Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 80 (2006) 24 – 38

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Receiver operating characteristics curves and related decision measures: A tutorial

Christopher D. Brown [a,*], Herbert T. Davis [b]

[a] *Ahura Corporation, 46 Jonspin Road, Wilmington, MA, 01887, United States*
[b] *223 Mission Ridge Corrales, NM, 87048, United States*

Received 5 February 2005; received in revised form 12 May 2005; accepted 20 May 2005
Available online 12 July 2005

Abstract

Chemometric and statistical tools for data reduction and analysis abound, but the end objective of most analytical undertakings is to make informed decisions based on the data. Decision theory provides some highly instructive and intuitive tools to bridge the gap between data and optimal decisions. This tutorial provides a user-centric introduction to receiver operator characteristic curves, and related measures such as predictive values, likelihood ratios, and cost curves. Important considerations for choosing between these tools are discussed, as well as the primary methods for determining confidence intervals on the various measures. Numerous worked examples illustrate the calculations, their interpretation and potential drawbacks.

Keywords: Receiver operator characteristic; ROC curve; Classification; Likelihood ratio; Decision theory; Bayesian; Cost; Limit of detection

1. Introduction

The mathematical and statistical methods of chemometrics are rather like sterile surgical instruments: they are often shiny, sometimes oddly shaped, and as a collection, can be very valuable in the gathering of information, but in the end the cutting, diagnosis and treatment is left to the experts. This is because existing chemometric tools are focused on efficiently distilling information, but the decisions made based on this information are highly domain-specific and context-dependent; they depend on the available choices, the quality of the information guiding the decision, and the possible outcomes. There are, however, several perspicuous techniques from decision theory that can guide the chemometric surgeon by quantifying and graphically representing decision-dependencies and probable outcomes. The methods reviewed in this tutorial, then, share at least this attribute — they are ultimately concerned with the context and consequences of decisions from data, rather than the data itself.

A great many chemical decisions pertain to dichotomous conditions. A protein marker is present or not, the structure of a molecule is X or Y, the reaction obeys first-order kinetics or second, one should terminate the reaction, or let it progress. The information available to aid in these decisions is rarely perfect in chemical problems, so no matter how assiduous the chemist, the ultimate result is she may be right, or she may be wrong.

Data themselves are impotent. It is not until data are interpreted against a set of rules that a decision is made, and action will be taken, usually under the assumption that the decision is correct. When the data are ambiguous decisions may well be erroneous, and the ramifications of this error define the 'loss' incurred. Decision theory is specifically concerned with how choices are, or should be made under conditions of uncertainty such that loss is minimized. Pure parameter estimation, in contrast, is detached from decision-based loss, as an erroneously estimated value may not actually result in an incorrect decision.

* Corresponding author. Tel.: +1 505 797 7106.
E-mail address: chrisbrown@chemist.com (C.D. Brown).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.05.004

How does one decide on a suitable rule for a decision process? In published accounts of dichotomous decisions (yes/no, stop/start) based on continuous variables (e.g., temperature, concentration, density, probability), one commonly sees the *classification rate* – the proportion of correct decisions to total decisions – as a measure of the goodness of the rule. There are two critical flaws with this goodness measure; namely, it considers all incorrect decisions to be equally hazardous, and treats all outcomes as equally likely. These assumptions are oftentimes inappropriate in practical applications, and even if they are, there are a number of auxiliary measures of decision performance that can enhance insight.

In this tutorial we have aspired to provide an overview of the practical aspects of some decision theory measures, including receiver operator characteristic (ROC) curves, area under the ROC curve, and related measures such as positive/negative predictive values, likelihood ratios, and cost function analysis. Where possible, the theoretical motivation of these measures is briefly discussed, but theoretical depth was inevitably sacrificed in favor of our primary goal, which was to provide the interpretation for the measures and critical instructions for practical use. There are several other excellent introductory references on this topic (many of them in the medical literature, e.g., [1–3]), and the monographs by Green and Swets [4], and Pepe [5] cover many aspects in greater depth and breadth. We have also aspired in the text to provide readable references for the topics as they are introduced herein.

2. Classification rate, correct positive and correct negative fractions

2.1. Introduction to the measures

The root of a dichotomous decision process is a threshold (t)-based rule on a continuous variable, y, that will drive the decision, D, as positive or negative according to

$$D = \begin{cases} + & \text{if} \quad y \geq t \\ - & \text{if} \quad y < t \end{cases} \tag{1}$$

For instance y might be a scalar instrumental measurement (e.g., pH, counts), a concentration estimate from a multivariate calibration model, or a probability from a logistic or discriminant analysis model. In the decision vernacular the 'positive' label is conventionally assigned to the decision that results in the most drastic action (e.g., critical process fault), but the decision could represent any dichotomy YES/NO, A/B, etc. Nevertheless, for clarity and simplicity of terminology we will continue to use the positive/negative demarcation. For convenience, we will also assume that large values of y are more indicative of positive decisions.

Assume that y does indeed have some ability to adequately discriminate between positive and negative events; the distributions cartooned in Fig. 1 illustrate one such scenario. There are an infinite number of possible decision thresholds, and we have labeled three possibilities in the figure. At threshold t_1, calling all events positive with $y \geq t_1$ would correctly identify nearly every positive event, although a large proportion of negative events would inappropriately be called positive. At threshold t_2 more of a balance is struck, as both positive and negative events are missed, and finally at t_3 most negative events are correctly identified, but a large proportion of the positive events are erroneously deemed negative as well.

At a candidate threshold, t_i, the outcomes of decision process ($D\pm$) over n trials can be evaluated against a reference ($R\pm$), which for the time being we take to be 'truth' (determined by hindsight, design, or other means). Four possible outcomes can result for each trial: the decision can be correctly positive (CP), correctly negative (CN), incorrectly positive (IP) or incorrectly negative (IN). (These are also often called true positive, true negative, false positive and false negative, respectively). A contingency table (sometimes called a confusion matrix) like that in Table 1 is often used to tabulate the outcomes. The cells CN, CP, IN, IP represent the number of trials that resulted in a particular outcome, so CN+CP+IN+IP=n. If the decision procedure was flawless, all n trials would be correctly categorized as positive or negative and IP=IN=0. More realistically, some trials will result in incorrect decisions IP or IN. The classification rate, CR, discussed briefly above, is simply

$$\mathrm{CR} = (\mathrm{CN} + \mathrm{CP})/n, \tag{2}$$

the number of correct trial outcomes out of the total number of trials. We will continue to use the term classification rate because it is so widely used in the literature, but as time is not involved some fields of research think it is more appropriately termed a fraction. The new terms that follow will therefore be referred to as fractions.

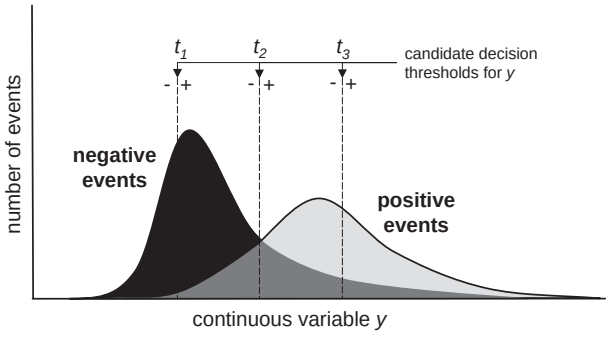

Fig. 1. An illustration of the occurrence of positive and negative events when ordered by the continuous variable y.

Table 1
A 2×2 contingency table

for threshold t		Reference		
		-	+	
Decision	-	CN	IN	
	+	IP	CP	
		CN+IP=n^-	IN+CP=n^+	n

A deeper assessment of the decision process is given by the calculation of *correct negative fraction* (CNF) and *correct positive fraction* (CPF):[1]

$$\mathrm{CNF} = \frac{\mathrm{CN}}{\mathrm{CN}+\mathrm{IP}} = \frac{\mathrm{CN}}{n^-} \tag{3}$$

$$\mathrm{CPF} = \frac{\mathrm{CP}}{\mathrm{CP}+\mathrm{IN}} = \frac{\mathrm{CP}}{n^+}. \tag{4}$$

Individually, the CNF and CPF values express the fraction of truly negative events that were correctly deemed negative, and the fraction of truly positive events that were correctly deemed positive. Consequently, CPF and CNF do not depend on the actual numbers of positive or negative events in the trials. The CR, in contrast, is dependent on these quantities, as some simple substitutions reveal.

$$\mathrm{CR} = p^+\mathrm{CPF} + p^-\mathrm{CNF}. \tag{5}$$

Here p^+ and p^- are the fractions of positive and negative events observed in the trials:

$$p^+ = n^+/n = (\mathrm{CP}+\mathrm{IN})/n$$

$$p^- = n^-/n = (\mathrm{CN}+\mathrm{IP})/n. \tag{6}$$

(And it is implied that $p^+=(1-p^-)$ and vice versa). Provided the positive and negative trials used to determine CNF and CPF were a representative sampling of the respective event distributions (that is, the trials didn't involve only unusually easy or unusually difficult cases), then CNF and CPF also have a probabilistic interpretation: CPF estimates the probability that the decision will be positive if the reference is truly positive, and CNF estimates the probability that the decision will be negative if the reference is truly negative.

$$E(\mathrm{CNF}) = \Pr(D = -|R = -) \tag{7}$$

$$E(\mathrm{CPF}) = \Pr(D = +|R = +). \tag{8}$$

Due to its dependence on the proportion of positive/negative events, the CR only has a probabilistic interpretation for the population if p^+ and p^- are representative of the population parameters $\pi^+=\Pr(\mathrm{R}=+)$and $\pi^-=\Pr(\mathrm{R}=-)$. If this holds, CR is an estimate of the probability that any given decision will be correct.

$$E(\mathrm{CR}) = \Pr(D = R). \tag{9}$$

This places a rather obvious limitation on classification rate. It is only meaningful as an estimator of the probability of a correct decision if the proportions of positive/negative events are reasonable, a contingency that often is not noted in literature reports of classification rates. Indeed in many focused studies the positive/negative proportions are intentionally skewed towards 0.5 so that a reasonable number of both positive and negative cases can be investigated. If the population parameters, π, are known from other sources (e.g., literature, other studies) they should be substituted for the sample estimates in Eq. (5).

Two scenarios shown in Fig. 2 illustrate calculations of Eqs. (2)–(6).

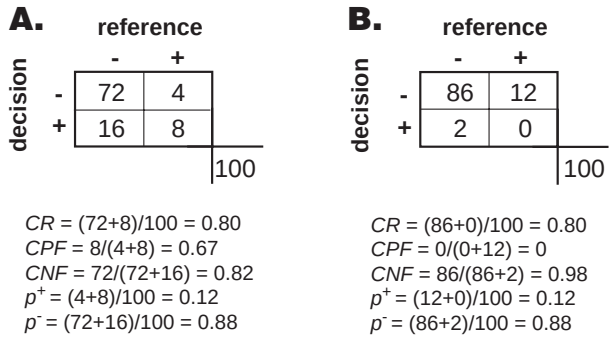

Fig. 2. Two scenarios illustrating the calculation of various decision measures.

2.2. Univariate confidence intervals

The CR, CPF, CNF, and p are all proportions, and hence all have binomial sampling distributions, and several standard sources such as Numerical Recipes [6] provide necessary details for exact interval calculation. Fig. 3 graphically presents exact confidence intervals for proportions of 0.75 and 0.5 at a variety of sample sizes. Otherwise,

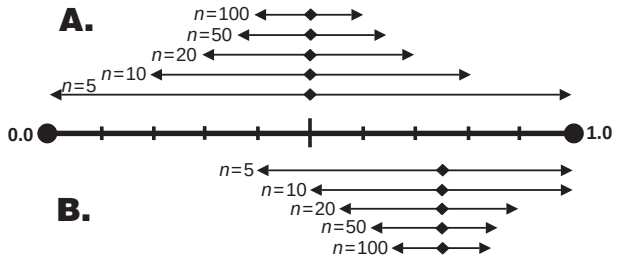

Fig. 3. Illustration of exact binomial 95% confidence intervals for true proportions A) 0.5, and B) 0.75 at a variety of sample sizes. The true proportion is indicated by the diamond, and the extent of the interval is conveyed by the double-headed arrow.

[1] Synonyms for CNF and CPF abound. CPF is variously called sensitivity or true positive rate (clinical use), hit rate and recall (signal detection theory, machine learning), while synonyms for CNF include specificity and true negative rate (clinical use), and (1-CNF) is often termed false-positive rate or false-alarm rate (signal detection theory, machine learning). In statistical hypothesis testing, CNF is the significance level $1-\alpha$, and CPF is the statistical power, $1-\beta$.

we provide two well-known, but approximate expressions for confidence intervals on proportions.

If the number of events is reasonably large, the binomial distribution can be approximated by the normal distribution, and the following large-sample formula for the upper (τ_U) and lower (τ_L) confidence bounds on an estimated proportion ($\hat{\tau}$) can be applied:

$$\tau_U, \tau_L = \hat{\tau} \pm \left[z_{\alpha/2} \sqrt{\hat{\tau}(1-\hat{\tau})/n} + 1/(2n) \right] \quad (10)$$

where $z_{\alpha/2}$ is the value of the normal deviate encompassing $1-\alpha$ of the normal curve (1.645, 1.960, 2.241, 2.576 for 90, 95, 97.5, and 99% confidence respectively), and n is the number in the denominator of the proportion. Snedecor and Cochran [7] advocate for the continuity correction term ($1/(2n)$), which generally makes the normal approximation more accurate. This formula can be extremely erratic and inaccurate if $n\tau(1-\tau)$ is small (less than 10).

The Wilson interval [8], though more daunting formulaically, has much better statistical properties, and is reasonable for any $n\hat{\tau}(1-\hat{\tau})$:

$$\tau_U, \tau_L = \frac{n}{n + z^2_{\alpha/2}} \cdot \left[\hat{\tau} + \frac{z^2_{\alpha/2}}{2n} \pm z^2_{\alpha/2} \sqrt{\frac{\hat{\tau}(1-\hat{\tau})}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \right]. \quad (11)$$

For the example in Fig. 2A, 95% confidence intervals using these two approaches are

	Point estimate	Eq. (10)	Eq. (11)
CR	0.80	[0.72 0.88]	[0.71 0.87]
CPF	0.67	[0.36 0.98]	[0.39 0.86]
CNF	0.82	[0.73 0.91]	[0.73 0.89]
p^+	0.12	[0.05 0.19]	[0.07 0.20]

The approach in Eq. (10) tends to provide overly wide interval estimates, and in some cases (e.g., CPF) dramatically so. We refer the reader to a recent and very comprehensive Ref. [9] for a detailed overview and discussion of the various binomial confidence interval estimation approximations.

2.3. Joint confidence intervals

As the accuracy of the decision is fully described the by the pair CPF and CNF, it is often more appropriate to report their joint confidence interval. As CPF and CNF are calculated from independent trials (positive and negatively respectively), they are statistically independent. Elliptical joint confidence regions could certainly be derived conditional on distributional assumptions, but these are excessively complex for the scope of this article. Instead, we report a simple method discussed in Pepe [5], and originally proposed by Hilgers [10] for the construction of distribution-free rectangular coverage regions.

Since each univariate interval has $(1-\alpha)$% coverage, the rectangular confidence region only covers $(1-\alpha)\times(1-\alpha)$%. For 95% univariate intervals, for example, the rectangular area would only provide 90% coverage. To define a rectangular region which truly covers $(1-\alpha)$%, therefore, the individual univariate confidence intervals are simply determined with $\sqrt{1-\alpha}$% confidence. Returning to the example in Fig. 2A, the rectangular confidence interval for the CPF, CNF pair suggests that with 95% confidence the CPF is between 0.36 and 0.88 and the CNF is between 0.71 and 0.89.

3. ROC curves, and area under the curve

3.1. Receiver operator characteristic curves

The receiver operating characteristic (ROC) curve was introduced in World War II military radar operations as a means to characterize the operators' ability to correctly identify friendly or hostile aircraft based on a radar signal. The loss incurred if a hostile aircraft is deemed friendly by mistake could be catastrophic, but at the same time military aircraft could not be sent to intercept an overwhelming number of benign vessels. The ROC curve was devised as a graphical means to explore the trade-offs between these competing losses at various decision thresholds when a particular quantitative variable, y, is used to guide the decision.

The simplicity and usefulness of the ROC approach was recognized shortly thereafter for signal detection studies in psychophysics [11] and was a major focus of the 1964 monograph by Green and Swets [4] and Egan's text of 1975 [12]. Swets and Pickett's 1982 publication on the evaluation of diagnostic systems [13] precipitated a flood of applications and theoretical advances in medical decision theory that has not yet subsided. The method seems to have slowly diffused into clinical chemistry applications [14–16] and somewhat surprisingly, only recently the machine learning [17] community. The ROC curve, by all indications, remains essentially unused in non-clinical analytical and chemometric applications. Whether the under-utilization is due to a lack of familiarity, lack of readily available software, or both is presently unclear.

The CPF and CNF we discussed above are characteristics of the threshold used to make the decision, as well as the intrinsic confusion in the data itself at that threshold. Clearly a different threshold could be chosen, and CPF and CNF can again be determined. Over many candidate thresholds, a table of CPF's and CNF's is assembled

t_1	t_2	t_3	...	t_k
CPF_1	CPF_2	CPF_3	...	CPF_k
CNF_1	CNF_2	CNF_3	...	CNF_k

The classic ROC curve is generated by plotting the CPF_i on the vertical axis, and $1-CNF_i$ on the horizontal axis, leading to a summary graph like that shown in Fig. 4.

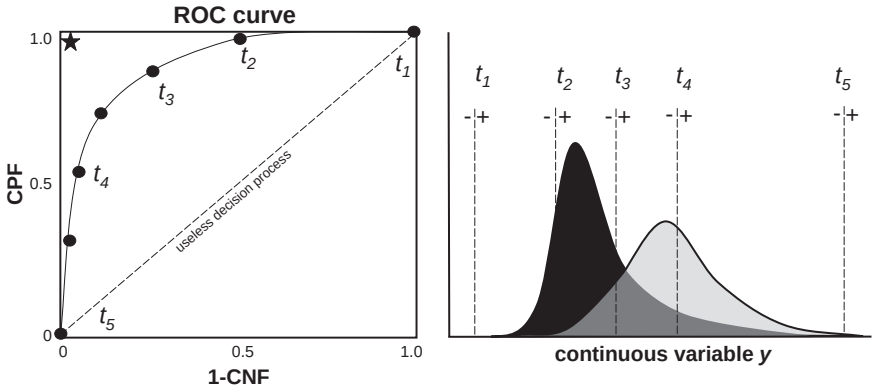

Fig. 4. Generation of the ROC curve by evaluating the CPF and CNF at various decision thresholds on y.

An ideal decision variable would have a ROC curve that passed through CPF = 1 and CNF = 1, which would correspond to a point in the top left corner of the ROC axis (indicated by a star in the figure). The top right (t_1: CPF = 1, CNF = 0), and bottom left vertices (t_5: CPF = 0, CNF = 1) represent extreme decision thresholds under which every trial is unilaterally deemed positive, or negative; that is, there is really no y-based-decision to speak of. If the table above is ordered by threshold value (either in increasing or decreasing order) then CPF and 1-CNF will both be monotonic increasing or monotonic decreasing regardless of the shape of the marginal distributions of events. Therefore the derivative at any point on the ROC curve must be greater than or equal to zero, which highly restricts the form the ROC curve can take.

The ROC is also invariant to strictly increasing transformations of y, which makes it a convenient representation insensitive to scale. For example, since a positive decision implies $y \geq t$, CPF can also be written as

$$E(\mathrm{CPF}) = \Pr(y \geq t | R = +). \tag{12}$$

This statement still holds true under any monotonically increasing function f:

$$E(\mathrm{CPF}) = \Pr(f(y) \geq f(t) | R = +). \tag{13}$$

It also results in another advantage to be discussed further below.

A dotted line is also drawn through the diagonal of the ROC axes in Fig. 4, traversing between the unilateral decision thresholds at t_1 and t_5. This is the line of chance. Decision processes that result in a CPF/CNF pair that falls on this line are no better than a proverbial coin-flip. To illustrate, a decision process with CPF = 0.2 and 1 − CNF = 0.2 implies that out of 100 positive events we will correctly detect 20, but also incorrectly call 20 out of 100 negative events positive, so a positive decision is no more suggestive of a truly positive event than a truly negative event.

The ROC curve, then, provides a very simple graphical view of the trade-space that is possible for a given decision variable, and the general discriminating power of the variable for detecting positive/negative events independent of the event rates. The classification rate at a certain decision threshold, by contrast, can often appear exceptionally good or exceptionally bad, but if viewed as a CPF, CNF pair on the ROC curve it will be immediately apparent if it corresponds to chance detection, or is capable of satisfying an objective. The ROC curve is also free of parametric assumptions. The marginal distributions of y for the positive and negative events can take any form—symmetric/asymmetric, unimodal/multimodal—and the ROC curve obediently follows as a simple descriptor of the decision space.

In applications where detecting positive events with a specified probability is critical—such as screening processes which are intended to trigger follow-on investigations—the analyst can very rapidly examine the ROC curve for a decision process and assess the minimum incorrect positive fraction that will satisfy the required CPF threshold. Frequently the exact choice of where to operate the decision process depends on secondary factors, such as the cost or risk associated with missing positive events or incorrectly identifying negative events as positive, a matter better suited for the application of cost functions, which are discussed in Section 5.1.

Although we have used the term ROC curve above, the tabulation of CPF's and CNF's merely provides a series of points on the ROC axes. More points can be generated by evaluating the CPF and CNF at more finely sampled decision thresholds, but if the number of positive or negative events is small in a data set under study, there will only be a small number of 'allowable' CPF or CNF's. For example, with only 5 positive events in a study, the points on the y-axis of the ROC curve will be limited to discrete values of 0/5, 1/5, 2/5, 3/5, 4/5, 5/5. A smooth approximation to these points is sometimes desirable.

The literature reflects a host of both formal (parametric) and informal (non-parametric) approaches for fitting the ROC curve. It is particularly challenging to fit an ROC curve with a one-size-fits-all function because the marginal distributions of the positive and negative events on y dictate the shape of the ROC curve, and hence the appropriate

model. In some instances the y distributions of the positive and negative events are near-normal, or normalizable by monotonic transformation (e.g., logarithm) and the so-called binormal ROC model [5] can be used. As noted above, the ROC is invariant to monotonic transformation, so the choice of normalizing transformation will not affect the shape of the ROC curve. Although the binormal fitting method is not particularly complicated, it can appear somewhat obtuse. In short, the critical parameters for fitting the binormal model are the (optionally transformed) distribution means (μ^+,μ^-) and standard deviations (σ^+,σ^-) or estimators thereof, and the fitted ROC curve as a function of threshold value t is obtained from manipulations of the standard normal cumulative distribution using these parameters.

$$\mathrm{ROC}(t) = \Phi\left(a + b\Phi^{-1}(t)\right). \tag{14}$$

Φ is the standard normal cumulative distribution function, and a and b are defined as

$$a = \frac{\hat{\mu}^+ - \hat{\mu}^-}{\hat{\sigma}^+} \quad b = \frac{\hat{\sigma}^-}{\hat{\sigma}^+}. \tag{15}$$

We provide a short MATLAB (The Mathworks, Natick, MA) program in the appendix for estimating the binormal ROC curve by this method, and two examples of fitting using this approach are given in Fig. 5.

The reader should proceed cautiously with parametric ROC curve fitting methods, for while the points on the ROC curve are free of distributional assumptions, fitting methods can be quite reliant on distributional assumptions. In general, smooth approximations to the measured CPF, CNF pairs are of aesthetic or interpretational value only, so there is often little impetus to expend much effort in the venture.

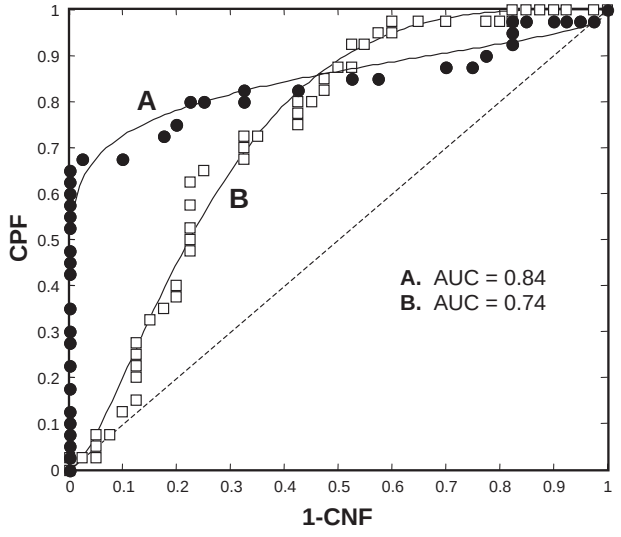

Fig. 5. Two examples of a smooth fit to the ROC curve estimated using the binormal model. AUC is the area under the fitted curve, discussed later in Section 3.2.

3.2. The area under the ROC curve

A single point on the ROC curve represents the characteristics of a decision process at a fixed threshold, and the ROC curve defines the operating characteristics that are achievable based on y, but what can be said about the discriminating power of the continuous variable in general? The area under the ROC curve (AUC) is frequently used as just such a summary measure. The perfect ROC curve, which traverses the point CPF=CNF=1, has an AUC of 1, while a test with an ROC curve along the line of 'uselessness' has an AUC of 0.5. Hanley and McNeil [18] showed that, although the AUC seems like a crude summary measure, it actually has a probabilistic interpretation. It is equal to the probability that y values for a randomly selected pair of positive and negative events will be correctly ordered, which is rather fantastically termed the probability of stochastic domination in non-parametric statistics. The AUC, then, is the probability of stochastic domination of the y-values for positive events over negative events, and due to the condensation of information, the reader should note that very differently shaped ROC curves can have the same AUC. Pepe [5] prefers a more general interpretation of the AUC as the average CPF over the entire range of possible CNF's.

An interesting parallel meaning for the AUC arises if we, for the time being, ignore the ROC curve entirely, and think just about the distributions of y for positive and negative events. At the most elementary level, one might want to know whether the distributions of positive and negative events had means that were statistically distinguishable. If the distributions were normal, a two-sample t-test would be a logical starting point to answer this question. However, if the normal assumption is tenuous non-parametric methods are preferred, and the non-parametric analog to the t-test is the Wilcoxon rank-sum test, or synonymously the Mann–Whitney U test, [7] which is precisely a non-parametric test for stochastic domination.

Assume the following y-values were recorded for a series of trials with known reference:

y values (R=-)	0.213	0.153	1.21	−0.110	0.001	0.524	0.847	−0.046	
y values (R=+)	1.42	0.775	0.966	0.412	1.22	0.856	0.210	0.735	1.18

To perform the U test we first order the y values jointly from lowest to highest:

1	2	3	4	5	6	7	8	
−0.110	−0.046	0.001	0.153	0.210	0.213	0.412	0.524	...
(−)	(−)	(−)	(−)	(+)	(−)	(+)	(−)	
9	10	11	12	13	14	15	16	17
0.735	0.775	0.847	0.856	0.966	1.18	1.21	1.22	1.42
(+)	(+)	(−)	(+)	(+)	(+)	(−)	(+)	(+)

and sum the ranks assigned to the values in each group:

$$W^- = (-)\,\text{rank sum} : 1+2+3+4+6+8+11+15 = 50$$

$$W^+ = (+)\,\text{rank sum} : 5+7+9+10+12+13+14+16+17 = 103.$$

The Mann–Whitney U statistic is calculated from the group with the smallest rank-sum, in this case the negative group:

$$U = n^+ \cdot n^- + \frac{n^- \cdot (n^- + 1)}{2} - W_- = 9 \cdot 8 + \frac{8 \cdot (8+1)}{2} - 50 = 58 \quad (16)$$

where n^x is the number of positive or negative events. What is the U statistic telling us? Had all y values been perfectly ordered in correspondence with negative and positive events, the negative rank-sum would have been 36, and U would have been 72. This would also, as discussed above, correspond to a perfect ROC curve with an AUC of 1. In fact, if the Mann–Whitney U is normalized to the maximum U that can be observed under 'perfect ranking', it is exactly the AUC as first noted by Bamber [19]:

$$\text{AUC} = \frac{U}{n^+ n^-}. \quad (17)$$

Alternatively, if the data conforms to the binormal model of Eq. (14), the AUC is estimable directly as

$$\text{AUC} = \Phi\left(\frac{\hat{\mu}_D - \hat{\mu}_{\bar{D}}}{\sqrt{\hat{\sigma}_D^2 + \hat{\sigma}_{\bar{D}}^2}}\right) \quad (18)$$

This method of approximating the AUC is also included in the aforementioned MATLAB program in the appendix, and the AUC's for the binormal estimated ROC's in Fig. 5 are inset in the figure.

Simple trapezoidal integration under the CPF, CNF pairs is also an option. The trapezoid rule uniformly underestimates the true definite integral, but with reasonably small segment widths it is often quite sufficient, and it is certainly very easy to implement. As we proceed to confidence interval estimation, however, it will be evident that the rank-sum approach offers some convenient advantages.

3.3. Confidence intervals

3.3.1. Confidence intervals on a fitted ROC curve

As one might have guessed from the expressed difficulties of ROC curve fitting, confidence intervals on such curves are even less definable. In general the practitioner is faced with a smorgasbord of literature approximations too varied to enumerate here. In linear regression the confidence interval on the regression line is much narrower than the standard error on the individual points used to construct it (assuming the number of observations is well in excess of the number of parameters); accordingly, the confidence interval on the ROC curve must be narrower than the interval estimates on the individual CPF, CNF pairs. One can simply plot the points on the ROC curve with their associated joint confidence intervals (as we have done in an example in Section 6.1 below) and take this as a bounding condition. Interval expressions do exist for ROC curves generated from the binormal model, with some further complicating assumptions, but we refer the reader to more complete discussions of this approach in Ref. [5].

Beyond this, the only generally applicable course of action is to empirically determine the confidence intervals on the ROC curve by bootstrap [20] resampling, although it is not even transparent what should be bounded by the bootstrap samples, since the ROC lacks a general parametric form. Some researchers have merely presented bounds within which $1-\alpha\%$ of the bootstrapped ROC curves fell. Due to described difficulties in ROC curve fitting, however, it is quite rare to see confidence intervals presented for a fitted ROC curve for anything but a binormal scenario. It is much more typical to see the uncertainty in the ROC curve cast in terms of the uncertainty in the AUC, which we discuss next.

3.3.2. Confidence intervals on the AUC

The correspondence between the AUC and the Mann–Whitney U statistic is a tremendous advantage for confidence interval estimation. Hanley and McNeil [18] suggested that the standard error of the AUC can be conservatively approximated as

$$\sigma^2_{\text{AUC}} \cong \frac{\text{AUC}(1-\text{AUC}) + (n^+ - 1)\left(q_1 - \text{AUC}^2\right) + (n^- - 1)\left(q_2 - \text{AUC}^2\right)}{n^+ n^-} \quad (19)$$

where

$$q_1 = \text{AUC}/(2-\text{AUC})$$

$$q_2 = 2\text{AUC}^2/(1+\text{AUC}). \quad (20)$$

Therefore, under the normal assumption, the $1-\alpha$ confidence interval for the AUC is

$$\text{AUC}_L, \text{AUC}_U = \text{AUC} \pm z_{\alpha/2}\sqrt{\sigma^2_{\text{AUC}}} \quad (21)$$

where z is the typical standard normal deviate, and σ_{AUC}^2 is determined from Eq. (19).

To reiterate, the Hanley/McNeil method is conservative, providing confidence intervals that are excessively wide. If the binormal model has been used to estimate the AUC, expressions for its standard error are again available [5]. Bootstrap resampling remains an option for any of these quantities.

4. Predictive measures: positive and negative predictive values; likelihood ratios

4.1. Positive and negative predictive values

To this point the discussion has focused on methods derived from correct positive fractions, and correct negative fractions, which are estimators of probabilities of decisions conditioned on events: if a positive event has occurred, what is the probability that I will make the correct (positive) decision? If a negative event has occurred, what is the probability that I will make the correct (negative) decision? Although these probabilities are informative, they are often not the probabilities that are truly of interest to the investigator. In practice, one only has the data, and not the truth, and the more relevant questions are: if I've made a positive decision, what is the probability that the event is truly positive? If I've made a negative decision, what is the probability that the event is truly negative? In short, they are questions of trepidation—I'm about to make a certain decision, how likely am I to be right/wrong? Probabilities of this form are often termed positive and negative predictive values in the literature, and are contrasted to CPF and CNF below:

The derivation of the expressions for PPV and NPV are

$\Pr(D\|R)$	$\Pr(R\|D)$
CPF : $\Pr(D=+\|R=+)$	PPV : $\Pr(R=+\|D=+)$
CNF : $\Pr(D=-\|R=-)$	NPV : $\Pr(R=-\|D=-)$

straightforward applications of Bayes theorem to the CPF and CNF using the prior probability of positive and negative events (p^+, p^-), which results in the formulae

$$\mathrm{PPV}=\frac{p^+\mathrm{CPF}}{p^+\mathrm{CPF}+p^-(1-\mathrm{CNF})} \tag{22}$$

$$\mathrm{NPV}=\frac{p^-\mathrm{CNF}}{p^-\mathrm{CNF}+p^+(1-\mathrm{CPF})}. \tag{23}$$

Rearrangement provides alternate expressions in terms of the quantities in the contingency Table 1:

$$\mathrm{PPV}=\frac{\mathrm{CP}}{\mathrm{CP}+\mathrm{IP}} \tag{24}$$

$$\mathrm{NPV}=\frac{\mathrm{CN}}{\mathrm{CN}+\mathrm{IN}}. \tag{25}$$

Naturally the PPV and NPV also provide the probabilities of erroneous decisions, since $1-\mathrm{PPV}$ is the probability that a positive decision will be incorrect, and $1-\mathrm{NPV}$ is the probability that a negative decision will be incorrect.

It is critical for the reader to recognize that the PPV and NPV depend on the proportion of positive and negative events, while the CPF and CNF do not, which is both their advantage, and detriment. To illustrate the advantage, assume a decision process yields a CPF=CNF=0.75, a characteristic which does not depend on the proportion of positive/negative events. The PPV and NPV at various different proportions of positive events are

p^+	0.05	0.20	0.35	0.50	0.65	0.80	0.95
PPV	0.136	0.429	0.618	0.750	0.848	0.923	0.983
NPV	0.983	0.923	0.848	0.750	0.618	0.429	0.136

Even if the CPF of a decision process is 0.75, if positive events occur relatively rarely (1 in 20, $p^+=0.05$) then a positive *decision* is really only correct 13.6% of the time. Interpreted differently, a positive decision means there is a 13.6% chance that a positive event has really occurred (assuming $p^+=0.05$). A negative decision, however, implies that a negative event is extremely likely (NPV=98.3%). The PPV and NPV, therefore, reflect the intrinsic power of a positive and negative decision—a negative decision, in this case, is extremely emphatic, while a positive decision should obviously be regarded with some skepticism. Consequently, the PPV and NPV are often much more useful to the end-user of a decision process than CPF and CNF. For instance, in medical screening/diagnosis the PPV and NPV are the critical parameters to the clinician.

To the detriment of the PPV and NPV, like the classification rate, the dependence of the PPV and NPV on positive/negative event rates makes them rather meaningless study-specific values unless the event rates p^+, p^- are very similar to those expected in real-world deployment of the decision process. As noted above, focus-studies, for example, are often designed such that a sufficient number of both positive and negative events can be examined in a manageably small number of trials, so the event rates tend to be balanced near 0.5; Mother Nature is rarely so even-handed. However, exploiting the fact that CPF and CNF are independent of event-rates, Eqs. (22) and (23) can still be used to estimate the actual PPV and NPV in various general event-rate situations. All that is required is substitution of p^+/p^- with the event-rates of interest, for example, literature, expert or population values. This procedure is sometimes used in case-control clinical studies where the study disease rate is greatly elevated relative to the prevalence in the population at large. The CPF and CNF as estimated from the case-control study can be used with population statistics to estimate the PPV and NPV of the procedure if it were to be broadly deployed.

Therefore, the utility of the PPV and NPV depends heavily on the study and application context. When the prior probabilities of positive and negative events are known to be accurate, the PPV and NPV are much more relevant for prediction than the CPF and CNF. If the event rates are unknown, or cannot be defined—for example, it is difficult to estimate the prior probability of a terrorist strike, or the

event rates in a previously unstudied population—they have very limited utility.

Although not an ROC curve in the literature vernacular, PPV can be plotted against (1-NPV) at various decision thresholds, which, provided the event-rates are appropriate, can be a more informative prediction view of the decision space than the ROC curve alone can provide.

4.2. Likelihood ratios (Bayes factors)

Over the past two decades, likelihood ratios (synonymous with Bayes factors) have gained in popularity in the medical literature for describing the value of a clinical decision process [21,22]. A likelihood ratio is simply the ratio of probabilities of a specific occurrence under different hypotheses. For a dichotomous decision process the alternative hypotheses are that a positive event has truly occurred, or a negative event has truly occurred. The likelihood ratio for a positive decision is therefore

$$\mathrm{LR}^{+} = \frac{\Pr(D = +|R = +)}{\Pr(D = +|R = -)} = \frac{\mathrm{CPF}}{1 - \mathrm{CNF}} \tag{26}$$

and similarly, the likelihood ratio for a negative decision is

$$\mathrm{LR}^{-} = \frac{\Pr(D = -|R = +)}{\Pr(D = -|R = -)} = \frac{1 - \mathrm{CPF}}{\mathrm{CNF}}. \tag{27}$$

The scale for likelihood ratios is $(0,\infty)$, and a $\mathrm{LR}=1$ indicates that the decision is powerless. A perfect decision process would have $\mathrm{CPF}=1$ and $\mathrm{CNF}=1$, and hence $\mathrm{LR}^{+}=\infty$ and $\mathrm{LR}^{-}=0$.

The increasing attention paid to likelihood ratios is mostly due to their simple interpretation: if a decision is positive, a positive event is LR^{+} times more likely than a negative event. Therefore, LR's of 1 indicate uninformative decision processes ($\mathrm{CPF}=1-\mathrm{CNF}$, which is always a point on the line of uselessness in the ROC plot). Mathematically, LR's apply to odds (O), rather than probabilities (Pr), but the two are easily inter-converted as

$$O = \frac{\Pr}{1 - \Pr} \quad \Pr = \frac{O}{1 + O}. \tag{28}$$

The pre-decision odds of a positive event is

$$O^{+}_{\mathrm{pre}} = \frac{p^{+}}{1 - p^{+}}. \tag{29}$$

After the data, y, arrives a positive or negative decision will be made according to the decision rule. If this decision rule has characteristics LR^{+}, the post-decision odds of a positive event given a positive decision is

$$O^{+}_{\mathrm{post}} = \mathrm{LR}^{+} \cdot O^{+}_{\mathrm{pre}}. \tag{30}$$

Another advantage of likelihood ratios is that they do not depend on the event-rates, and yet they are very similar to the PPV and NPV in terms of providing evidence for prediction.

As an example, consider a test with $\mathrm{CPF}=0.60$ and $\mathrm{CNF}=0.95$. The positive likelihood ratio is 12.00, which means that, given a positive decision, the true event is 12.00 times more likely to be positive than negative. If the prior odds of a positive event were 0.25 ($p^{+}=0.2$), the posterior odds of a positive event given a positive decision are 3, which is a posterior probability of 0.75 (3/1+3). Note that if instead the prior odds of a positive event were 2 ($p^{+}=0.667$), the posterior odds given a positive decision would now be 24 with the same positive likelihood ratio of 12.00.

As an interesting aside, if the distributional data follow a binormal model, the LR^{+} can be interpreted as the slope of the line tangential to the ROC curve at a given CPF–CNF pair.

4.3. Confidence intervals

PPV and NPV are proportions, and therefore their confidence intervals can be determined exactly from the binomial distribution, or approximately using Eqs. (10) or (11). Likelihood ratios, being the ratio of two proportions, are not so simple. The most common approach is to work with the logarithm of the likelihood ratio, and assume asymptotic normality [23]. The procedure is as follows. First estimate the variance of the log-likelihood ratios as

$$\sigma^{2}_{\ln(LR^{+})} = \frac{1 - \mathrm{CPF}}{\mathrm{CP}} + \frac{\mathrm{CNF}}{\mathrm{IP}} \tag{31}$$

$$\sigma^{2}_{\ln(LR^{-})} = \frac{1 - \mathrm{CNF}}{\mathrm{CN}} + \frac{\mathrm{CPF}}{\mathrm{IN}}. \tag{32}$$

(Quantities in the denominators are again those from Table 1). The confidence interval for the likelihood ratio is then given by

$$\left[\mathrm{LR}\cdot\exp\left(-z_{\alpha/2}\sqrt{\sigma^{2}_{\ln(LR)}}\right) \quad \mathrm{LR}\cdot\exp\left(+z_{\alpha/2}\sqrt{\sigma^{2}_{\ln(LR)}}\right)\right]. \tag{33}$$

(The same expression holds for both LR^{+} and LR^{-}, with the appropriate replacement of terms).

5. Other considerations

5.1. Cost functions

In the introduction we discussed the notion that different decision errors, i.e., incorrect negatives, incorrect positives, could result in very different losses. The ROC curve and related measures do not however quantify the losses; they merely characterize the probability of the various potential errors of the decision process. In many instances the loss incurred can be reduced to monetary terms, and the most instructive characterization of the available trade-offs in a decision operating space is achieved by considering monetary characteristics of decision processes. Such cost functions allow one to assess the decision space relative to *expected cost*—the average expense per trial—or *expected*

margin—the expected loss or savings per trial when a decision process is used, relative to the expected cost if no decision process is used.

The expected cost function depends on the rate of positive/negative events (p^+) and as many as 5 cost terms:

Ξ	cost of making the decision
Ξ_{CP}	cost of a correct positive decision
Ξ_{CN}	cost of a correct negative decision
Ξ_{INF}	cost of an incorrect negative decision
Ξ_{IP}	cost of an incorrect positive decision

Normally these costs expressed on a per trial basis. The expected cost of a decision process operating with characteristics CPF and CNF is

$$E(\text{Cost})_{\text{decision}} = \Xi + [\Xi_{CP}\text{CPF}\cdot p^+ + \Xi_{CN}\text{CNF}\cdot p^-] + [\Xi_{IN}(1-\text{CPF})\cdot p^+ + \Xi_{IP}(1-\text{CNF})\cdot p^-] \quad (34)$$

The square brackets above simply collect cost terms that result from correct versus incorrect decisions. If the decision process was not employed at all ($\Xi=0$) and all events were assumed negative (CNF=1, CPF=0), the expected cost would simply be

$$E(\text{Cost})_{\text{no decision}} = \Xi_{IN}\cdot p^+ + \Xi_{CN}\cdot p^-. \quad (35)$$

The expected margin is

$$M = \left(\frac{E(\text{Cost})_{\text{no decision}}}{E(\text{Cost})_{\text{decision}}} - 1\right)\cdot 100\% \quad (36)$$

where positive margins indicate that the decision is providing a cost-benefit, and negative margins indicate that the decision is actually losing money.

Oftentimes in cost function analysis the terms associated with correct decisions are ignored (the cost is assumed to be zero), which in effect presumes that the cost of the decision (Ξ) and the costs of the correct decisions (Ξ_{CP}, Ξ_{CN}) have no role in defining the operating characteristics of the decision process. This is certainly true for some circumstances, but not all. For example, few automobile owners would invest in a highly accurate oil-dipstick that costs $500 per use, as any purported improvement in accuracy would likely not justify the cost. Additionally, our presentation here suggests that the costs of certain outcomes are fixed regardless of CPF, and CNF. In practice this may not be the case, as overhead and infrastructural costs may increase or decrease depending on the number of positive/negative decisions being made.

An example of cost function analysis is presented in Fig. 6, and the ROC curve for the candidate decision process is indicated by the black squares in Fig. 6A. Assume the question regards the implementation of a rapid spectroscopic analysis method that could be run on each lot of manufactured material. It has been estimated that the amortized cost of the QC method per test will be $10 ($\Xi$). If the method correctly identifies faulty batches, the added cost in downtime is $15 ($\Xi_{CP}$), while correct negatives (no fault) incur no added cost ($\Xi_{CN}=0$). If batches are incorrectly deemed faulty, the incremental cost of identifying the false-positive is $50 ($\Xi_{IP}$). Critically, if faulty batches are deemed normal, the flaw is often not caught until the customer has received shipment, and the cost in this situation is estimated to be $250 ($\Xi_{IN}$). As it is a rather finicky manufacturing process, the expected abundance of faulty batches is quite high: about 0.2 (p^+). The question is, is it worth implementing this spectroscopic method? If so, at which decision characteristics?

In Fig. 6B we have plotted the classification rate as a function of CPF, which suggests the maximum classification rate is achieved around CPF=0.43, although one could apparently operate anywhere between CPF=0 and CPF=0.7 and achieve roughly a CR of 0.8 owing to the relative infrequent occurrence of faulty batches. Fig. 6C presents quite a different view of the decision space, where we have plotted expected margin (%) versus the same CPF axis for direct comparison to Fig. 6B. The same information could be discerned from the isomargin contours we have overlaid

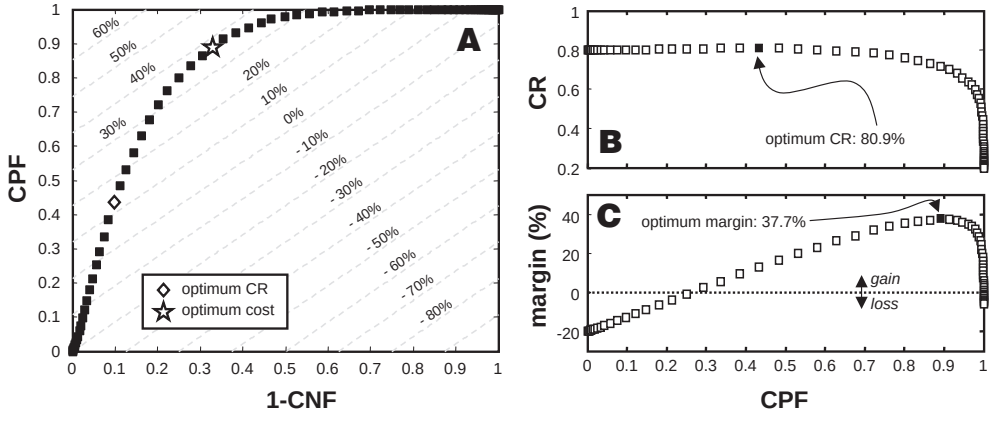

Fig. 6. Cost analysis of candidate decision processes. A. The ROC curve for the method, overlaid on iso-margin contours. B. The classification rate of the method versus CPF, showing a maximum at CPF=0.44, and C. The margin versus CPF, suggesting an optimum at CPF=0.89.

on the ROC curve, but this is a somewhat simpler way to view it. Fig. 6C indicates that it is possible to achieve an estimated 37.7% cost savings by using the spectroscopic method, and that the decision process should operate at a CPF of approximately 0.9 (CNF of about 0.68) to achieve optimal savings. If one had chosen the maximum CR as the threshold at which to operate the decision process, it would operate at a margin of only 11%. Worse still, it might be (incorrectly) inferred that since the optimum CR=0.8, and the proportion of faulty batches is about 0.2, the decision process is barely improving on chance detection, and therefore can provide no cost-benefit whatsoever.

Cost curves, as we have portrayed them in the example above, are most relevant in large-scale applications of decision processes where the critical objective is to reduce cost. They are, however, equally relevant many other applications in which the various outcomes are viewed to have very different consequences. For example, comparative risk values can be substituted for the monetary cost terms in Eq. (34) to stress, for example, the criticality of positive event detection for chemical/biological weapons or protein markers for immediately life-threatening disease. Conversely, comparative risk can be chosen to reflect the severe repercussions of incorrect positive results in forensic applications.

5.2. Imperfect references

It is sometimes difficult to perfectly define whether an event has or has not actually occurred. This situation often presents itself because of cost limitations, time constraints, or other practical limitations on experimentation. For example, in medical decision making, new screening/diagnostic procedures are often compared to the existing 'gold-standard' methodology, which is often an imperfect 'silver-standard' rather than gold. For example, the current standard for the diagnosis of diabetes is the oral glucose tolerance test, which has an estimated ROC AUC of only 0.82. The perfect reference would be longitudinal outcome (e.g., development of diabetic retinopathy, neuropathy, etc.) but in the early phases of evaluating a new methodology longitudinal studies, which often take many years to execute, are prohibitively expensive. Therefore, it is a practical reality that some decision processes must be evaluated against a reference that may be less than ideal. Nevertheless, the objective of the analyst remains the same: assess the true performance characteristics of candidate decision process. Fortunately, methods exist for such objective determinations even if the reference is imperfect. If the characteristics (CPF, CNF) of the reference method are known, and the reference and candidate decision processes are independent, a solution is tractable and exact [24–26]. We give a synopsis of this solution below. Beyond this scenario, latent class analysis methods have found increasing favor [27,28]. Discrepant resolution is also commonly applied, but this approach produces biased estimates [29] of CPF and CNF and is increasingly recommended against [30,31].

Assume that a number of trials have been run, and a contingency table has been generated for the decision versus the (imperfect) reference condition at a particular decision threshold t. The *observed* CNF, CPF and p^+ can be calculated directly from the contingency table (quantities which we denote by the subscript *obs*). If the reference method has known $\mathrm{CPF}_{\mathrm{ref}}$ and $\mathrm{CNF}_{\mathrm{ref}}$ characteristics, the following chain of formulae afford unbiased estimates of the true decision $\mathrm{CPF}_{\mathrm{true}}$ and $\mathrm{CNF}_{\mathrm{true}}$, based on the observed $\mathrm{CPF}_{\mathrm{obs}}$ and $\mathrm{CNF}_{\mathrm{obs}}$, and p^+_{obs}. First, the true fraction of positive events must be estimated:

$$\hat{p}^+ = \frac{p^+_{\mathrm{obs}} - 1 + \mathrm{CNF}_{\mathrm{ref}}}{\mathrm{CPF}_{\mathrm{ref}} - 1 + \mathrm{CNF}_{\mathrm{ref}}} \tag{37}$$

With this information, determine the following four quantities A–D:

$$A = \frac{\hat{p}^+ \mathrm{CPF}_{\mathrm{ref}}}{\hat{p}^+ \mathrm{CPF}_{\mathrm{ref}} + (1-\hat{p}^+)(1-\mathrm{CNF}_{\mathrm{ref}})}$$

$$B = \frac{(1-\hat{p}^+)(1-\mathrm{CNF}_{\mathrm{ref}})}{\hat{p}^+ \mathrm{CPF}_{\mathrm{ref}} + (1-\hat{p}^+)(1-\mathrm{CNF}_{\mathrm{ref}})}$$

$$C = \frac{\hat{p}^+(1-\mathrm{CPF}_{\mathrm{ref}})}{\hat{p}^+(1-\mathrm{CPF}_{\mathrm{ref}}) + (1-\hat{p}^+)\mathrm{CNF}_{\mathrm{ref}}}$$

$$D = \frac{(1-\hat{p}^+)\mathrm{CNF}_{\mathrm{ref}}}{\hat{p}^+(1-\mathrm{CPF}_{\mathrm{ref}}) + (1-\hat{p}^+)\mathrm{CNF}_{\mathrm{ref}}}. \tag{38}$$

Finally, using A–D and the $\mathrm{CPF}_{\mathrm{obs}}$ and $\mathrm{CNF}_{\mathrm{obs}}$, estimates of the true performance of the decision process can be determined:

$$\hat{\mathrm{CPF}}_{\mathrm{true}} = \frac{\mathrm{CPF}_{\mathrm{obs}} - B + B\cdot\mathrm{CNF}_{\mathrm{obs}}/D - B\cdot C/D}{A - B\cdot C/D} \tag{39}$$

$$\hat{\mathrm{CNF}}_{\mathrm{true}} = \frac{\mathrm{CNF}_{\mathrm{obs}} - C\cdot(1-\mathrm{CPF}_{\mathrm{ref}})}{D}. \tag{40}$$

The most common problem encountered with this approach is that the estimated p^+, $\mathrm{CPF}_{\mathrm{true}}$, $\mathrm{CNF}_{\mathrm{true}}$ can be less than 0, or greater than one. This is an unfortunate consequence of the independent binomial sampling variance in the reference and decision CPF, CNF.

5.3. Decision comparisons and combinations

We have omitted the important matter of statistical procedures for comparing the attributes of two candidate decision processes: comparisons of proportions, both paired and unpaired (e.g., CPF, CNF, PPV, NPV), comparisons of AUC's, and comparisons of LR's. There are also many situations in which independent decision processes can be combined to yield performance that exceeds that of the individual processes (commonly termed ensemble prediction or aggregation). The complexities and nuances of these comparisons are treated in Pepe [5] and Zhou et al. [32] with far more erudition than we could aspire to in this introduction.

6. Examples

6.1. Classification

A recent study assessed the performance of a prototype NIR spectroscopic classifier for indications of type-2 diabetes in a 145 subject case-control study, which is described in detail elsewhere [33]. The system had been calibrated in a separate study some months previous, so the objective of the case-control study was to validate the performance of the device. In typical case-control fashion, approximately equal numbers of diabetic and non-diabetic subjects participated (69 and 76, respectively), and repeat spectroscopic measurements were acquired for each subject. The output of the multivariate classifier was designed to be a quantitative indicator of the probability of type-2 diabetes.

The histograms of the classifier output for the diabetic (+) and non-diabetic (−) subject measurements are shown in the left of Fig. 7. The ROC curve shown in the right of Fig. 7 was determined from the subject-average measurements, and we have shown Wilson 95% joint coverage intervals for the individual CPF, CNF pairs using the method discussed in Section 2.2. (The ROC curve for the individual measurements is virtually identical, but has artificially narrow intervals because of the repeat measurements.) Again, we remind the reader that these coverage intervals cannot be construed as representing the confidence interval for the ROC curve, as it will be considerably narrower. The AUC and the confidence interval for the AUC (inset in the ROC figure) were calculated using the Mann–Whitney U statistic, again on the subject-average measurements.

The classification rate, if calculated from this study, would suggest an optimum of 0.77, at a CPF=0.87 and CNF=0.68. However the proportion of positive cases (0.48) is significantly elevated in this study relative to the population at large (~0.1), so the stated classification rate of 0.77 is inappropriate. If the population proportions are substituted for the study proportions in Eq. (5), using CPF=0.87, CNF=0.68 a realistic estimate of the classification rate in the population at large is 0.67. The designation of the threshold for the decision process does not end here, however. For the decision process to prove advantageous in large-scale testing for diabetes, a cost function analysis (see Section 5.1) is necessary to determine the operating point with optimum cost-benefit, and as with all candidate medical devices, it must prove efficacious in a realistic clinical setting.

6.2. Limit of detection

The limit of detection is an analytical figure of merit that has also historically been tethered to the assumptions of asymptotic normality. It is commonly defined as the analytical concentration which is three times greater than the standard deviation of concentration readings from a blank:

$$\mathrm{LOD} = 3\sigma_{\mathrm{blank}}. \tag{41}$$

Three standard deviations on the standard normal variate corresponds to a false-positive probability of approximately 0.0013. More appropriately, analytical detection decisions should not only relate to false-positives, but also incorrect negatives—failures to detect the substance when it is truly present—which has been discussed by several researchers in recent publications [34–36]. The ROC curve and the AUC is quite an attractive non-parametric option for describing the detection characteristics of an analytical system when parametric assumptions such as normality and homogeneity of variance are not likely applicable. An ROC-type approach was very recently employed by Christesen for characterizing biological detection schemes [37], and a recent DARPA report [38] strongly advocates for ROC curve analysis of chemical devices for chemical/biological agent detection. Decision theoretic alternatives to the ROC for limit-of-detection applications have also recently been discussed [39].

A combined ROC and AUC method for characterizing detection decisions using non-parametric means is illustrated in Fig. 8 using simulated data. Behavior like that shown in Fig. 8 is common for enzymatic assays, and often does not meet the requirements of either normality or homogeneity of variance over the concentration range. For

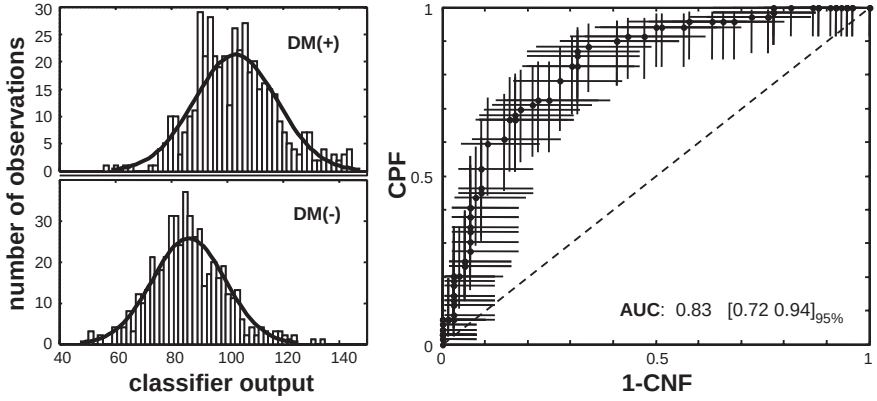

Fig. 7. Illustration of ROC use for assessment of a NIR spectroscopy-based classifier for indications of diabetes mellitus (DM). Bars on the ROC points reflect 95% coverage for each CPF, CNF pair using Wilson intervals for the proportions.

this simulated protocol, 10 samples were prepared at each of the following 6 discrete concentration levels: 0, 2, 4, 6, 8, 10 mM. The reference is therefore defined according to whether the analyte is present at non-zero concentration:

$$R = \begin{cases} - & \text{if} \quad c = 0 \\ + & \text{if} \quad c > 0 \end{cases} \tag{42}$$

and the analytically reported concentration (y) at each threshold t was converted to a positive/negative decision in standard fashion as

$$D = \begin{cases} - & \text{if} \quad y < t \\ + & \text{if} \quad y \geq t \end{cases} \tag{43}$$

Each of the ROC traces (which we have fit using the binormal model, with a logarithm transformation of y to account for the non-normality in the data) in Fig. 8 allows examination of the positive and negative detection characteristics at a fixed reference concentration level, the object being to determine the reference concentration level at which the new analytical method can reliably detect non-zero concentrations with reasonable power (CPF) and false-positives (1 − CNF). Manual inspection of these traces suggests that if 'reasonable power' is determined to be 80% (CPF = 0.8), that 6 mg/dL is likely the minimum concentration that can be detected without excessive false-positives (5%, 1 − CNF = 0.05). The ROC points are quite choppy for such small sample sizes, however, so alternatively one can examine the AUC and its estimated confidence intervals, as we show in Fig. 9.

As an alternative to defining LOD in terms of type I and II errors, one could strive for a 'superiority specification', for which the AUC is very well suited. For instance in Fig. 9 we have drawn a horizontal line at an AUC of 0.95, which, by virtue of its correspondence to the U test, implies that reference concentration levels with AUC's above this line ensure greater than 95% stochastic dominance—the method reading at a true reference concentration of 'X' will exceed the reading from a blank greater than 95% of the time.

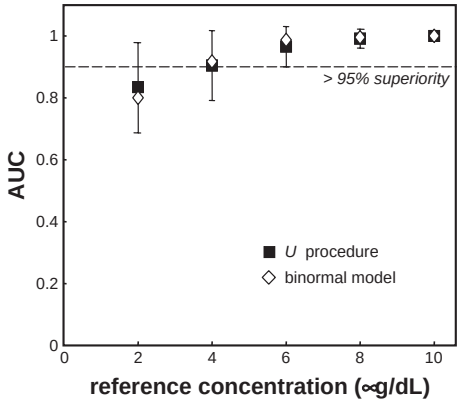

Fig. 9. AUC determined using the Mann-Whitney U procedure (■) and its 95% confidence intervals, when using the ordinate concentration as a threshold of detection. The estimated AUC's using a binormal assumption are also shown (◇).

Although the point estimate for the 6 μg/dL reference concentration exceeds 95%, based on the interval estimates there roughly 50% chance that the true AUC may be lower that 95%. The 8 μg/dL level is a safer bet. This procedure isn't likely to be considered for standard adoption, but it is simply another illustration of how these tools can be employed to inform practical chemical decisions.

7. Summary

Receiver operator characteristic curves, and related decision theory measures are rapidly becoming the standard approach for describing the merits of proposed dichotomous decision processes in many scientific fields; journals dealing with medical and clinical screening and diagnosis, for example, now strongly recommend ROC curve analysis where appropriate in their instructions to authors. As one reviewer noted, the fact that ROC curves have not been used in chemometrics may be simply due to the absence of ROC analysis options in commercial chemometric software. The simplicity of the

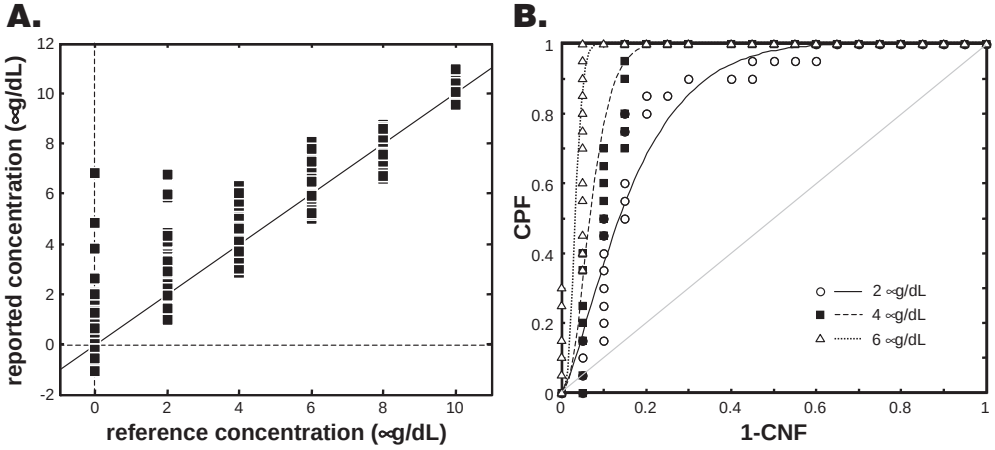

Fig. 8. Data from a hypothetical LOD experiment to evaluate operating characteristics of various candidate detection concentration thresholds. A. Predicted versus reference concentrations for the LOD experiment. B. ROC curves for detection of various concentrations in the experiment.

analyses means that it can be done in Excel (or even by hand) with very little effort, so optimistically this tutorial will inspire some readers to consider decision theoretic measures even if packaged software is lacking. There are also many free or low-cost programs available on the internet for ROC analysis.

The value of the ROC curve is primarily its independence from the proportion of positive/negative events (which is not the case for the classification rate), and a transparent graphical summary of the entire space of possible operating characteristics of a decision process in terms of the correct positive and negative fractions. The area under the ROC curve also makes for an extremely convenient non-parametric summary measure independent of a decision threshold. In predictive applications where the prior probabilities of positive/negative events can be accurately specified, users may find more value in measures such as the positive and negative predictive values or likelihood ratios, which are better suited to characterizing the probability of events given a specific decision. Correct and incorrect decisions often come with associated costs or losses, and the ROC can easily be converted into expected cost curves, which describe the total expected cost of operating at the various possible decision thresholds and minimize at the most cost-effective decision operating point. We provided several examples to illustrate the use of these measures, although there are certainly many more potential chemometric applications; characterizing the efficiency of outlier detection methods, and model selection are ones that come immediately to mind.

Scientific intuition is usually relied on to bridge the gap between data and decision, and will always be indispensable for doing so, but hopefully this tutorial provides a practical introduction of some simple decision theoretic tools that can provide objective perspective for the data-to-decision process.

Acknowledgements

The authors would like to thank Veralight, Inc. (Albuquerque, NM) for consenting to publication of the non-invasive NIR diabetes classification data.

Appendix A. Matlab code for binormal receiver operator characteristic curve and AUC estimation

```
function [FPF,CPF,AUC] = binormROC(y,class);

% Function to compute binormal receiver operating characteristic curves.
% Requires the Matlab Statistics toolbox for the normal CDF and inverse
% CDF.
%
% Usage: [FPF,CPF,AUC] = binormROC(y,class);
%
% Inputs:
%      y - nx1 vector of responses
%      class - nx1 membership vector of 0/1 values:
%              negatives taken to be 0's
%              positives taken to be 1's
%
% Outputs:
%      FPF - vector of fitted false positive fraction values (1-CNF)
%      CPF - vector of fitted correct positive fraction values
%      AUC - estimated area under the ROC curve

k0 = find(class == 0);              % extract negative events
k1 = find(class == 1);              % extract positive events

mu0 = mean(y(k0));                  % Calculate moments of (-)
s0 = std(y(k0));

mu1 = mean(y(k1));                  % Calculate moments of (+)
s1 = std(y(k1));

a = (mu1-mu0)/s1;                   % Parameters for binormal curve
b = s0/s1;                          % as per equation 15

FPF = linspace(0.01,0.99);          % ordinate axis for the fit
tmp = norminv(FPF,0,1);
CPF = normcdf(a+b*tmp);             % estimated CPFs from the cumulative
                                    % normal density

AUC = normcdf(a/sqrt(1+b*b));       % binormal AUC estimate as per
                                    % equation 18
```

References

[1] J.A. Swets, Science 240 (1988) 1285–1293.
[2] R.M. Centor, Medical Decision Making 11 (1991) 102–106.
[3] N.A. Obuchowski, Radiology 229 (2003) 3–8.
[4] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, John Wiley & Sons, New York, NY, 1974.
[5] M.S. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, New York, NY, 2003.
[6] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, 2nd ed., Cambridge University Press, New York, NY, 1992.
[7] G.W. Snedecor, W.G. Cochran, Statistical Methods, 8th ed., Iowa State University Press, Ames, Iowa, 1989.
[8] E.B. Wilson, Journal of the American Statistical Association 22 (1927) 209–212.
[9] L.L. Brown, T.T. Cai, A. DasGupta, Statistical Science 16 (2001) 101–133.
[10] R.A. Hilgers, Methods of Information in Medicine 30 (1991) 96–101.
[11] W.P. Tanner, J.A. Swets, Psychological Review 61 (1954) 401–409.
[12] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.
[13] J.A. Swets, R.M. Pickett, Evaluation of diagnostic systems: methods from signal detection theory, Academic Press, New York, NY, 1982.
[14] A.R. Henderson, Annals of Clinical Biochemistry 30 (1993) 521–539.
[15] K. Linnet, Clinical Chemistry 34 (1988) 1379–1386.
[16] M.H. Zweig, G. Campbell, Clinical Chemistry 39 (1993) 561–577.
[17] F. Provost, T. Fawcett, R. Kohavi, Proceedings of the Fifteenth International Conference on Machine Learning 1998, San Francisco, CA, Morgan Kaufmann, San Mateo, CA, 1998, pp. 445–453.
[18] J.A. Hanley, B.J. McNeil, Radiology 143 (1982) 29–36.
[19] D. Bamber, Journal of Mathematical Psychology 12 (1975) 387–415.
[20] B. Efon, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC Press, New York, NY, 1993.
[21] E.J. Boyko, Medical Decision Making 14 (1994) 175–179.
[22] B. Dujardin, J. van den Ende, A. van Gompel, J.P. Unger, P. van der Stuyft, European Journal of Epidemiology 10 (1994) 29–36.
[23] D.L. Simel, G.P. Samsa, D.B. Matchar, Journal of Clinical Epidemiology 44 (1991) 763–770.
[24] S. Baker, Communications in Statistics. Theory and Methods 20 (1991) 2739–2752.
[25] P.N. Valenstein, American Journal of Clinical Pathology 93 (1990) 252–258.
[26] J. Gart, A. Buck, American Journal of Epidemiology 83 (1966) 593–602.
[27] S.L. Hui, X.H. Zhou, Statistical Methods in Medical Research 7 (1998) 354–370.
[28] J.S. Uebersax, W.M. Grove, Statistics in Medicine 9 (1990) 559–572.
[29] H.B. Lipman, J.R. Astles, Clinical Chemistry 44 (1998) 108–115.
[30] A. Hagdu, Lancet 348 (1996) 592–593.
[31] W.C. Miller, Journal of Clinical Epidemiology 51 (1998) 219–231.
[32] X.-H. Zhou, D.K. McClish, N.A. Obuchowski, Statistical Methods in Diagnostic Medicine, J. Wiley & Sons, New York, NY, 2002.
[33] C.D. Brown, H.T. Davis, M.N. Ediger, C.M. Fleming, E.L. Hull, M. Rohrscheib, Diabetes Technology and Therapeutics 7 (2005) 456–466.
[34] L.A. Currie, Chemometrics and Intelligent Laboratory Systems 37 (1997) 151–181.
[35] R. Boqué, F.X. Rius, Chemometrics and Intelligent Laboratory Systems 32 (1996) 11–23.
[36] H. van der Voet, Encyclopedia of Environmetrics, vol. 1, Wiley, Chichester, 2002, pp. 504–515.
[37] S. Christesen, K. Spencer, J. Sylvia, K. Gonser, "SERS of chemical agents in water — determining limits of detection," presentation at Federation of Analytical Chemistry and Spectroscopy Societies (FACSS) Conference; Portland, OR, 2004.
[38] http://www.darpa.mil/mto/people/pms/pdfs/CBS3FinalReport.pdf.
[39] H.T. Davis, E. Merrill, A proposed method to estimate receiver operating characteristic curves for chemical and biological standards, Proceeding of the SPIE Defense & Security Symposium, March 28–April 1, 2005, Orlando, Florida, 2005.