper left corner, the narrower is the **sampling distribution** of the various indices derived from the curve [39].

For clinical imaging studies involving interpretations, the most economical method of collecting a reader's impression of each case is through the use of a rating scale, i.e. graded levels of confidence that the case is D+. A discrete five-point scale – 1 = "definitely not diseased", 2 = "probably not diseased", 3 = "possibly diseased", 4 = "probably diseased", 5 = "definitely diseased" – is commonly used. Getting a reader to use all of the rating categories provided yields a more stable ROC curve estimate, but is not always easy to accomplish without causing other problems [23]. Use of ratings from the continuous 0–100% confidence scale [31, 49] has several advantages: it more closely resembles reader's clinical thinking and reporting; its use of a finer scale leads to somewhat smaller **standard errors** of estimated indices of accuracy; and it increases the possibility that the data will allow parametric curve fitting.

Obtaining an ROC Curve and Summary Indices Derived from it

For *rating scale data*, the 2 (D states) $\times$ k (rating categories) frequency table of the ratings yields $k - 1$ empirical (TPF, FPF) ROC points. As shown in Figure 1, these are obtained from the $k - 1$ possible **two-by-two tables** formed by different re-expressions of the $2 \times k$ data table. After TPF = 0 at FPF = 0, the lowest leftmost ROC point is derived using the strictest cutpoint, where only the most positive category would be regarded as positive; each subsequent point towards the top right ROC corner (TPF = 1, FPF = 1) is obtained by employing successively laxer criteria for test positivity. For objective tests that yield *numerical data*, the same procedure – with each distinct observed numerical test value as a category boundary and with k no longer fixed a priori but rather determined by the numbers of 'runs' of D+ and D− in the aggregated data – is used to calculate the series of empirical ROC data points. The sequence of points can then be joined to form the empirical ROC curve or a smooth curve can be fitted.

As a summary measure of accuracy, one can use: (i) TPF[FPF], the TPF corresponding to a single selected FPF; (ii) the area under the ROC curve; or (iii) the area under a selected portion of the curve, often called the partial area. Summary (i) is readily understood and most clinically pertinent. However, reported TPFs are often in reference to different FPF values, and it may be unclear whether a reference FPF was chosen in advance or after inspection of the curve. Moreover, the statistical reliability tends to be lower than that of other summary indices.

Summary (ii) has been recommended as an alternative [52]. It has an interpretation in signal detection theory as the proportion of correct choices in a two-alternative forced choice experiment [21], i.e. an

experiment where in each trial the subject is presented with a pair of stimuli, one from a randomly chosen D+ and one from a randomly chosen D−, and is asked to decide which derives from which. This method of reporting judgments is common in psychophysics and has statistical advantages when using synthetic images, where observer time is the limiting factor [8].

To some, the area index has a serious limitation. Since a large part of the area comes from the rightmost part of the curve, it includes FPFs of no clinical relevance, and so can be insensitive when used to compare the performance of two tests. One curve may have higher TPFs than another in the region of relevant FPFs, but they could conceivably cross. Since the area under the entire curve averages the sensitivity over the full (0, 1) range of FPFs, any superiority in the relevant FPF region may be lost, or even reversed, when the curves are ranked on the basis of the entire area. The average sensitivity (TPF) over a range of relevant FPFs, summary index (iii) [34, 53, 60], is a compromise between (i) and (ii).

All three indices can be calculated either from the empirical or a (parametrically fitted) smooth curve. The statistical precision of nonparametric estimates can be calculated using a general method applicable to all three indices [60]; in the case of (ii) other essentially equivalent but less cumbersome methods, based on ***U*-statistics**, are also available [11, 28]. The case of (i) is more subtle than most realize: the standard error of an estimated TPF must include, in addition to its own obvious **binomial** variation, the uncertainty associated with determining the position of the FPF point [32].

A smooth ROC curve can be fitted to rating scale data by fitting two overlapping distributions on a continuous but "latent" scale underlying the results for D− and D+ cases [12, 36] In the most commonly used, "binormal", model, the two distributions are taken to be, without loss of generality, N(0, 1) for D−, and N(μ, σ) for D+. The distributions of the ratings are thus **multinomial**, with **expectations** that are functions of the $k-1$ cutpoints and the two parameters μ and σ, allowing the $k+1$ (two relevant and $k-1$ **nuisance**) parameters to be fitted to the observed data table ($2k-2$ **degrees of freedom** in total, leaving $k-3$ degrees of freedom to test the fit) using the criterion of **maximum likelihood**. Small additions to empty cells can be used to avoid "degenerate" situations [13]. The extent to which the normal deviate (*see* **Normal Scores**) **transformations** (z[TPF], z[FPF]) of the empirical (TPF, FPF) pairs are linear provides a visual test of the fit, since under the binormal model their expectations satisfy

$$z(\text{TPF}) = (1/\sigma)z(\text{FPF}) - \mu/\sigma = bz(\text{FPF}) - a.$$

When $b = 1$, the curve in (TPF, FPF) space is symmetric about the negative diagonal, while $b < 1$ produces a curve which rises more steeply at first and "flattens out" at the end.

Since the various summary indices derived from the fitted curve are functions of the estimates of a and b, their statistical precision – used in tests (*see* **Hypothesis Testing**) and **confidence intervals** – can be calculated from the corresponding variance/covariances provided by the maximum likelihood procedure. Confidence intervals can also be calculated for the entire curve [33].

For rating scale data, the popularity of the binormal model over bilogistic [22, 47] or other competitors [16] is more historical than theoretical. Use of a binormal model for rating data does not imply that if one could observe the latent distributions, they would have this exact form [38]. Rather, the working assumption is that the two overlapping multinomial distributions can be mathematically predicted from the discretization of two **normal distributions** on some unspecified latent scale. Whereas any two overlapping distributions will uniquely determine a specific ROC curve, the reverse is not true: the "binormal" assumption concerns only the functional form of the ROC curve, which can always be examined empirically, and not the form of the underlying distributions themselves, which cannot be determined in many applications of ROC analysis [38]. Use of a small number of rating categories, with few degrees of freedom, to distinguish the fit of one specific form over another, leaves considerable freedom to fit different distributional forms. This freedom is not a function of sample sizes (numbers of cases) but of the number of rating categories [25].

More important than the choice of distributional family seems to be the need to allow for unequal **variances** ($b \neq 1$). Empirically, b tends to be less than 1 [51], possibly because of the presence of unidentified subtypes in the D+ sample. Thus, whereas one-parameter models, with $b = 1$, would have practical advantages, particularly for **meta-analyses** and for fitting an entire (but symmetric)

ROC curve to a single empirical (TPF, FPF) data point, they are not supported by empirical findings.

Care must be taken in the fitting of parametric curves to results recorded on a numerical scale: directly fitting $N(\mu_1, \sigma_1)$ for D− and $N(\mu_2, \sigma_2)$ for D+ can yield severe distortions when the data do not arise from normal distributions [19]. The method in which the raw data are first categorized and the categorized data analyzed as if they were rating data [41] – with the assumptions of overlapping normal distributions on an unspecified transform of the actual measurement scale – is a much more **robust** approach.

Comparison of ROC Curves

Because of the need to account for large differences in case difficulty, compared curves are usually based on the same set of cases, and one must therefore take the **correlation** of the estimated curves – and summaries derived from them – into account when calculating the standard error of differences.

Parametric methods for comparing two curves are based on the estimates of the binormal (or bilogistic, or other model) parameters (a, b) associated with each curve, and their variances and covariances [34, 42]. The equality of two curves can be assessed by testing the equality of the two vectors (a_1, b_1) and (a_2, b_2). Comparisons involving a summary index are made by computing, for each curve, the appropriate function of the parameter estimates, then using the **delta method** to calculate the standard error of the difference in indices.

A **nonparametric method** is now available to compare two curves based on continuous data from the same set of cases [59]. A criterion for positivity that is common for the two tests is induced using the **ranks** in the combined D+ and D− data for each test. Using this calibration, one first computes the difference in the numbers of errors made by the two tests at each possible level of test positivity and then calculates the average of the absolute differences over the different levels. The test statistic is referred to the permutation distribution obtained by randomly interchanging pairs of ranks. Nonparametric comparisons of the areas under two curves are based on correlated U-statistics [11] or equivalently on the **jackknife method** [27], while partial areas, and – ultimately – sensitivity at a single specificity value can be compared with a more general method [60].

Guidelines for **sample size determination** and **power** calculations are available for both parametric [software program ROCPWR from Charles Metz at the University of Chicago, or the article by Obuchowski & McClish [43]] and nonparametric approaches [29].

Comparison of Accuracy of Imaging Procedures

The methods just described deal only with simple comparisons of two tests that yield objective numerical results, and are not sufficient for imaging studies (*see* **Image Analysis and Tomography**) which produce interpretations of each case by multiple readers. For example, the performance with conventional versus laser printed films might be studied by having several readers interpret each image; a comparison of computed tomography, magnetic resonance and ultrasound images might involve different readers for each modality [46]. Several refinements and some alternatives to the method initially proposed for dealing with these more complex comparisons have been suggested. The challenge is to include and estimate properly each of the several relevant **components of variance** and covariance since the comparison is necessarily an average over cases, readers and (possibly) rereadings [40, 52]. Two methods [15, 44] deal with the problem by modeling the variation in the summary index in question, while another [54] models the raw rating data responses. From the investigations thus far [14, 55], both modeling approaches appear to give comparable answers, but commentators [48, 35] have called for some further work to investigate the performance of analysis strategies that use statistical tests to decide what is the appropriate error term and denominator **degrees of freedom** when reader x modality **interactions** are involved.

When studies of imaging procedures involve multiple centers, complex procedures, ethical concerns, "real-time" readings, and different experts in the different imaging modalities [18], the data can quickly become imbalanced and/or incomplete. Analysis problems are thus aggravated by the subjective (and thus possibly nonpoolable) nature of the ratings, the often large numbers of case–reader sets, each containing too few observations to allow parametric fitting of separate ROC curves, and the fact that,

unlike the usual response measures in **clinical trials**, the elemental ROC data are not absolute numbers that can be easily averaged or displayed individually in a descriptive way. If all available data are to be used, the only logical approach is to use **regression** methods. Since the first regression work in this area [57], there has been considerable activity in developing **parsimonious** approaches to the problem, including **random effects** models [2], Gibbs sampling (*see* **Markov Chain Monte Carlo**) [17], and **generalized estimating equations** [56, 61].

In some situations, the study material may involve more than one region of interest on an image. Sample reuse methods can help to calculate the precision with which statistical contrasts are made [1, 10, 24, 45] (*see* **Bootstrap Method**).

Software

Programs for parametric estimation are available from the WWW location `www-radiology.uchicago.edu/cgi-bin/software.cgi` maintained by the developer, Charles Metz. Special-purpose programs for nonparametric inference are more numerous, but most of the tasks can be accomplished using a spreadsheet [27]. Software for the multireader, multicase approach to imaging data is available from the authors of ref. [15]; software for the other approaches is still evolving and interested users should contact the various authors cited above.

Future Developments

Whereas methods for comparing ROC curves associated with tests that yield objective test results have now become routine, solutions to the complex analytic problems involved in the comprehensive comparison of accuracy of imaging procedures have not been completely achieved. The methods proposed in the last 5 years need further testing; the links between them need to be better understood; and user-friendly software to implement these newest approaches remains to be developed.

References

[1] Baum, R.A., Rutter, C.M., Sunshine, J.H., Blebea, J.S., Carpenter, J.P., Dickey, K.W., Quinn, S.F., Gomes, A.S., Grist, T.M. et al. (1995). Multicenter trial to evaluate vascular magnetic resonance angiography of the lower extremity. American College of Radiology Rapid Technology Assessment Group, *Journal of the American Medical Association* **274**, 875–880.

[2] Beam, C. (1995). Random effects models in the ROC curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches and issues, *Academic Radiology* **2**, Supplement 1, S4–S13.

[3] Begg, C.B. (1986). Statistical methods in medical diagnosis, *Critical Reviews in Medical Informatics* **1**, 1–22.

[4] Begg, C.B. (1987). Biases in the assessment of diagnostic tests, *Statistics in Medicine* **6**, 411–424.

[5] Begg, C.B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980s, *Statistics in Medicine* **10**, 1887–1895.

[6] Begg, C.B. & Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to verification bias, *Biometrics* **39**, 207–215.

[7] Begg, C.B. & McNeil, B.J. (1988). Assessment of radiologic tests: control of bias and other design considerations, *Radiology* **167**, 565–569.

[8] Burgess, A.E. (1995). Comparison of receiver operating characteristic and forced choice observer performance measurement methods, *Medical Physics* **22**, 643–655.

[9] Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Statistics in Medicine* **13**, 499–508.

[10] Dagirmanjian, A., Ross, J.S., Obuchowski, N., Lewin, J.S., Tkach, J.A., Ruggieri, P.M. & Masaryk, T.J. (1995). High resolution, magnetization transfer saturation, variable flip angle, time-of-flight MRA in the detection of intracranial vascular stenoses, *Journal of Computer Assisted Tomography* **19**, 700–706.

[11] DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver-operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837–845.

[12] Dorfman, D.D. & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data, *Journal of Mathematical Psychology* **6**, 487–496.

[13] Dorfman, D.D. & Berbaum, K.S. (1995). Degeneracy and discrete receiver operating characteristic rating data, *Academic Radiology* **2**, 907–915.

[14] Dorfman, D.D. & Metz, C.E. (1995). Rejoinder in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S76–S78.

[15] Dorfman, D.D., Berbaum, K.S. & Metz, C.E. (1992). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method, *Investigative Radiology* **27**, 723–731.

[16] Egan, J.P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, New York.

[17] Gatsonis, C. (1995). Random-effects models for diagnostic accuracy data, *Academic Radiology* **2**, Supplement 1, S14–S21.

[18] Gatsonis, C. & McNeil, B.J. (1990). Collaborative evaluation of diagnostic tests: experience of the Radiologic Diagnostic Oncology Group, *Radiology* **175**, 571–575.

[19] Goddard, M.J. & Hinberg I (1990). Receiver operating characteristic (ROC) curves and non-normal data: an empirical study, *Statistics in Medicine* **9**, 325–337.

[20] Gray, R. & Begg C.B. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making* **4**, 151–164.

[21] Green, D.M. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

[22] Grey, D.R. & Morgan, B.J.T. (1972). Some aspects of ROC curve fitting: normal and logistic models, *Journal of Mathematical Psychology* **9**, 128–139.

[23] Gur, D., Rockette, H.E., Good, W.F., Slasky, B.S., Cooperstein, L.A., Straub, W.H., Obuchowski, N.A. & Metz, C.E. (1990). Effect of observer instruction on ROC study of chest images, *Investigative Radiology* **25**, 230–234.

[24] Hajian-Tilaki, K.O., Hanley, J.A., Joseph, L. & Collet, J.P. (1997). Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks, *Academic Radiology* **4**, 222–229.

[25] Hanley, J.A. (1988). The robustness of the binormal model used to fit ROC curves, *Medical Decision Making* **8**, 197–203.

[26] Hanley J.A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging* **29**, 307–335.

[27] Hanley, J.A. & Hajian-Tilaki K.O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update, *Academic Radiology* **4**, 49–58.

[28] Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under an ROC curve, *Radiology* **143**, 129–133.

[29] Hanley, J.A. & McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same set of cases, *Radiology* **148**, 839–843.

[30] Hunink, M.G., Richardson, D.K., Doubilet, P.M. & Begg, C.B. (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing, *Medical Decision Making* **10**, 201–211.

[31] King, J.L., Britton, C.A., Gur, D., Rockette, H.E. & Davis, P.L. (1993). On the validity of the continuous and discrete confidence rating scales in receiver operating characteristic studies, *Investigative Radiology* **28**, 962–963.

[32] Linnet, K. (1987). Comparison of quantitative diagnostic tests: type I error, power and sample size, *Statistics in Medicine* **6**, 147–158.

[33] Ma, G. & Hall, W.J. (1993). Confidence bands for receiver operating characteristic curves, *Medical Decision Making* **13**, 191–197.

[34] McClish, D.K. (1989). Analyzing a portion of the ROC curve, *Medical Decision Making* **9**, 190–195.

[35] McClish, D.K. (1995). Invited discussion in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S61–S64.

[36] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

[37] Metz, C.E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine* **8**, 283–298.

[38] Metz, C.E. (1986). ROC methodology in radiological imaging, *Investigative Radiology* **21**, 720–733.

[39] Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies, *Investigative Radiology* **24**, 234–245.

[40] Metz, C.E. & Shen, J.H. (1992). Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis, *Medical Decision Making* **12**, 60–75.

[41] Metz, C.E., Shen, J.H. & Herman, B.A. (1990). New methods for estimating a binormal ROC curve from continuously distributed test results. Paper presented at the annual meeting of the American Statistical Association, Anaheim.

[42] Metz, C.E., Wang, P-L. & Kronman, H.B. (1984). A new approach for testing the significance of differences between ROC curves from correlated data, in *Information Processing in Medical imaging*, F. Deconink, ed. Martinus Nijhoff, The Hague, pp. 432–445.

[43] Obuchowski, N.A. & McClish, D.K. (1997). Sample size determination for diagnostic accuracy studies involving binormal r.o.c. curve indices, *Statistics in Medicine* **16**, 1529–1542.

[44] Obuchowski, N.A. (1995). Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using analysis of variance with dependent observations, *Academic Radiology* **2**, Supplement 1, S22–S29.

[45] Obuchowski, N.A. (1996). Nonparametric analysis of clustered ROC data. Presentation at Eastern North American Biometrics Meeting.

[46] Obuchowski, NA. & Zepp, RC. (1996). Simple steps for improving multiple-reader studies in radiology, *American Journal of Roentgenology* **166**, 517–521.

[47] Ogilvie, J.C. & Creelman, C.D. (1968). Maximum likelihood estimation of ROC curve parameters, *Journal of Mathematical Psychology* **5**, 377–391.

[48] Rockette, H.E. (1995). Contributed comments in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S70–S71.

[49] Rockette, H.E., Gur, D. & Metz, C.E. (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques, *Investigative Radiology* **27**, 169–172.

[50] Swets, J.A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models, *Psychological Bulletin* **99**, 100–117.

[51] Swets, J.A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psychological Bulletin* **99**, 181–198.

[52] Swets, J.A. & Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

[53] Thompson, M.L. & Zucchini W. (1989). On the statistical analyses of ROC curves, *Statistics in Medicine* **8**, 1277–1290.

[54] Toledano A. & Gatsonis, C.A. (1995). Regression analysis of correlated receiver operating characteristic data, *Academic Radiology* **2**, Supplement 1, S30–S36.

[55] Toledano A. & Gatsonis, C.A. (1995). Rejoinder in Symposium on Advances in Statistical Methods for Diagnostic Radiology, *Academic Radiology* **2**, Supplement 1, S81–S82.

[56] Toledano, A.Y. & Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data, *Statistics in Medicine* **15**, 1807–1826.

[57] Tosteson, A.N.A. & Begg, C.B. (1988). A general regression methodology for ROC curve estimation, *Medical Decision Making*, **8**, 204–215.

[58] Turner, D.A. (1978). An intuitive approach to receiver operating characteristic curve analysis, *Journal of Nuclear Medicine* **19**, 213–220.

[59] Venkatraman, E.S. & Begg, C.B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika* **83**, 835–848.

[60] Wieand, S., Gail, M.H., James, K.L. & James, B.R. (1988). A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data, *Biometrika* **76**, 585–592.

[61] Zhou, X.H. (1996). Empirical Bayes combination of estimated areas under ROC curves using estimating equations, *Medical Decision Making* **16**, 24–28.

(*See also* **Diagnostic Test Evaluation Without a Gold Standard**; **Diagnostic Tests, Evaluation of**)

J.A. HANLEY

Record Linkage

At the core of all **descriptive epidemiology** studies lies a data set, with many variables, which has been gathered to answer a specific hypothesis. Often it is only as the project develops that the researcher realizes the potential of exploring alternate study endpoints by adding in other data about the same respondents. The tried and tested technique is for a clerk to look at the individual records, sorted in some logical order, and put the records together, applying intuitive decision rules based on human judgment. As record systems have been computerized over the past 20 years, one of the greatest impacts of increased processing power has been to facilitate linkages between related data sets, even when they do not share a unique identifier.

Three main techniques are used for record linkage, Newcombe [13] and Jamieson et al. [9] describe in detail the technical issues relating to exact matching and probability linkage (*see* **Matching, Probabilistic**). They can be summarized as follows:

1. *Unique*. Records are linked together where unique identifiers such as insurance number or health service number match exactly. The files of records are computer sorted into the same order, and matched together within blocks. It is a fairly simple process, but may only identify 80%–85% of true matches due to errors in recording of identifiers.
2. *Fuzzy*. For data sets which do not have unique identifiers, key identifiers such as surname, date of birth, sex, date of interview/treatment, and postal district are used for linkage. To cope with coding errors, fuzzy matching identifies records which are "almost" the same, such as surname spelling incorrect, or year and month of birth correct, but day wrong. Computer programs either present a choice of matches for the user to choose the best match, or have incorporated a simple scoring system and determine the best match from the score. Computer **algorithms** are well developed for matching on individual variables, and this technique provides 85%–90% of true matches. It requires human intervention and there may be operator **bias**.
3. *Probability*. This is the most sophisticated form of linkage, in which decision rules on records matching are programmed based on the probability of two records being from different people having the same identifier. These probabilities are aggregated to a score and checked against a threshold to determine whether a match is made. The computer system needs to be tailored for the data sets to be matched and is processor intensive, but provides linkages of 95%–99% true matches with false positive rates of 1%–2%.

The following are examples of the uses made of record linkage within healthcare systems and demonstrate the value of this powerful technique. Many of the examples come from uses made of the Scottish Medical Record Linkage Database [7], which contains morbidity records from Scottish hospitals, and mortality records from the General Register Office (Scotland) from 1968 onwards – almost 4 million people with 12 million episodes.

Medical record linkage poses problems of data **confidentiality** and privacy, because the linked data are comprehensive and the techniques use personal identifying data. Most analytic studies do not require access to patient identifiable data once the linkages have been made. For administrative data sets, strict controls need to be in place to ensure that the data are not released to individuals and used for purposes other than those registered in government legislation.

The issue of infringing civil liberties, by invasion of privacy through wrongful use of information, is currently taxing most governments. Researchers need to be aware of legislation and appropriate use of data. For example, in Scotland, access to identifiable data is controlled by medically qualified data holders, and a Privacy Advisory Committee [10] has been established, with membership drawn from senior medical officers, legal professions, and the public, to ensure that ethical approval is in place for record linkage studies.

Evidence-Based Medicine

The perception remains that descriptive epidemiology has little to contribute to the development of **evidence-based medicine** with its focus on randomized **clinical trials** (see McPherson [12]). Probability-based record linkage techniques can make a major contribution in assessing the efficacy of treatment regimes at the macro level. For example, using exact

matching on health service administration numbers, Evans et al. [4] at the Tayside Medicines Monitoring Unit are able to use **case–control** methodology to review the association of topical nonsteroidal anti-inflammatory drugs with hospital admission for upper gastrointestinal bleeding and perforation. The Scottish Health Service are now automating the linkage between prescribing data, patient hospitalization, and death profiles. This will establish a facility for **post-marketing surveillance**, in which possible adverse drug reactions can be quickly analyzed and assessed before the public are alarmed by the media (*see* **Pharmacoepidemiology, Overview**).

Another area in which linkage is being used effectively is the follow-up of very low birthweight children and the impact of their improved survival on health care costs. In California [2], probability-linked data from the California Birth Cohort and Medicaid claims in years after birth have been used to evaluate competing hypotheses for racial and ethnic differences (*see* **Ethnic Groups**) in mortality and health care costs, and to assess the need for hospital services from the improved neonatal survival of these children.

Outcome Measurement

Evaluating the effects of medical care is not a new idea, but it has received increased emphasis over the past decade because of concerns for the quality and cost of medical care (*see* **Quality of Care**). While most attention has been placed on determining the effectiveness of new treatment regimes through randomized control trials, the inclusion criterion for patients can be so selective (*see* **Eligibility and Exclusion Criteria**) that the true efficacy of the treatment can only be assessed when it comes into general usage. Application of record linkage techniques using **administrative databases** for follow-up of cohorts of patients with specific disease patterns, or procedures, permits analysis of outcomes measures which would otherwise be prohibitively expensive.

The Clinical Resource and Audit Group of the Scottish Office Department of Health have pioneered the publication of routine clinical outcome measures in the UK since 1993. The three reports [17] to date have been produced following detailed consultation with health service professionals, to gain consensus on the measures and to assess the feasibility of using them to monitor the effectiveness and appropriateness of health purchasing strategies. Without a unique patient identifier, probability linkage is the key to determining readmission rates, including to other institutions, and postoperation survival after discharge.

While the measures tend to be presented as interval estimates (*see* **Estimation, Interval**), standardized for **confounding** factors, such as age, sex, deprivation, and co-morbidities, administrative data do not yet contain **robust** measures of severity of disease. As with all descriptive epidemiology techniques, the outcome measures highlight topics for more detailed investigation via randomized controlled trials or clinical audit.

Survival Rates

As the search continues for new, meaningful outcome measures, one of the main uses of record linkage has been analysis of survival patterns for disease, especially for cancer (*see* **Survival Analysis, Overview**). Most civil registration authorities provide an exact matching service for bona fide researchers. However, increased computing power has meant that this process can now be automated to include probability or "fuzzy" matching techniques, which increase the reliability of the links. Within the Scottish Cancer Registry, we found that exact matching with manual techniques under-ascertained almost 5% of deaths [16], because the procedures were built on zero tolerance of **false positive** rates. This resulted in one study for the nuclear industry showing a "healthy worker" effect (*see* **Occupational Epidemiology**), until another three deaths were determined by automated probability linkage among the cohort of employees.

The availability of population-based data in specific **diseases registers**, such as cancer, diabetes, and renal failure, with linkage to death registrations enables the development of survival tables [16] (*see* **Life Table**), which are of use not only to the professional dealing with individual patients but also to the patients and their carers. Society is becoming more attuned to the concepts of **risk**, and one of the most common questions asked when life-threatening disease is diagnosed is "What is my chance of surviving 1 year, 5 years, or 10 years?". Insurance companies are very interested in improved estimates of