Phantoms have been developed for testing ultrasound imaging performance based on detectability of spherical simulated lesions as a function of depth. Each phantom consists of a block of liver-mimicking material with c = 1540 m/s, α/f = 0.5 dB/cm/MHz and backscatter coefficient simulating that of liver. Embedded in a vertical midplane of the phantom is a regular array of spherical simulated lesions, all having the same diameter and object contrast. Object contrast = $10 \log_{10} (\eta_{targ} / \eta_{BKGD})$ where η_{targ} is the backscatter coefficient of the sphere (target) material and η_{BKGD} is that of the surrounding (background) material. The values of the sphere diameters are 2, 3, 4 and 5 mm, and the object contrasts are -3, -6, -9 and -14 dB.

Automation allows determination of the depth ranges over which lesions of a given diameter and object contrast are detectable. First, a target image is digitized where the scan slice is centered on the plane containing the sphere centers. The positions of all sphere centers are determined by analyzing the target image. Second, using the same scanner settings, four independent images of the background material are digitized (no targets in the scan slice). Third, a *lesion signal-to-noise ratio* (LSNR) is computed at each lesion site. The numerator of the LSNR equals the mean pixel value over the target area minus the mean pixel value in the area surrounding the target. The denominator of the LSNR equals the standard deviation of a large set of pixel value means from the four background images, each mean being over an area equal to the target area; the positions of these areas are within 1 cm of the target coordinates in the target image.

To determine detectability automatically a threshold LSNR value is needed. To establish this threshold, a two-alternative-forced-choice (TAFC) was done employing 5,600 image pairs, where one image contained a barely detectable -- or undetectable -- target image and the other was of background. Three human observers agreed rather well. A TAFC fraction correct of 7/10 corresponded to an LSNR value of -2, and -2 was chosen as the detectablity threshold.

For each scanner configuration (transducer, frequency, focus depth, etc.) a "figure of merit" (FOM) was computed at each depth. The FOM equals the number of phantoms for which spheres were detectable at that depth. The maximum FOM was 8, the number of phantoms in the study.

To assess the clinical relevance of the FOM, images of *in vivo* human thyroid nodules were digitized for assessment by radiology residents. In addition to imaging the nodules with the transducer on the skin, a set of tissue-mimicking stand-off pads were interposed, thus allowing a variety of nodule depths to be represented. Each nodule image was displayed with no scanner or depth information. Each resident scored the "quality" of the nodule image from 1 (worst) to 5 (best).

For each scanner configuration the FOM scores from the phantom/automation analysis were scaled from a minimum of 1 to a maximum of 5 putting the clinical and phantom scoring on the same scales. Kappa coefficient values indicate a good correlation between clinical and phantom/automated system scores.