AbstractID: 9795 Title: Truth and Scoring Issues in Computer Aided Detection and Diagnosis

To assess the performance of computer-aided diagnosis (CAD) systems, the output of these systems needs to be compared to some form of knowledge that has traditionally been called "ground truth". This presentation will describe: (a) a description of methods that are currently being used to establish "ground truth", and (b) a description of methods to score or assess the performance of a CAD system with respect to that "ground truth". These issues will be illustrated with examples from breast cancer imaging using mammography and from lung cancer imaging using helical CT.

For each of these disease/modality combinations, there are two very different types of CAD problems with very different requirements for truth and very different scoring methods. The first is the detection problem that seeks to determine if there is *something* suspicious in the image. For this problem, the primary issue is the spatial localization of suspicious objects in the image(s). Truth is often established by expert readers or by a panel of expert readers. However, as recent presentations related to CT imaging of the lung have shown, obtaining consensus may not be easy and can be influenced by the exact instructions to the readers. The difference between "finding every nodule" and "finding every nodule suspicious for lung cancer" yields very different results due to radiologists' *interpretation* of the clinical significance of each nodule.

Scoring of detection is typically done based on a comparison of the CAD output to the spatial location of the nodules identified by the expert(s). A true positive may be identified with various criteria such as having the CAD system identify: (a) the centroid of the expert defined lesion; (b) at least 50% of expert defined lesion area, (c) a bounding box that fully encompasses the expert defined lesion, etc. However, as one may infer from the previous paragraph, the difficulty in establishing truth also influences how the scoring may be done.

For the diagnosis task, truth – in the form of clinical diagnosis - is often established through further tests such as: (a) biopsy, (b) surgical resection, or (c) lack of any growth observed on radiographic follow-up for a certain period of time. However, there have been discussions in the literature regarding the accuracy of each of these methods as well. For this task, scoring is typically done by comparing the CAD output to the results of these additional tests with a true positive occurring when the CAD output agrees with the clinical diagnosis. However, the output of the CAD system may not be a binary variable (benign/malignant) and this case will be discussed as well.

Educational Objectives:

1. To describe some commonly used methods for determining truth for CAD systems, both for detection and diagnosis tasks
2. To describe some of the scoring methods used for evaluating CAD system performance with respect to that truth
3. To describe some of the subtle issues that can influence both the determination of truth and scoring for both the diagnostic and detection tasks commonly being utilized in CAD.