"*Unless the authors can explain how JAFROC correctly defines the number of true negative findings in any image, the results are not valid. (The report appears to consider only "marked" locations."*

I thank the reviewer for the frankly expressed opinion. The issue of true-negatives is a frequent source of confusion in the context of free-response studies. The number of true negatives in such studies cannot be measured, and therefore the number of true negatives is not definable. This is because one does not know which normal regions were considered for marking but were not marked. Even the radiologist does not know, since unconscious decisions to not mark could have been made at some normal regions. This is well-known in the psychophysical literature and is embodied in a recent model of image perception (Kundel *et al.* 2007).

It is an obvious truism that any method for analyzing data can only deal with the data that is available. In the case of a FROC study these are the marked regions. Current methods for analyzing FROC data, namely JAFROC (Chakraborty *et al.* 2004), IDCA (Edwards *et al.* 2002), non-parametric (Samuelson *et al.* 2006), and search-model (Yoon *et al.* 2007) all use only the marked regions (and the unmarked lesions).

"*If so, why are locations that are not "marked" ignored?*"

Regions that are not marked are not ignored, but this is in a subtle sense, and I appreciate the reviewer's question. In JAFROC scoring, by default a radiologist receives maximum credit for not marking, but is debited for marking normal regions. The JAFROC figure of merit decreases with more marked normal regions, especially if these are rated high.

"*Wouldn't this mean that in a totally clear lung by all three methods there were no true negatives?"*

This gets us back to the issue, addressed above, that in a free-response study the number of true negatives is unknown and un-measurable. However, the absence of marks does not mean that there were no true negatives; rather it means that any decision on that image was a true negative. But the reviewer makes an interesting point. If a particular image is totally clear by chest tomosynthesis and chest radiography (MDCT was the reference) then for this image the radiologist receives the maximum credit, and based on this image alone there is no difference between the two modalities. If there was a mark in one modality, this image would contribute to a slight decrease in the JAFROC figure of merit for that modality.

"A secondary flaw in the analysis is that the interpretations of multiple locations within the same image are not independent. If the authors do chose to investigate sensitivity (see below), this interdependence ought to be taken into account."

This is an important issue and it has been repeatedly and justifiably raised against *pre-JAFROC* methods that have assumed independence for the ratings of marks on the same image. However, JAFROC does not assume independence of marks on the same image. In JAFROC analysis when a case is jackknifed *all* marks on the case are removed from the analysis and each case yields *one* pseudovalue. No assumption regarding the correlation structure is made. The approach is in the same spirit as the Rutter bootstrapping approach (Rutter 2000) to the analysis of ROI data.

JAFROC has been extensively validated using simulators that have included correlations of the ratings (Chakraborty *et al.* 2004; Chakraborty *et al.* 2008) on an image. The method has the expected NH behavior even in the presence of strong correlations. But I do not expect the reviewer to take my word for it. To quote a recent paper(Wagner *et al.* 2007) (emphasis added):

"*They have presented a solution to the FROC problem using a jackknife resampling approach that respects the correlation structure in the images.* They refer to their algorithm and software as JAFROC. For scorekeeping, they include all lesion ratings on abnormal images (unmarked lesions receive the lowest or default rating, and false-positives on lesion-containing images do not contribute); the summary figure of merit is defined as the probability that a lesion rating on an abnormal image exceeds all false-positive ratings on a normal image. *Under these conditions, their paradigm successfully passes a rigorous statistical validation test.* "

An eminent statistician, has written (Dodd *et al.* 2004) "*To accommodate the correlations within an image,* Chakraborty and Berbaum ... have suggested a jackknife approach to resampling in the same spirit as recent work of Rutter .... Extensive simulations have been conducted and on those trials the Jackknife AFROC approach preserves the power advantage of the earlier AFROC method while maintaining the appropriate rejection rate...."

The same data might be used to make essentially the same claims if the authors restricted their analysis to the comparison of the sensitivities of the two techniques and, for patients with and without nodules by MDCT, the number of false positive marks.
With all due respect this would be a step-backwards. I am not aware of any active work in this area that is considering this approach. Comparisons of sensitivities (or numbers of false positive marks) are susceptible to criterion shift and bias and their co-variation presents an obvious problem, all of which effects result in a statistical power penalty. As the reviewer is well-aware this is the rationale for ROC analysis where the AUC figure of merit takes into account the co-variation. The JAFROC figure of merit plays a similar role in FROC analysis – it is the area under the AFROC curve (Chakraborty 1989; Chakraborty *et al.* 1990) and takes into account the co-variation of sensitivity with the false positive rate.

References

D. P. Chakraborty, "Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16** (4), 561-568 (1989).

D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis and validation," Medical Physics **31** (8), 2313-2330 (2004).

D. P. Chakraborty and L. H. L. Winter, "Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment," Radiology **174**, 873-881 (1990).

D. P. Chakraborty and H. J. Yoon, "Investigation of methods for analyzing location specific observer performance data," Proc. SPIE Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment **6917** (2008).

L. E. Dodd, R. F. Wagner, S. G. r. Armato, M. F. McNitt-Gray, S. Beiden, H. P. Chan, D. Gur, G. McLennan, C. E. Metz, N. Petrick, B. Sahiner and J. Sayre, "Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium," Acad Radiol **11** (4), 462-475 (2004).

D. C. Edwards, M. A. Kupinski, C. E. Metz and R. M. Nishikawa, "Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model," Med Phys **29** (12), 2861-2870 (2002).

H. L. Kundel, C. F. Nodine, E. F. Conant and S. P. Weinstein, "Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study," Radiology **242** (2), 396-402 (2007).

C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data.," Acad. Radiol. **7** (6), 413-9 (2000).

F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," 2006 IEEE International Symposium on Biomedical Imaging: From Nano to Micro, 1312-1315 (2006).
R. F. Wagner, C. E. Metz and G. Campbell, "Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review," Academic Radiology **14** (6), 723-748 (2007).
H. J. Yoon, B. Zheng, B. Sahiner and D. P. Chakraborty, "Evaluating computer-aided detection algorithms," Medical Physics **34** (6), 2024-2038 (2007).