# The Relationship of Instrument Parameters to Performance Within a Survey Peer Group

G. A. Hermann, D. W. Tholen, and N. E. Herrera

*Presbyterian-University of Pennsylvania Medical Center, The College of American Pathologists Computer Center, and the Danbury Hospital*

**Using simulators of transmission imaging, an interlaboratory survey assessed the discriminatory performance of 86 subscribers, each of whom imaged a liver phantom in anterior and right lateral projections. Analysis was by receiver operating characteristic (ROC) with $A_z$, the area under the ROC curve, used as a measure of accuracy unconfounded by decision bias. $A_z$ values were then defined as the dependent variable in a statistical model that related performance to several instrument design and operating parameters. Six of 14 postulated parameters explained approximately half of observed subscriber variability. These were: year of camera manufacture or upgrade, number of photomultiplier tubes, collimator type, total counts collected, use of a Co-57 disk source for imaging the phantom, and computer processing of the image. The findings confirm previous inferences drawn from controlled intralaboratory experimentation, but hitherto unsubstantiated by clinical imaging data.**

**J Nucl Med: 1371–1374, 1984**

Basic and applied scientists have defined the importance of certain fundamental characteristics of scintillation cameras, such as contrast, energy, spatial and temporal resolution, field uniformity, plane sensitivity, and collimation (*1–3*). However, demonstration that optimization of these parameters produces superior diagnostic performance has proved elusive. Previous interlaboratory surveys, for example, have failed to show a significant correlation between subscriber performance and such elements as collimator type, counts collected, energy window width, or year of instrument manufacture (*4*). Hoffer and co-workers (*5*), examining Anger cameras of varying intrinsic spatial resolution, were unable to detect significant differences in observer performance as indexed by lesion detectability.

If it is reasonable to expect that improvement in camera design and a proper choice of operating parameters produce better clinical results, this lack of correlation should be explained. Either the association linking design and operation with clinical results is too weak to be demonstrated, given available data, or/and previously accepted descriptors of performance are defective. Swets and Pickett (*6*), surveying several supposed indices of performance such as sensitivity, specificity, accuracy (overall probability of a correct response), and predictive value, and rejected all of them. They recommend instead the area under the binormal relative operating characteristic (ROC) curve termed $A_z$. This index represents the proportion of the total area of the graph that lies beneath the curve, and is a measure of signal detectability. It may be thought of as the probability of assigning a higher rating score to a target image when it and a nontarget image are randomly obtained and systematically compared (*7*). The principles and the utilities of ROC analysis have been reviewed extensively in recent publications (*8,9*). Goin et al. (*10*) have used this approach to demonstrate significant differences in observer success among digital scintigraphic display modes.

Thus the purpose of this study was to survey a representative group of imaging departments, to evaluate their results in terms of ROC analysis, and to correlate individual rankings with a group of independent variables currently accepted as important determinants of clinical diagnostic accuracy. Since many of these determinants are under the control of the operator, results of this study might eventually be useful in establishing guidelines for improvement of imaging quality through interlaboratory survey techniques.

## MATERIALS AND METHODS

One hundred sixty-eight responses to Nuclear Medicine Imaging Series TSA 1983 of the College of American Pathologists, constituted the data base. This interlaboratory survey program began in 1973 and uses peer-group participation in semiannual exercises designed to assess both individual performance and the

Received Mar. 23, 1984; revision accepted June 29, 1984.

For reprints contact: G. A. Hermann, Custer Laboratories, P-UPMC, 51 N. 39th Street, Philadelphia PA 19104.

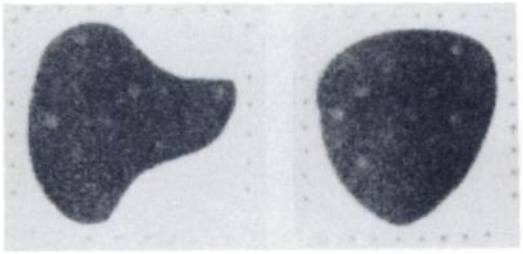

**FIG. 1.** Transmission scintiphotos of liver phantom in anterior and right lateral projections, produced by subscriber. Each projection contains 12 1.0-cm targets with T/B ratios varying from 0.7 to 0.9. Perimeter points are for spatial orientation of subscriber.

overall state of the art in clinical gamma imaging. The rationale, techniques, and findings of this ongoing quality-assurance program have been described in detail (*11*,*12*).

Subscribers received a transmission model containing planar targets randomly distributed to simulate an anterior and a right lateral projection of a human liver containing several simulated mass lesions. Twenty-four of 81 subregions contained targets 1.0 cm in diameter with target-to-background ratios from 0.7 to 0.9. All targets were 1.0 cm from the imaging surfaces and included 0.6 cm of tissue-equivalent Masonite front scatter. Since the model itself is not radioactive, each subscriber provided a flood source of appropriate activity using either Tc-99m or a Co-57 disk, so that the image simulator was interposed between source and detector. Participants were requested to image according to their routine clinical protocol for liver. The resulting images were read according to rating-scale methods (*13*), with assignment of a score from one to four to each subregion. A score of one was accorded those regions the observer felt definitely lacked a target, and a score of four to those areas felt definitely to contain a target. Scores of two and three corresponded to areas probably lacking and probably containing targets.

The subscribers returned their ratings, copies of images (Fig. 1), and a questionnaire with replies specifying 14 operating variables: gamma-camera manufacturer, year of manufacture or update, number of photomultiplier tubes, crystal thickness, collimator type, field uniformity source, energy window width, total counts collected, information density, computer-processing capability, display film type, radionuclide phantom flood, frequency of quality-control testing, and type of quality-control test patterns used.

ROC operating points were calculated for each observer using rating score data. The area under each curve, computed by the trapazoidal rule (*14*), defined an index of discriminatory performance free from extradiscriminatory decision biasing. Most laboratories submitted more than one result because of multiple departmental instruments. In such cases the one best performance was selected for inclusion in the final statistical evaluation. There were 86 of these.

Initially all 14 operating factors were examined separately in one-way analysis-of-variance models to look for performance differences. The six parameters identified as explaining significant variability in performance, in order of decreasing impact were: number of photomultiplier tubes, year of manufacture or update, field uniformity source, collimator type, computer processing, and total counts recorded. These six were then included in a series of two-factor ANOVAs to test for bivariate relationships and for interactions. None of the interaction terms were statistically significant. The factors were then used in a main-effects multivariate least squares model with indicator variables for each factor level

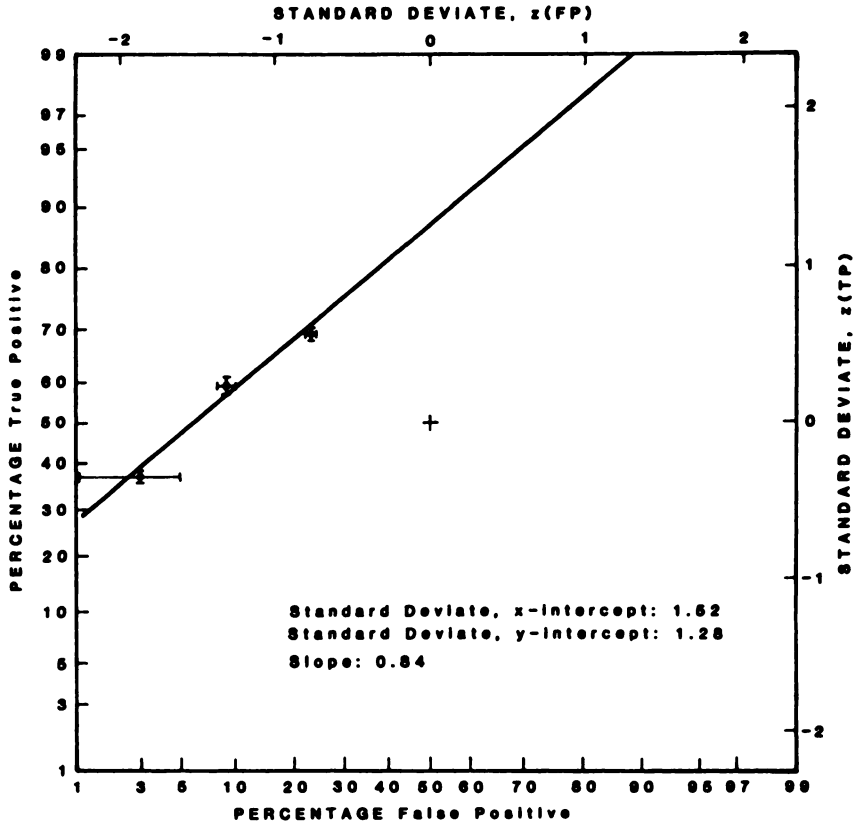


**FIG. 2.** Receiver Operating Characteristic (ROC) plot of pooled rating data. Axes are scaled in normal deviate (z) units with center of ROC space located z(TP), z(FP) = 0,0. Bivariate 95% confidence limits for three decision loci are included.

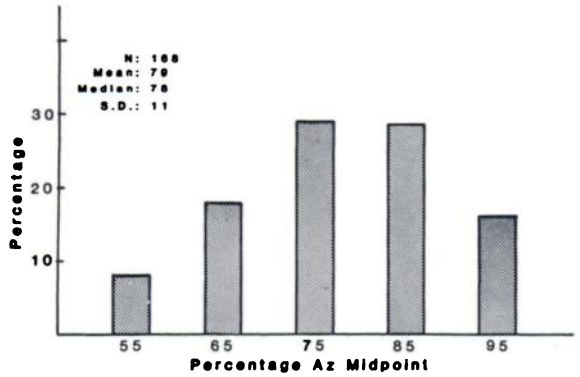


**FIG. 3.** Bar chart of area measure $A_z$ of all reader/instrument combinations.

(*15*). Such indicators have values of one for all observations with their levels, and zero for all other observations. This approach requires that there be one less variable than there are levels of a factor. For example, a factor such as collimator type with three levels (low-energy, all-purpose, high-resolution, and diverging) requires two variables to describe uniquely all responses. This approach is similar to multifactor ANOVA, but avoids some of the problems caused by small group sizes.

Coefficients resulting from this model estimate the effect of the factor level in question (.e.g, low-energy, all-purpose collimator) relative to the level with no associated variable (e.g., high-resolution collimator), which is termed the reference level. The magnitude and sign of the coefficient directly estimate how performance ($A_z$) would change if an observer were to switch from the reference level to the level of interest, holding all other factors at the same levels. Although the coefficients change with differing reference levels, all possible models have the same total predictive ability ($R^2$), and the effect of any factor level relative to any other level remains consistent in all models. For the model reported here, the reference levels are those levels of the imaging factors that are associated with better performance.

**TABLE 1. LEAST SQUARES REGRESSION MODEL WITH REFERENCE POINTS SET AT LEVELS ASSOCIATED WITH BEST PERFORMANCE***

| Factor | Level | Reference Level | Coefficient |
|---|---|---|---|
| Year | 1970–77 | 1980–81 | −0.079 |
| | 1978–79 | | −0.094 |
| | 1980–81 | | — |
| | 1982–83 | | −0.004 |
| No. PM tubes | 19 | 61 | −0.081 |
| | 37 | | −0.037 |
| | 61 | | — |
| | 75 | | −0.058 |
| | 91 | | −0.081 |
| Collimator | LEAP | Hi-resol. | −0.075 |
| | Hi Resol. | | — |
| | Diverging | | −0.175 |
| Total counts | 300–500k | 1000k | −0.024 |
| | 600–900k | | −0.019 |
| | 1000k | | — |
| | 1500–2000k | | −0.012 |
| Radionuclide | Tc-99m flood | Co-57 disk | −0.048 |
| Source-image | Co-57 disk | | — |
| Computer | Yes | Yes | — |
| | No | | −0.100 |
| Intercept | | | 1.051 |
| $R^2$ | | | 0.532 |
| | | | 0.086 |
| | | | 73 |

* Number of observations.

## RESULTS

Figure 2 illustrates the ROC produced from 168 responses by plotting the pooled rating data on axes linear with respect to the standard deviates of the true-positive and false-positive probabilities. Results from the best 86 responses used in the ANOVA and regression studies differ only slightly. Gaussian distributions plotted on such axes yield straight lines whose slopes are proportional to nontarget-to-target variance ratios. Our data are consistent with underlying sensory distributions that are Gaussian and of unequal variances. Decision points are usually plotted on arithmetic scales linear with respect to the true-positive and false-positive probabilities, and produce the familiar curvilinear ROC function.

Figure 3 is a bar graph of the area measured for the 168 responses. It shows that the $A_z$ values themselves are nearly Gaussian in distribution for the sample. A corresponding chart of the 86 best responses is also Gaussian.

The linear regression data are displayed in Table 1, which contains the coefficient associated with each factor level and the reference levels that apply to this model. The table also includes two summary statistics for the regression model: $R^2$, the percentage of the variability explained by the model, and the standard error of the regression.

The coefficient may be interpreted as the mean effect of a particular level of a factor relative to the reference point. For this model, these are the estimated effects of switching from the reference points to the critical level. The intercept is an estimate of performance with all factors at their reference point. The intercept for this model is 1.051, which exceeds "perfect" performance ($A_z$ = 1.00), and reflects a lack of fit of the model. This also implies that "perfect" performance can be predicted with fewer-than-optimal operating parameters. For example, if a participant has an instrument with all factors at the reference levels except for using a Tc-99m flood source for imaging (effect = b = −0.048), the model predicts an $A_z$ of 1.003.

An example of the model's use to predict performance is the following: If a participant's operating parameters included a camera built in 1980, 75 photomultiplier tubes, a high-resolution collimator, 2 million counts, a Co-57 disk source of imaging, and a computer for image processing, the model predicts performance to be:

$$A_z = 1.051 - 0 - 0.058 - 0 - 0.012 = 0.981$$

If the same participant were to change to 600k counts, the estimate changes to:

$$A_z = 0.981 - (0.019 - 0.012) = 0.981 - 0.007 = 0.974$$

Thus the model can be used to estimate $A_z$ values for any combination of levels of these six imaging factors.

## DISCUSSION

The results of the study suggest that the number of photomultiplier tubes, the year of camera manufacture or update, the type of imaging flood source, the collimator type, computer-processing

capability, and the total counts recorded all play measurable roles in determining imaging performance. Note that the term "performance" here refers only to an ability to differentiate multiple areas that contain targets from those that lack them. More complex decision tasks, such as location and diagnostic classification, needed in clinical practice were not required of our observers. Conclusions drawn from these data, therefore, may not describe clinical results perfectly. Nevertheless, successful diagnostic imaging presupposes a basic ability to discriminate among regions containing and those lacking abnormalities.

The six factors listed above explain about half of the variability of performance of the subscribers. There are at least three reasons for this. First, an unknown number of other factors that play a role in the discriminatory process, such as reader ability, were not included in the model. Second, there are small group sizes for some factors, as is frequent with data drawn from noncontrolled survey studies, and this leads to a loss of statistical power. Finally, an obvious but unmeasured correlation exists among several of the factors. For instance, newer cameras may tend to have more PM tubes and are more likely to have computers associated with them than the older cameras, or subscribers using high-resolution collimators may tend to collect fewer counts than those using the low-energy, all-purpose types.

The creditability of proposed performance indices rests upon the reasonableness of their descriptions. Our results generally confirm results found previously in controlled experiments. Newer cameras perform better than their older counterparts. Those collecting 500k or fewer counts clearly produced poorer results than those who used more, with performance improving monotonically with increasing counts. Subscribers who processed their images with computers before interpretation did better than those who did not.

Higher $A_z$ values associated with the use of a Co-57 disk for the actual imaging was a mildly unexpected result. This might be caused by the sporadic problems with mixing and/or wall bowing known to occur with conventional liquid floods. Certainly some participants performed quite well using Tc-99m sources, suggesting that the problem is not inherent in the procedure itself. This finding, while not germane to most clinical imaging, does offer a caveat to those who use liquid floods for uniformity correction: mix well and check for wall bowing.

## REFERENCES


1. ADAMS R, HINE GJ, ZIMMERMAN CD: Deadtime measurements in scintillation cameras under scatter conditions simulating quantitative nuclear cardiography. *J Nucl Med* 19:538-544, 1978
2. HASMAN A, GROOTHEDDE RT: Gamma-camera uniformity as a function of energy and count-rate. *Br J Radiol* 49:718-722, 1976
3. MUEHLLEHNER G, WAKE RH, SANO R: Standards for performance measurements in scintillation cameras. *J Nucl Med* 22:72-77, 1981
4. HERMANN GA, HERERRA NE, HAUSER W: Rationale, techniques and results of a quality control programme of imaging procedures, *Medical Radionuclide Imaging.* Volume I IAEA-SM-210/212. Vienna, International Atomic Energy Agency pp 55-66, 1977
5. HOFFER PB, NEUMANN R, QUARTARAROI, et al: Improved intrinsic resolution: does it make a difference? *J Nucl Med* 25:230-236, 1984
6. SWETS JA, PICKETT RM: *Evaluation of Diagnostic Systems.* New York, Academic press, pp 24-33, 1982
7. HANLEY JA, MCNEIL BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29-36, 1982
8. METZ CE: Basic principles of ROC analysis. *Semin Nucl Med* 8:283-298, 1978
9. SWETS JA: The relative operating characteristic in psychology. *Science* 182:990-1000, 1973
10. GOIN JE, PRESTON DF, GALLAGHER JH, et al: Comparison of five digital scintigraphic display modes. *Med Dec Making* 3:215-227, 1983
11. HERMANN GA, HERRERA NE, HAUSER W: The College of American Pathologists phantom series-an assessment of current nuclear imaging capabilties. *Am J Clin Pathol* 74: 591-594, 1980
12. HERMANN GA, HERRERA N, SUGIURA HT: Comparison of interlaboratory survey data in terms of receiver operating characteristic (roc) indices. *J Nucl Med* 23:525-531, 1982
13. GREEN DM, SWETS JA: *Signal Detection Theory and Psychophysics.* New York, Robert E. Krieger, 1974, pp 40-43
14. BAMBER D: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psych* 12:387-415, 1975
15. NETER J, WASSERMAN W: *Applied Linear Statistical Models.* Homewood, Illinois, Richard D. Irwin, Inc., 1974, pp 320-326