doi:10.1111/j.1365-2753.2005.00598.x

Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis

Ariel Linden DrPH MS
President, Linden Consulting Group, Portland, OR, USA and Oregon Health and Science University, School of Medicine, Department of Preventive Health/Preventive Medicine, Portland, OR, USA

Correspondence
Ariel Linden,
Linden Consulting Group
6208 NE Chestnut Street
Hillsboro OR 97124
USA
E-mail: alinden@lindenconsulting.org

Keywords: diagnostic accuracy, disease management, ROC, sensitivity, specificity

Accepted for publication:
14 December 2004

Abstract

Diagnostic or predictive accuracy concerns are common in all phases of a disease management (DM) programme, and ultimately play an influential role in the assessment of programme effectiveness. Areas, such as the identification of diseased patients, predictive modelling of future health status and costs and risk stratification, are just a few of the domains in which assessment of accuracy is beneficial, if not critical. The most commonly used analytical model for this purpose is the standard 2 × 2 table method in which sensitivity and specificity are calculated. However, there are several limitations to this approach, including the reliance on a single defined criterion or cut-off for determining a true-positive result, use of non-standardized measurement instruments and sensitivity to outcome prevalence. This paper introduces the receiver operator characteristic (ROC) analysis as a more appropriate and useful technique for assessing diagnostic and predictive accuracy in DM. Its advantages include; testing accuracy across the entire range of scores and thereby not requiring a predetermined cut-off point, easily examined visual and statistical comparisons across tests or scores, and independence from outcome prevalence. Therefore the implementation of ROC as an evaluation tool should be strongly considered in the various phases of a DM programme.

Introduction

Disease management (DM) is a system of coordinated interventions aimed to improve patient self-management as well as increase doctors' adherence to evidence-based practice guidelines. The assumption is that by augmenting the traditional episodic medical care system with services and support between doctor visits, the overall cost of health care can be reduced (DMAA 2004).

Diagnostic or predictive accuracy concerns are common in all phases of a DM programme, and ultimately play an influential role in the assessment of programme effectiveness. For example, (1) accurate identification of diseased patients is essential for programme inclusion. Most programmes rely on medical and pharmacy claims data for this purpose; however, this method is notoriously unreliable (Hannan *et al*. 1992; Jollis *et al*. 1993). (2) Predictive models are typically used as an initial stratification tool to forecast patients' use of services in future periods. These tools also typically rely on past claims experience and are thereby limited in both the accuracy of those data as well as the statistical model used for the prediction (Weiner 2003). (3) During the initial patient contact, the DM nurse typically performs an assessment of the patient's disease severity level to determine the intensity of DM services that will be required. Misclassification may result in the patient receiving either too much or too little on-going attention. (4) Throughout the programme intervention, accuracy is needed in assessing a patient's level of self-management and their progression across the stages of behavioural change (Linden & Roberts 2004). DM nurses may differ in their psycho-social behavioural modification skills and thus in their ability to effect and accurately rate a patient's level of progress.

This paper introduces the concept of receiver operating characteristic (ROC) analysis as a means of assessing accuracy in the programmatic domains described above. Several examples will be presented with discussion so that this technique can be easily replicated in DM programmes. For those organizations that purchase DM services, this paper will provide a substantive background with which to discuss the inclusion of ROC as an integral component of the programme evaluation with their contracted vendors.

Conventional measures of accuracy

The most common method for assessing diagnostic accuracy classifies the outcome measure as a binary variable in which the result either occurs or does not (e.g. disease or no disease, reached high cost level or not, improved health status or not, etc.) and presents the analytical model using standard 2 × 2 tables such as that shown in Table 1. As an example, the data in this table allows the analyst to calculate the proportion of patients whose diagnosis (or lack thereof) was correctly predicted by the model (true positives and true negatives). Sensitivity is the proportion of true positives that were correctly predicted by the model as having the diagnosis: $A/(A + C) \times 100\%$. Specificity is the proportion of true negatives that were correctly predicted by the model as not having the diagnosis: $D/(B + D) \times 100\%$. False-negatives (FN) are those patients with the diagnosis not predicted as such by the model: $C/(A + C) \times 100\%$. False-positives (FP) are those patients not having the diagnosis but categorized as such by the model: $B/(B + D) \times 100\%$. The positive predictive value (PPV) refers to those patients with the diagnosis who were predicted by the model to have the diagnosis: $A/(A + B)$. The negative predictive value (NPV) refers to those patients without the diagnosis who were predicted by the model not to have the diagnosis: $D/(C + D)$. The disease prevalence rate is the proportion of the sample with the diagnosis or disease: $[(A + C)/(A + B + C + D) \times 100\%]$.

Table 1 An assessment of accuracy of a model in predicting the presence or absence of a diagnosis

Model Prediction	*Diagnosis*		
	Positive	*Negative*	*Total*
Positive	True Positive A	False Positive B	A + B
Negative	False Negative C	True Negative D	C + D
Totals	A + C	B + D	A + B + C + D

A perfect test, measure or predictive model would have 100% sensitivity and 100% specificity, thereby correctly identifying everyone with the diagnosis and never mislabelling people without the diagnosis. In reality however, few measures are that accurate. The primary limitation of this traditional method is the reliance on a single defined criterion or cut-off, for determining a true-positive result (or conversely a true-negative result). Bias is introduced when the cut-off is set at an inappropriate level. This may occur when the criterion is not determined through evidence-based research, or when that criterion is not generalizable across various populations or subgroups (Linden *et al*. 2004).

For example, the National Institutes of Health (NIH) have established cut-off points using body mass index (BMI) for overweight and obesity at 25 kg m^{-2} and 30 kg m^{-2} respectively (NIH 1998). However, the use of BMI to predict percent body fat (considered the gold-standard criterion for the diagnosis of obesity) has also been shown to have several limitations. Several studies have shown that ethnicity, age and sex may significantly influence the relationship between percent body fat and BMI (MacDonald 1986; Wang *et al*. 1994; Gallagher *et al*. 1996; Wellens *et al*. 1996; Deurenberg *et al*. 1998). Therefore, relying on the NIH cut-off criteria may lead one to inaccurately label an individual as obese when in fact they are not, or fail to classify an individual as overweight when in fact they are.

Another source of bias is introduced when the measurement instrument is not standardized. For example, diabetic control is typically assessed using an assay to measure hemoglobin A1c (HbA1c) levels in the blood. This measure represents the average blood glucose level of approximately the past 4 weeks, strongly weighted toward the most recent 2 weeks (Mortensen & Volund 1988). However, several different laboratory tests have been introduced that measure slightly different subtypes with different limits for normal values and thus different interpretive scales. Even though the American Diabetes Association (ADA) has established guidelines for standardizing HbA1c assays (ADA 2000) many laboratories still use various methods. Therefore, in comparing HbA1c values across patients one must consider the normal range for each laboratory.

Another significant limitation of this method is that the predictive values of the test or model are highly sensitive to the prevalence rate of the observed outcome in that population evaluated (Altman *et al*. 1994). When the population has a high prevalence of that outcome, PPV increases and NPV decreases. Conversely, when there is low outcome prevalence, PPV decreases and NPV increases. So, for example, in a diabetic population where nearly everyone's HbA1c value is within normal range (typically set at <7.0%), it would be much easier to predict a person's likelihood of being in diabetic control, and much harder to predict who will have HbA1c values outside of that normal range.

Table 2 presents a hypothetical example in which the accuracy of a predictive model to identify asthmatics is assessed. The results of this analysis indicate that 83.3% of true asthmatics were correctly identified as such by the predictive model (sensitivity) while 1 out of every 2 healthy individuals were incorrectly classified as being an asthmatic (specificity). These findings indicate that in this particular situation, the predictive model was much better at detecting the presence of asthma than correctly noting the absence of the disease.

Also in this example, the probability of presence of disease among those who were test positives (PPV) was 0.71, the probability of absence of disease among those who were test negatives (NPV) was 0.67, and the asthma prevalence was 60%. As noted earlier, these values are very sensitive to changes in the disease prevalence. Table 3 illustrates this effect of changes in prevalence on PPV and NPV holding the sensitivity and specificity of the model constant at 83.3% and 50% respectively. As shown, when the prevalence rate of asthma in the population tested is 0.94, the probability of correctly classifying an individual as being an asthmatic is 96%. Conversely, when the prevalence is only 0.13 of the sample, the PPV falls to a mere 20%.

Principles of receiver operator characteristic curves

Given the limitations of the conventional measures of accuracy described above, a more robust tool is needed for measuring diagnostic and predictive accuracy in DM. ROC analysis was initially developed in the field of statistical decision theory but its use was broadened in the 1950s to the field of signal detection theory as a means of enabling radar operators to distinguish between enemy targets, friendly forces and noise (Proc IEEE 1970; Collision 1998). The introduction of ROC analysis into the biomedical field came via the radiological sciences where it has been used extensively to test the ability of an observer to discriminate between healthy and diseased subjects using a given radiological diagnostic test, as well as to

Table 2 A hypothetical example for assessing a predictive model's accuracy in correctly identifying asthmatics

	Diagnosis of Asthma		
Model Prediction	*Positive*	*Negative*	*Total*
Positive	500	200	700
Negative	100	200	300
Totals	600	400	1000

Table 3 The effect of different prevalence rates on positive predictive value (PPV) and negative predictive value (NPV), when holding sensitivity and specificity constant

Sensitivity (%)	*Specificity (%)*	*Prevalence*	*PPV*	*NPV*
83	50	0.13	0.20	0.95
83	50	0.60	0.71	0.67
83	50	0.94	0.96	0.17

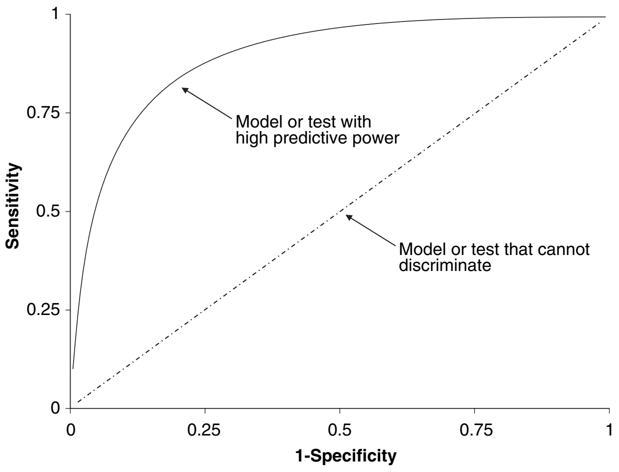

Figure 1 Hypothetical ROC curves with high and no discriminatory ability. ROC, receiver operator characteristic.

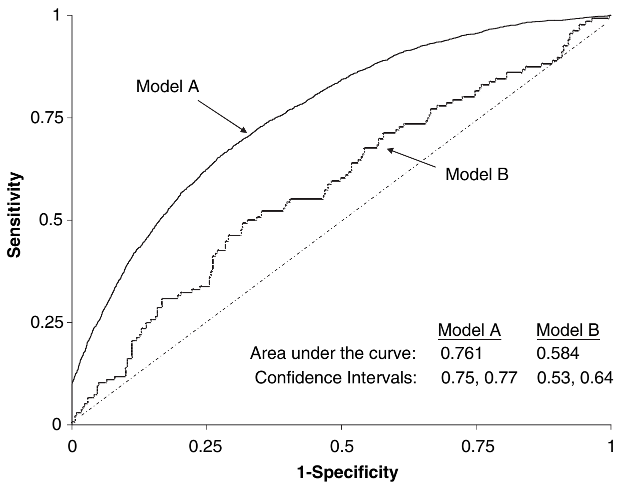

Figure 2 A comparison of the areas under the curve for two ROC curves. ROC, receiver operator characteristic.

compare the efficacy among the various tests available (Lusted 1971; Metz 1978; Swets 1979; Hanley & McNeil 1982; Metz 1986; Hanley 1989; Metz & Shen 1992).

ROC analysis involves first obtaining the sensitivity and specificity of every individual in the sample group (i.e. both subjects with and without the diagnosis or chosen outcome) and then plotting sensitivity vs. 1-specificity across the full range of values. Figure 1 illustrates a hypothetical ROC curve. A test that perfectly discriminates between those with and without the outcome would pass through the upper left hand corner of the graph (indicating that all true positives were identified and no false positives were detected). Conversely, a plot that passes diagonally across the graph indicates a complete inability of the test to discriminate between individuals with and without the chosen diagnosis or outcome. Visual inspection of the figure shows that the ROC curve appears to pass much closer to the upper left hand corner than to the diagonal line.

So that one does not have to rely on visual inspection to determine how well the model or test performs, it is possible to assess the overall diagnostic accuracy by calculating the area under the curve (AUC). In keeping with what was stated above, a model or test with perfect discriminatory ability to will have an AUC of 1.0, while a model unable to distinguish between individuals with or without the chosen outcome will have an AUC of 0.50.

While knowing the AUC provides a general appreciation for the model's ability to make correct classifications, its true value comes when comparisons are made between two or more models or tests. Figure 2 provides a comparison between two predictive models. As illustrated, Model A is both visually and statistically better at identifying individuals with and without the outcome than Model B. Therefore, if a choice was to be made between selecting one model or the other, Model A would be the choice.

Disease management programmes typically rely on medical (inpatient and outpatient services) and pharmacy claims data to identify individuals with a given disease. However, a complete array of data is not always available. Moreover, diagnosis codes used in claims data may be incorrect or non-specific, which may lead to a number of inaccurate classifications. ROC analysis can be used in this situation to examine the efficacy of the model using the different decision thresholds. For example, asthmatics may be identified from pharmacy claims data if a prescription was filled for a bronchodilator, beta-agonist, or both. However, a patient presenting with an upper respiratory infection may also be prescribed these medications. Thus, there is a risk of a false positive identification for asthma. This concern may be somewhat reduced by requiring that at least two prescriptions be filled over the course of a given period in order to classify that individual as an asthmatic. Similarly, an initial diagnosis of asthma may be made

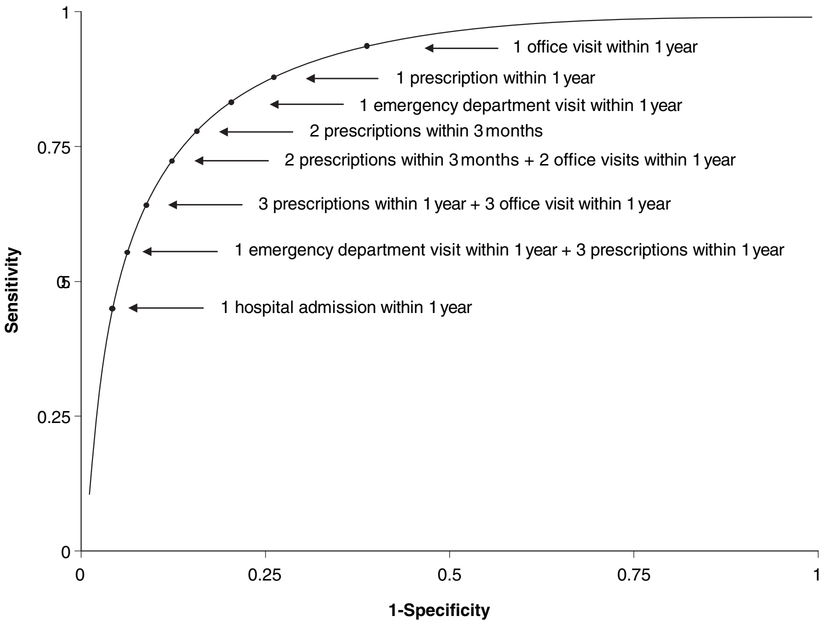

Figure 3 An ROC curve showing hypothetical levels of accuracy between various methods for identifying asthmatics. ROC, receiver operator characteristic.

during an emergency department (ED) visit for an individual presenting with obstructive airways, even though later it may be determined that the narrowed airways was the result of an allergy or an inhaled irritant. A diagnosis of asthma made during a hospital stay is most likely to be more accurate than the preceding two methods (pharmacy or ED). However, few asthmatics are ever hospitalized for asthma, and therefore relying on inpatient claims for identification may limit the number of persons found for programme participation. Given these circumstances, it is possible to perform an ROC analysis comparing the various identification modalities for efficacy.

Figure 3 presents a hypothetical ROC curve in which the various asthma identification criteria, or decision thresholds, are plotted. It can also be hypothesized that the gold-standard used for comparison was a positive pulmonary function test. Owing to the pairing of sensitivity and 1-specificity (or false positive fraction), there will always be a trade-off between the two parameters. As shown, the most lax decision threshold is that in which 1 office visit is all that is required to classify an individual as an asthmatic. While the ability to identify nearly all the true asthmatics in the population is high (sensitivity is approximately 95%), it comes at the price of a high false positive rate (approximately 70%). At the other extreme, using the strict decision threshold of 1 hospitalization within a year, elicits a lower sensitivity and an exceptionally low false positive fraction.

Using the results from this analysis, a DM programme can decide which criteria provide the best trade-off between sensitivity and specificity. This is a particularly important issue as there is a high short-term price to pay for ‘over-identifying’ potential patients (e.g. high false-positive rate) and a potentially high long-term price to pay for ‘under-identifying’ them. For example, based on the information in Fig. 3, if the programme chooses to use the decision threshold of ‘1 office visit within a year’ as the classifying criterion for asthma, up to 70% of people identified may be false positives. Most DM programmes perform an initial telephonic screening interview to weed out the false positives, so the cost of a high false-positive rate can be narrowed to the resources needed at this juncture. However, if the strict decision threshold of ‘1 hospitalization within 1 year’ was chosen for classifying an asthmatic, many true asthmatics would initially be missed, leading to the high costs of hospitalizations later down the line. As a practical matter, most researchers would choose the decision threshold point that lies on the ROC curve closest to the upper left-hand corner. This would provide the best compromise between a true positive and false positive classification. In Fig. 3 that point would coincide with the '2 prescriptions within 3 months' criterion.

There are some situations in DM where subjective judgement is necessary and thereby require some modification to the data needed to generate the ROC curve. For example, a congestive heart failure (CHF) programme may risk-stratify participants according to the New York Heart Association (NYHA) Function Classification System (The Criteria Committee of the New York Heart Association 1964), which places patients in one of four categories based on how much they are limited during physical activity (scoring is from I to IV, with better functional status denoted with a lower score). However, there maybe differences between nurses in how they would score a given patient. Congruence among raters may be high for classifying patients as levels I and IV, but might be poor for classifying patients in the middle range of II and III. Moreover, nurses may inadvertently introduce 'interviewer bias' into the scoring by posing questions to the patient in such a way as to elicit an inaccurate response ('you are feeling better today, aren't you Mr Jones?'). Similarly, nurses performing the interview telephonically may classify patients differently than nurses conducting inperson interviews.

These concerns for the accuracy of subjective judgement across the NYHA scale may be tested empirically vis-à-vis ROC analysis. As with any test of accuracy, a gold-standard must first be determined. For the NYHA classification, the true patient status may be established by expert agreement or by clinical indication. One such marker that has shown promise in recent studies is the brain natriuretic peptide (BNP). Plasma concentration levels have been documented to correlate highly with the NYHA categories and thus make this a useful clinical tool to assess disease severity (Redfield 2002; Maisel *et al*. 2003).

Each DM nurse then interviews a sampling of participants to ascertain their NYHA level. Comparisons are made between the nurse's assessment and the gold-standard determination. Sensitivity and 1-specificity is calculated for each of the four NYHA levels and plotted on an ROC curve. The resulting visual display should resemble that which was presented earlier in Fig. 3. A subsequent analysis can then be performed to determine which nurse had the highest overall accuracy. This analysis would be similar to that which was presented in Fig. 2, with each curve representing an individual nurse. The AUC for each curve would be determined and the largest AUC may be established as the 'best-practice.' The process just described can be an invaluable tool for organizations to measure inter-rater reliability and to ensure that programme participants are accurately and consistently stratified.

Discussion

In this paper the utility of ROC analysis was demonstrated as a means of assessing accuracy in the programmatic domains of any DM programme. There are several reasons why DM programme evaluators should consider using ROC analysis in lieu of the more conventional methods. First, it thoroughly investigates model or test accuracy across the entire range of scores. A predetermined cut-off point is not required because each possible decision threshold is calculated and incorporated into the analysis.

Second, unlike conventional 2 × 2 tables, ROC analysis allows for visual examination of scores on one curve or a comparison of two or more curves using a similar metric. This allows the analyst to easily determine which decision threshold is most preferred, based on the desired trade-off between sensitivity and specificity or between cost and benefit (Metz 1986), or to establish which model or test has the best accuracy based on the largest AUC.

Third, prevalence of the outcome in the sample population is not a limiting factor as it is with the conventional measures of accuracy. That said, it has been suggested that meaningful qualitative conclusions can be drawn from ROC experiments performed with as few as 100 observations – 50 for each group of positive and negatives (Metz 1978).

While the calculations for establishing the sensitivity and 1-specificity coordinates for individual decision thresholds are not especially complicated, an exhaustive iterative process is required to determine all points along the ROC continuum. As such, this procedure is better left to commercially available software packages that perform these functions for even large data-sets within seconds. In addition, typical outputs include AUC, tests of significance and confidence intervals.

Conclusions

Receiver operator characteristic analysis is an excellent tool for assessing diagnostic or predictive accuracy in several different areas of DM. Among other things, it can be used to determine (1) the most suitable data elements needed to properly identify an individual with the disease (2) which predictive model is most accurate in forecasting future costs, and (3) accuracy in risk-stratification and inter-relater reliability. There are many applications and advantages to using the ROC analysis in place of the more conventional approaches. Therefore its implementation as an evaluation tool should be strongly considered throughout the various phases of a DM programme.

References

Altman D.G. & Bland M. (1994) Diagnostic tests 2: predictive values. *British Medical Journal* **309**, 102.

American Diabetes Association (2000) Tests of glycemia in diabetes. *Diabetes Care* **23**, S80–S82.

Collision P. (1998) Of bombers, radiologists, and cardiologists: time to ROC. *Heart* **80**, 215–217.

Detection theory and applications. (1970) *Proc IEEE* **58**, 607–852.

Deurenberg P., Yap M. & van Staveren W.A. (1998) Body mass index and percent body fat: a meta analysis among different ethnic groups. *International Journal of Obesity and Related Metabolic Disorders* **22**, 1164–1171.

Disease Management Association of America [homepage on the Internet] Washington DC: Definition of Disease Management. DMAA (cited 2004 June 23) Available from: http://www.dmaa.org/definition.html

Gallagher D., Visser M., Sepulveda D., Pierson R.N., Harris T. & Heymsfield S.B. (1996) How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *American Journal of Epidemiology* **143**, 228–239.

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

Hanley J.A. & McNeil B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hannan E.L., Kilburn H., Jr, Lindsey M.L. & Lewis R. (1992) Clinical versus administrative data bases for CABG surgery. Does it matter? *Medical Care* **30**, 892–907.

Jollis J.G., Ancukiewicz M., Delong E.R., Pryor D.B., Muhlbaier L.H. & Mark D.B. (1993) Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Annals of Internal Medicine* **119**, 844–850.

Linden A., Adams J. & Roberts N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface* **7**, 38–45.

Linden A. & Roberts N. (2004) Disease management interventions: what's in the black box? *Disease Management* **7**, 275–291.

Lusted L.B. (1971) Decision making in patient management. *New England Journal of Medicine* **284** (8), 416–424.

MacDonald F.C. (1986) Quetelet index as indicator of obesity. *Lancet* **1**, 1043.

Maisel A.S., McCord J., Nowak R.M. *et al.* (2003) Bedside B-Type natriuretic peptide in the emergency diagnosis of heart failure with reduced or preserved ejection fraction. Results from the Breathing Not Properly Multinational Study. *Journal of the American College of Cardiology* **41** (11), 2018–2021.

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

Metz C.E. (1986) ROC methodology in radiologic imaging. *Investigational Radiology* **21**, 720–733.

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.

Mortensen H.B. & Volund A. (1988) Application of a biokinetic model for prediction and assessment of glycated haemoglobins in diabetic patients. *Scandinavian Journal of Clinical Laboratory Investigation* **48**, 595–602.

National Institutes of Health & National Heart, Lung and Blood Institute. (1998) Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults—the Evidence Report. National Institutes of Health. *Obesity Research* **6** (Suppl. 2), 51S–209S.

Redfield M.M. (2002) The Breathing Not Proper trial: enough evidence to change heart failure guidelines? *Journal of Cardiac Failure* **8** (3), 120–123.

Swets J.A. (1979) ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology* **14**, 109–121.

The Criteria Committee of the New York Heart Association (1964) Physical capacity with heart disease. In *Diseases of the Heart and Blood Vessels: nomenclature and Criteria for Diagnosis* (6 edn) by the Criteria Committee of the New York Heart Association (chairman Charles E. Kossmann), pp. 110–114. Little, Brown & Co, Boston.

Wang J., Thornton J.C., Russell M., Burastero S., Heymsfield S. & Pierson R.N. (1994) Asians have lower body mass index (BMI) but higher percent body fat than do whites: comparisons of anthropometric measurements. *American Journal of Clinical Nutrition* **60**, 23–28.

Weiner J.P. (2003) Predictive modeling and risk measurement: paradigms, potential and pitfalls. Symposium on 'Predictive Modeling': sponsored by the National Bluecross/BlueShield Association, Chicago. January 30.

Wellens R.I., Roche A.F., Khamis H.J., Jackson A.S., Pollock M.L. & Siervogel R.M. (1996) Relationships between the body mass index and body composition. *Obesity Research* **4**, 35–44.