A review based on 848 cases. *Radiology* 118: 101–107, 1976

3. HAMILTON RG, ALDERSON PO, MCINTYRE PA: Technetium-99m phytate as a bone marrow imaging agent. *J Nucl Med* 18: 563–565, 1977

Motion-Corrected Hepatic Scintigraphy

The article by Turner et al. (*1*) is a model of the performance of the multiple-reader study. One must be cautious, however, in accepting the conclusion that the method of correction for obtaining liver images yields greater accuracy in reading than uncorrected scans. The authors have discussed the difficult problem of significance of separation of receiver operator characteristic curves in an honest and straightforward manner. Since error bars overlap for Observers Nos. 2, 3, 4, and 5 in ROC curves related to uncorrected images for virtually all points plotted, one might conclude from the study that four out of five readers found no significant difference in reading liver scans, even when employing the analog motion correction device.

For example, Turner et al. in previous correspondence noted conditional probabilities for Reader No. 3 as clearly separated on the two ROC curves for uncorrected and corrected plus uncorrected images for (pS|s) equals 0.59 ± 0.07 and 0.76 ± 0.06 at $p(S|n) = 0.04$. However, 95% confidence limits for this certainty must be examined. Two standard deviations above the lower probability of 0.59 (uncorrected scan) equals a probability of 0.73, while 2 standard deviations below the upper point of 0.76 equals 0.64, an obvious overlap. Similarly, for Observer No. 4 Turner et al. felt that the point on the ROC curve of both techniques with $p(S|n)$ of 0.21 ± 0.06 at $p(S|s) = 0.84$ was clearly different than that of $p(S|n) = 0.06 \pm 0.03$ at $p(S|s) = 0.86$ for the uncorrected image. However, the probability of 0.21 ± 0.06 means that 2 standard deviations (95% confidence limits) below that probability is a probability of 0.09, while 2 standard deviations above 0.06 ± 0.03, is 0.12. Clearly, there is an overlap again at the 95% confidence limits.

For each of five readers the uncorrected image yields a ROC curve that lies below that of the corrected image. At first glance one might feel that this must be statistically significant, and using the binomial expression $(0.5)^n$ (where $n = 5$), Turner et al. suggest that the probability of getting such a result, with the uncorrected curve giving poorer results than the corrected curve five out of five times is 0.03. However, this binomial test should have two "tails," since either the corrected or the uncorrected technique could have been better. The actual probability is therefore 0.0625 that five of five readers would find one or the other technique preferable.

The above discussion is not to deny that Turner et al. *may* be correct in their conclusion, for the data suggest that motion correction may provide more accurate hepatic imaging. Because of the overlap in 95% confidence limits for four out of five readers, however, and the probability in excess of 0.05 that five out of five readers might prefer the same technique, it is suggested that other definitive studies of this interesting technique should be performed before we all make this modification on our gamma cameras.

EDWARD B. SILBERSTEIN
University of Cincinnati
Cincinnati, Ohio

REFERENCE

1. TURNER DA, FORDHAM EW, AMJAD A, et al: Motion corrected hepatic scintigraphy: Objective clinical evaluation. *J Nucl Med* 19: 142–148, 1978

Reply

We wish to thank Dr. Silberstein for his interest in our work and his kind remark regarding it.

Statistical testing of the separation of receiver operating characteristic (ROC) curves is difficult. Although the problem is being vigorously investigated in several quarters, and an answer may be close at hand, no method that is entirely satisfactory for clinical experiments has yet been described. In particular, we are unaware of a method that appropriately tests the separation of ROC curves generated from statistically dependent (correlated) sets of observations. For reasons outlined below, it appears to us that Dr. Silberstein's analysis is inappropriate.

In our experiment (*1*), all five observers performed better reading corrected scintigrams (CS) than reading uncorrected scintigrams (US). The strikingly superior performance of Observer 1 reading CS in contrast with his performance reading US certainly should be statistically significant. Although the improvement in performance of Observers 2 and 4 with motion correction and Observers 3 and 5 reading both types of study together was smaller, it was not trivial, especially in the region of clinical interest (i.e., the left side of the curves). The statistical significance of this improvement is uncertain, however, because we have no appropriate way of testing it. Dr. Silberstein has suggested that we look at the error bars and infer significance or lack of significance of curve separation from the presence or absence of overlap of the bars. This is inappropriate, however, because the data from which the curves were generated are correlated (*1*). Since the data are paired, the error bars underestimate the significance of the separation of the curves, and no valid conclusion about the significance of that separation can be drawn from them.

Dr. Silberstein has stated that the fact that five out of five observers performed better with motion correction than without is not significant by the sign test. He refers to a "two-tailed" sign test that yields a P value of 0.0625. In the first place, the propriety of his use of a two-tailed test is open to question. We have asked the question, "Is motion corrected scintigraphy better than uncorrected scintigraphy?" The appropriate test in this case has one "tail" and yields a P value of 0.03, a result generally considered to be statistically significant (i.e., $p < 0.05$). Furthermore, even if one chooses to use the "two-tailed" test, a P value of 0.0625 is very close to 0.05 and, therefore, a very important result, although not technically a "significant" one.

We have interpreted our data as suggesting that analog motion correction can improve the inherent detectability of mass lesions in the liver, provided that the motion correction device is properly calibrated, the spatial resolution of the imaging system and the counting rate are sufficient, the count density is high enough, and so on (*1*). In spite of the difficulties relating to the statistical analysis of the data, we continue to hold that opinion.

DAVID A. TURNER
Rush Medical College
Chicago, Illinois