Using receiver operating characteristic analysis to evaluate the accuracy in predicting future quality grade from ultrasound marbling estimates on beef calves

J. R. Brethour[1]

Kansas State University Agricultural Research Center, Hays, KS 67601

ABSTRACT: Ultrasound estimates of marbling score were collected on 144 calves that averaged 219 d of age and 219 kg. Those estimates were correlated (r^2 = .32, P < .001) with carcass marbling at slaughter 252 d later. Receiver operating characteristic (ROC) analysis provides a method to determine critical operating points for selecting outcome groups with specified percentages of the desired quality grades. When an *a priori* estimate of the prevalence of Choice or premium choice can be made, the ROC procedures incorporate error rates to estimate the percentage of the herd that will be selected in an outcome group. A unique feature of the ROC output is the ability to conduct a cost/benefit analysis. The ROC analysis indicated that the initial marbling estimates were 78 ± 4% accurate in classifying future quality grade and predicting whether an animal would grade USDA Choice. A similar accuracy (76 ± 4%) was observed in predicting whether an animal would attain USDA Average Choice or higher (premium choice). A decision matrix was also examined in which sensitivity and specificity in predicting Choice were 90% and 46%, respectively. Relative values of those diagnostic measures were reversed in predicting premium choice (39% and 86%, respectively). Evaluation of the different methods indicated that ROC analysis may have advantages over either traditional regression analysis or contingency tables for evaluating ultrasound procedures.

Key Words: Carcass Quality, Cattle, Ultrasound

 J. Anim. Sci. 2000. 78:2263–2268

Introduction

Ultrasound estimates of intramuscular fat made on feedlot cattle several months before slaughter have correlated with carcass marbling (Brethour, 1990). Pilot studies have indicated that ultrasound estimates of marbling taken on calves at weaning might predict USDA quality grade 10 mo later (Brethour, 1992, 1998). The ability to predict potential quality grade of calves at weaning might enable clustering cattle into groups with similar carcass characteristics, which might facilitate more appropriate management and marketing strategies.

Correlation coefficients are used commonly to evaluate the accuracy of technologies that predict attributes. However, correlation coefficients are biased by sample variability and do not provide quantitative information that allows economic interpretations. In addition, a user might be more interested in categorization and whether ultrasound technology simply will predict whether an animal will grade USDA Choice. That might be evaluated with a traditional two-way contingency table, in which an arbitrary marbling threshold at evaluation time is selected for a decision point. However, the measure of accuracy from such tables is prevalence-dependent. A powerful diagnostic procedure that is especially effective with a continuous predictor variable (such as an ultrasound marbling estimate) is receiver operating characteristic (**ROC**) (Metz, 1978; Swets and Pickett, 1982). This analysis provides a measure of accuracy that is independent of both the prevalence of an attribute and the decision threshold. It is an important diagnostic tool in medical research but has been used little in agriculture (Saveland and Neunenschwander, 1990; Pyorala and Pyorala, 1997; Twengstrom et al., 1998).

This paper reports the relationship of ultrasound marbling estimates on calves with an average age of 219 d when evaluated and the carcass measures at an average of 252 d later and compares evaluation of results by correlation analysis, contingency tables, and ROC analysis.

Materials and Methods

This study was conducted at the KSU Agricultural Research Center at Hays, KS with 144 calves (130

[1]Correspondence: phone: 785-625-3425; fax: 785-623-4369; E-mail: Jbrethou@oz.oznet.ksu.edu.

Received September 28, 1999.
Accepted May 2, 2000.

steers, 14 heifers) produced from the center cow herd. Breed composition of cows included proportions of Angus to Simmental ranging from 12 to 87%, and calf sires included Angus, Limousin, and South Devon. Average birth date of calves was March 4, 1997, and average age and BW at insonation were 219 d (SD = 20 d) and 259 kg (SD = 37 kg), respectively.

On October 9 and 10, 1997, a few days after calves were weaned, they were evaluated for marbling using an Aloka 210 (Wallingford, CT) ultrasound system equipped with a 3.5-MHz linear array transducer (UST 5021-125 mm window). A sagittal tomogram of the longissimus was obtained caudal to the last rib and approximately 8 cm distal to the back. Marbling was estimated by an image analysis procedure (Brethour, 1994) that has been validated for an accuracy of approximately .4 marbling score unit in the range of estimates encountered in this study.

Calves were group-fed in pens of approximately 25 and received a high-energy ration composed of rolled sorghum grain, sorghum silage, soybean meal, urea, and a vitamin-trace mineral premix. Animals were implanted (Synovex, Fort Dodge Animal Health, Overland Park, KS) soon after weaning. They were marketed in three groups according to BW and ultrasound backfat measures on May 27 (231 d after evaluation, 88 cattle), July 9 (274 d, 27 cattle), and July 30, 1998 (295 d, 19 cattle). The purpose of marketing in groups was to minimize carcasses of less than 249 kg (two cattle) or with more than 20 mm of backfat (five cattle). Average carcass weight was 345 kg (SD = 41 kg), and ADG from evaluation to slaughter averaged 1.20 kg. Carcass backfat thickness averaged 11 mm; 71.5% of the carcasses graded USDA Choice and higher and 22.9% graded Average Choice and higher.

Cattle were slaughtered at a commercial packing plant (IBP, Emporia, KS) and carcass marbling scores were assigned to the nearest .1 unit by experienced USDA graders after a 24-h chill at 0°C. Marbling scores were coded so that 4.0 = slight^{00} (Low Select) and 5.0 = small^{00} (Low Choice). A linear model that included date of birth, sex, percentage Angus, initial ultrasound marbling estimate, and slaughter date was used to adjust actual marbling scores in the three marketing groups to an average slaughter date. A carcass was deemed to be USDA Select if the adjusted marbling score was less than 5.0, USDA Choice if it was 5.0 or higher, and "premium choice" (such as Certified Angus Beef) if it was 6.0 and above (equivalent to USDA Average Choice). The term "commodity" was selected for the category of carcasses that graded less than Average Choice.

Linear, exponential, and modified power functions (Spain, 1982) were compared for the relationship of the ultrasound marbling estimate at weaning and carcass marbling score using routines available on a Lotus 123 spreadsheet. The values on the abscissa (Figure 1) were jittered slightly (.02 marbling score unit) in order to display those data points that were otherwise superimposed. Contingency tables were constructed and tested with χ^2 using the Yates correction for continuity (Shavelson, 1996). The critical operating point for predicting Choice in Figure 2A was a weaning marbling score of 3.8, because that resulted in the highest estimate of accuracy (true positive fraction [**TPF**] or animals correctly predicted to be Choice plus true negative fraction [**TNF**] or animals correctly predicted to be less than Choice). However, that procedure was not feasible in building a contingency table for detection of premium choice, because the highest accuracy would have been obtained by predicting that none would attain that

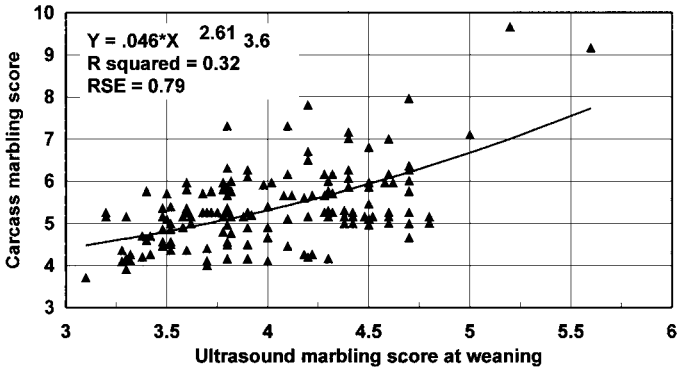

Figure 1. Ultrasound marbling estimate on calves (average age = 210 d) and carcass marbling scores 252 d later.

2 X 2 Contingency Table — Detection of USDA Choice — A

Predicted Choice Graded Choice True Positive Fraction 93 head (65%)	Predicted Choice Graded Select False Positive Fraction 22 head (15%)
Predicted Select Graded Choice False Negative Fraction 10 head (7%)	Predicted Select Graded Select True Negative Fraction 19 head (13%)

Sensitivity = 90.3% [TPF / (TPF + FNF)]
Specificity = 46.3% [TNF / (FPF + TNF)]
Accuracy = 78% (TPF + TNF)
Chi suared = 22.2 (P<.01)

2 X 2 Contingency Table — Detection of "premium choice" — B

Predicted premium choice Graded premium choice True Positive Fraction 13 head (9%)	Predicted premium choice Graded commodity False Positive Fraction 15 head (10%)
Predicted commodity Graded premium choice False Negative Fraction 20 head (14%)	Predicted commodity Graded commodity True Negative Fraction 96 head (67%)

Sensitivity = 39.4% [TPF / (TPF + FNF)]
Specificity = 86.5% [TNF / (TNF + FPF)]
Accuracy = 76% (TPF + TNF)
Chi suared = 9.29 (P<.01)

Figure 2. Contingency table for discriminating ability to predict USDA Choice (A) or premium choice (B) from ultrasound evaluations on young cattle.

grade, a pitfall in the use of contingency tables. A decision point of marbling score = 4.5 was chosen for this table (Figure 2B), because that represented the value in the power function model that corresponded to the lowest carcass marbling score (6.0) required for premium choice. Sensitivity is TPF as a percentage of TPF plus the false negative fraction (**FNF**) or animals that were erroneously predicted not to grade Choice. Specificity is TNF as a percentage of TNF plus the false positive fraction (**FPF**) or cattle that were incorrectly predicted to grade Choice.

Software (Labroc1) for performing the ROC analysis was provided by Charles E. Metz, Department of Radiology, The University of Chicago. This software is available online (Metz, 1999). The program performs a maximum likelihood estimation of a binormal ROC curve from a set of continuously distributed data. The ROC curve plots the true positive fraction as a function of the false positive fraction. The program also enables plotting the points on the ROC curve as a function of the corresponding operating points (estimated marbling score). Although this relationship is globally nonlinear, linear and quadratic models enabled usable approximations in the region of interest. In addition, the effect of different *a priori* estimates of the prevalence of Choice in a herd on projected outcomes at different operating points was examined. The ROC procedures were used to test the ability to correctly predict the likelihood of grading Choice (or premium choice), which was assigned the positive classification.

Results and Discussion

Correlation Analysis

The relationship of the ultrasound marbling estimates at weaning with carcass marbling score is shown in Figure 1. A modified power function (r^2 = .315, P < .001) was better (P < .05) than a linear model (r^2 = .296, P < .001) in expressing the relationship of carcass marbling to initial ultrasound estimate. An exponential model was numerically lower than the power function (r^2 = .308). The nonlinear models corroborate with previous observations that the rate of marbling increase during feeding is higher among cattle with higher initial marbling levels. However, two outliers that were predicted correctly to grade very high biased the correlation analysis. When they were omitted, the r^2 values dropped to approximately .205, and no differences occurred among the three models. This disproportionate effect of two outliers demonstrates a pitfall with correlation analysis. Even though the correlation coefficients were highly significant, the residual standard error (**RSE**) for carcass marbling was only 17% lower than the SD (.79 vs .94).

Three major components of error exist in predicting future quality grade. One is the accuracy of ultrasound technology to measure a present value of marbling in the live calf. The average amount of marbling at evaluation was 4.0 (slight^{00}, SD = .47), which is approximately equivalent to 3% ether extract. Herring et al. (1998) reported that the most precise ultrasound systems currently used for measuring marbling have standard errors of prediction ranging from .7 to .9 marbling score units, although the range of marbling scores was wider in their study. Differences among animals, as well as effects of environment and health, may affect the rate of marbling increase during the feeding period. Also, the determination of carcass marbling is subjective and varies among graders, carcass temperature, lighting, bloom time, and ribbing technique.

Contingency Tables

The 2 × 2 contingency tables for predicting USDA Choice (Figure 2A) and premium choice (Figure 2B) both show significant discrimination ability (χ^2 = 22.2 and 9.3, P < .001 and .01, respectively). The sensitivity for predicting USDA Choice was high (90%), but the specificity was low (46%). The reverse was true for predicting premium choice (39 and 86%, respectively). An exchange between high sensitivity and specificity is normal in diagnostic performance (Friedlander, 1999). This means that the procedure was proficient in detecting a large proportion of the herd that would grade USDA Choice but poor at classifying with much certainty the cattle that would grade USDA Select. However, more than half of those animals that attained premium choice were missed, but cattle that did not attain that grade were efficiently culled (only one animal that graded Select was in the FPF group). Poor ultrasound image acquisition and interpretation occasionally may result in marbling estimates that are lower than they should be. Good sensitivity may be more important than specificity to identify as many Choice candidates in a herd as possible. In most applications being certain that an animal in the TNF group and expected to grade Select actually does so would not seem important. However, high specificity in sorting for premium choice to avoid including commodity cattle in that group might be more important, even though a proportion of those capable of attaining the higher grade are missed. A disadvantage of the contingency table is that it does not allow an objective way to select the critical operating point for the decision process.

ROC Analysis

The ROC curves are presented in Figures 3A and 3B. These curves portray the true positive fraction as a function of the false positive fraction. The percentage of area under the curve in ROC space is an estimate of the prediction accuracy, and spaces were 78 and 76% for estimating Choice and premium choice, respectively. If there is no discrimination, the line will be a diagonal from the origin to the upper right. The curve in Figure 3A rises steeply at the beginning and then climbs more gradually through the rest of the chart; this may reflect

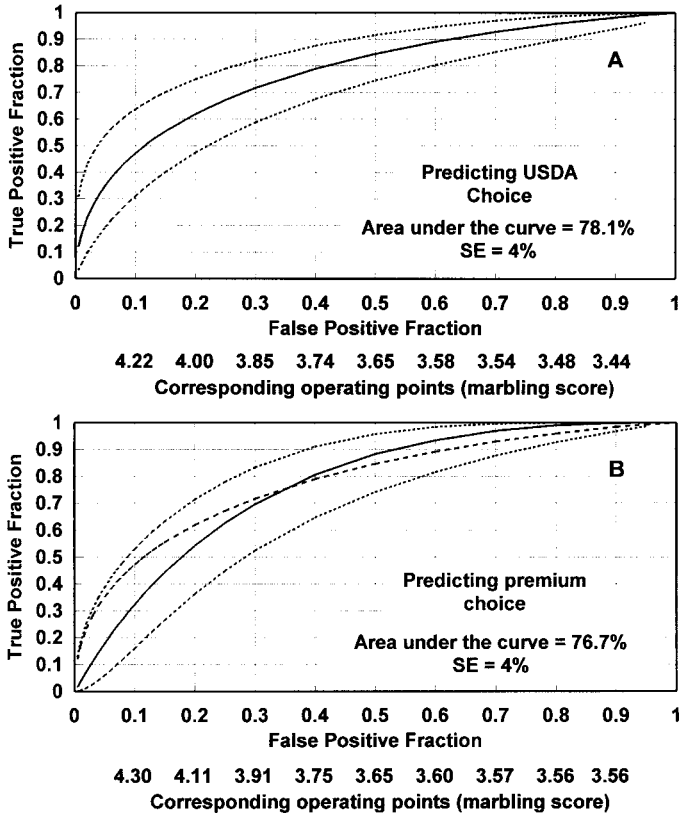

Figure 3. Receiver operating characteristic (ROC) analysis for predicting USDA Choice or premium choice (A or B, respectively) from ultrasound marbling estimates on calves (solid line). Dotted lines show the 95% confidence interval. Dashed line on Figure 3B is the superimposed ROC curve from Figure 3A. True positive fraction represents Choice or premium choice.

the greater sensitivity in selecting Choice cattle. However, the ROC curve in Figure 3B bulges higher at the right side of the chart. This may be caused by the greater specificity in identifying animals that will not grade premium choice.

The ROC analysis provides a method to choose a critical operating point to attain a desired proportion of Choice or premium choice in an outcome group (Figures 4A and 4B). That value depends on an estimate of the percentage Choice or premium choice in a population. For example, in a set of calves that should grade 70% Choice based on past performance, a critical operating point of approximately 3.7 marbling score units would enable selecting a set expected to grade 80% Choice. If the herd is expected to grade 50% Choice, then that critical point must be increased to nearly 4.2 to accomplish the expectation of an outcome group that would grade 80% Choice. An important consequence is that the likelihood of a calf with a specific marbling score attaining Choice is dependent on an *a priori* estimate of Choice prevalence in the herd.

Figures 5A and 5B are useful extensions of the analysis, because they show the proportion of the herd that will be selected in the outcome group to meet expectations. In the examples mentioned above, upgrading a herd from 70% Choice to 80% Choice should result in 70% of the animals being selected, whereas an attempt to create an outcome group grading 80% Choice in a herd that contains only 50% Choice results in only 30% selected. (The calves in this experiment graded 71.5% Choice. Setting the critical operating point at 3.7 marbling score units would have resulted in selecting 74% of the population and a group that graded 81% Choice). These tools provide a means to determine whether sorting for quality grade with ultrasound will be a feasible endeavor and accurately take into account the error rate.

Economic Analysis

An especially powerful feature of ROC procedures is the ability to select a decision point that maximizes the profitability of the application. The formula (Metz, 1978) for this task is as follows:

$$S = (P\,(\text{Ch}-)/P\,(\text{Ch}+) * ((V_{FP} - V_{TN})/(V_{FN} - V_{TP}))$$

where S = slope of the ROC curve, P (Ch–) = prevalence of Select, P (Ch+) = prevalence of Choice, V_{FP} = value of a false positive estimate, V_{TN} = value of a true negative estimate, V_{FN} = value of a false negative estimate, and V_{TP} = value of a true positive estimate.

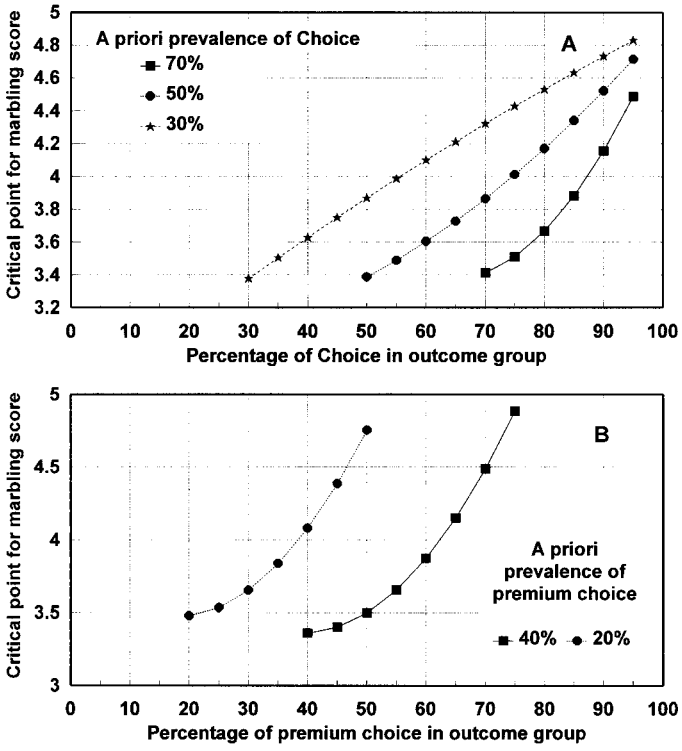

Figure 4. Critical point for marbling score to attain a specified proportion of USDA Choice (A) or premium choice (B).

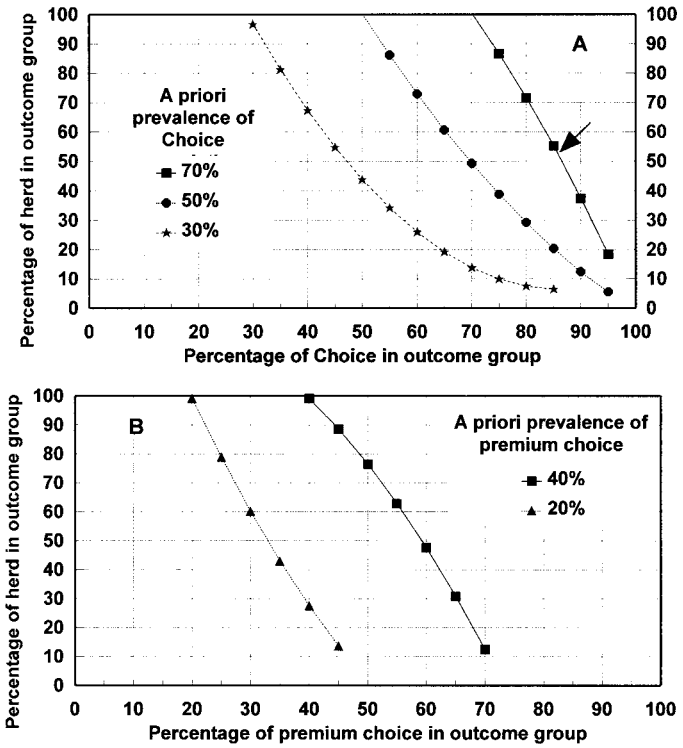

Figure 5. Percentage of the herd selected for the outcome group as a function of the percentage of Choice (A) or premium choice (B) desired and an *a priori* estimate of the composition of the total population. The arrow represents the point that corresponds to 86% Choice in the economic analysis example in the text.

Continuing the example, assign arbitrary values of \$65 profit for a calf sold at weaning (V_{TN}) and \$100 profit for a calf fed to Choice (V_{TP}), but a loss (–\$10) for a calf that is fed out but grades Select (V_{FP}). The V_{FN} value is the same as the V_{TN}, or \$65, because it also represents a calf sold at weaning because it was assigned to that group erroneously. (Admittedly, such values would be virtually impossible to assign in a real situation.) If the prevalence of Choice in the herd is estimated as 70% (P (Ch+)), the solution to the equation for the slope that maximizes profitability is .92. Figure 6 shows the critical operating point (marbling score of 3.85) and the values for TPF (.69) and FPF (.27) that correspond to that slope. The selected group would be expected to grade 86% Choice, which is the solution to the equation (.69 ∗ .7) / (.69 ∗ .7 + .27 ∗ .3). Figure 5A shows that this percentage results in 54% of the herd in the selected set.

The average value of sorting is expressed by

$$V_{AV} = V_{TP} * P\,(TP) + V_{TN} * P\,(TN) + V_{FP} * P\,(FP) + V_{FN} * P\,(FN)$$

where V_{AV} = average value after sorting, V_{TP}, V_{TN}, V_{FP}, and V_{FN} = values of TPF, TNF, FPF, and FNF, P (TP) = probability of true positive or TPF times prevalence of Choice, P (TN) = probability of true negative or (1 – FPF) times prevalence of Select, P (FP) = probability of false positive or FPF times prevalence of Select, and P (FN) = probability of false negative or (1 – TPF) times prevalence of Choice. In the example, the average value after sorting was \$75.83, which should be compared to average values of \$67 if all calves were retained and fed out and \$65 if all were sold at weaning. Exploring different scenarios indicated that response to sorting was greatest when the average value from feeding was the same as that for selling calves at weaning. Also, an approximately \$10 benefit (not including ultrasound costs) from sorting was the most that could be achieved using realistic estimates for values of the different categories and the error rates observed in this experiment.

Application of ultrasound technology to sort calves at young ages for future quality grade will be challenging. In this experiment, computer marbling estimates based on image analysis procedures were verified with visual image assessment by the author, who has had over 12 yr of experience in animal sonography. Although the Aloka 210 system used in this study is no longer manufactured, it seems more accurate for estimating marbling in young animals than newer systems. Many factors affect ultrasound estimates of marbling in the live animal, including technician competency and experience, algorithms for image analysis, conditions at insonation, and different ultrasound systems (substantial differences have been experienced among systems, even of the same model).

Implications

The use of ultrasound technology on young cattle to predict future quality grade seems feasible. However, the relative accuracy of 76 to 78% in predicting carcass grade categories may not be high enough to effect substantial monetary benefits from the procedure. This

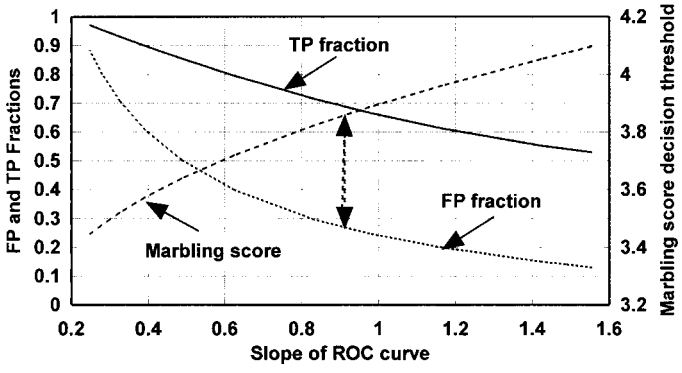

Figure 6. Mapping true positive (TP) and false positive (FP) fractions and the marbling score decision threshold onto the slope of the receiver operating characteristic (ROC) curve (used in economic analysis of sorting procedures). Double arrow represents the optimal slope of .92 presented in the text example.

may limit application until higher accuracy can be obtained. The receiver operating characteristic analysis enables incorporating error rates and performing cost/benefit analyses and is superior to either traditional regression analyses or contingency tables for evaluating ultrasound predictions.

Literature Cited

Brethour, J. R. 1990. Using ultrasound to identify qualitative traits such as marbling. Kansas Agric. Exp. Sta. Rep. Prog. 597:26–30. Manhattan.

Brethour, J. R. 1992. Progress in assessing and predicting marbling in live cattle with ultrasound. Kansas Agric. Exp. Sta. Rep. Prog. 653:10–14. Manhattan.

Brethour, J. R. 1994. Estimating marbling score in live cattle from ultrasound images using pattern recognition and neural network procedures. J. Anim. Sci. 72:1425–1432.

Brethour, J. R. 1998. Evaluating calves with ultrasound at weaning for future carcass potential. Kansas Agric. Exp. Sta. Rep. Prog. 808:1–4. Manhattan.

Friedlander, E. 1999. Profiles and pitfalls. Practical Topics in Lab Diagnosis. Available at http://www.pathguy.com/lectures/profilin.htm. Accessed Sept. 1, 1999.

Herring, W. O., L. A. Kriese, J. K. Bertrand, and J. Crouch. 1998. Comparison of four real-time ultrasound systems that predict intramuscular fat in beef cattle. J. Anim. Sci. 76:364–370.

Metz, C. E. 1978. Basic principles of ROC analysis. Seminars in Nuclear Med. 8:283–298.

Metz, C. E. 1999. ROC analysis. Available at: http://www-radiology.uchicago.edu/krl/toppage11.htm#software. Accessed Sept. 1, 1999.

Pyorala, S. , and E. Pyorala. 1997. Accuracy of methods using somatic cell count and N-acetyl-beta-D-glucosaminidase activity in milk to assess the bacteriological cure of bovine clinical mastitis. J. Dairy Sci. 80:2820–2825.

Saveland, J. M., and L. F. Neunenschwander. 1990. A signal detection framework to evaluate models of tree mortality following fire damage. Forest Sci. 36:66–76.

Shavelson, R. J. 1996. Statistical Reasoning for the Behavioral Sciences. Allyn and Bacon, Boston, MA.

Spain, J. D. 1982. BASIC Microcomputer Models in Biology. Addison-Wesley, Reading, MA.

Swets, J. A., and R. M. Pickett. 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.

Twengstrom, E., R. Sigvald, C. Svensson, and J. Yuen. 1998. Forecasting Sclerotinia stem rot in spring sown oilseed rape. Crop Prot. 17:405–411.