Gold Standards and Expert Panels: A Pulmonary Nodule Case Study with Challenges and Solutions

Dave P. Miller*, Kathryn F. O'Shaughnessy **, Susan A.Wood **, Ronald A. Castellino **

* Ovation Research Group
** R2 Technology, Inc.

ABSTRACT

Introduction: Comparative evaluations of reader performance using different modalities, e.g. CT with computer-aided detection (CAD) vs. CT without CAD, generally require a "truth" definition based on a gold standard. There are many situations in which a true invariant gold standard is impractical or impossible to obtain. For instance, small pulmonary nodules are generally not assessed by biopsy or resection. In such cases, it is common to use a unanimous consensus or majority agreement from an expert panel as a reference standard for actionability in lieu of the unknown gold standard for disease. Nonetheless, there are three major concerns about expert panel reference standards: (1) actionability is not synonymous with disease (2) it may be possible to obtain different conclusions about which modality is better using different rules (e.g. majority vs. unanimous consensus), and (3) the variability associated with the panelists is not formally captured in the p-values or confidence intervals that are generally produced for estimating the extent to which one modality is superior to the other.

Methods: A multi-reader-multi-case (MRMC) receiver operating characteristic (ROC) study was performed using 90 cases, 15 readers, and a reference truth based on 3 experienced panelists. The primary analyses were conducted using a reference truth of unanimous consensus regarding actionability (3 out of 3 panelists). To assess the three concerns noted above: (1) additional data from the original radiology reports were compared to the panel (2) the complete analysis was repeated using different definitions of truth, and (3) bootstrap analyses were conducted in which new truth panels were constructed by picking 1, 2, or 3 panelists at random.

Conclusions: The definition of the reference truth affected the results for each modality (CT with CAD and CT without CAD) considered by itself, but the effects were similar, so the primary analysis comparing the modalities was robust to the choice of the reference truth.

Key Words: Panel, Reference Truth, Consensus, CAD, Bootstrap, MRMC ROC Study, Lung nodules

1. INTRODUCTION

To meaningfully measure the accuracy or performance effects of any diagnostic imaging modality, a second and more accurate "gold standard" modality is needed to provide ground truth. In the case of lung nodule detection, the "truth" is comprised of three questions: Is the identified finding a real abnormality? Is it really a nodule? And is it clinically significant?

In projection chest radiography, detected nodules are typically in the order of 1-3 cm in size. At this size range, CT serves as the gold standard for the first two questions, and biopsy the gold standard for the third.

For conventional (thick-slice, i.e. ≥5mm collimation) CT, where suspected lesions are usually in the range of 5mm or greater, thin-slice CT (i.e. ≤3mm collimation) provides the higher standard for addressing whether they are true abnormalities and truly nodules. Biopsy remains the "proof" of clinical significance for larger nodules (those 1cm or larger), while intermediate nodules (i.e. 5-10mm) remain indeterminate and must be followed temporally or, if suspicion is high enough, scheduled for open biopsy.

Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment,
edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.544716

In the case of thin-slice CT, though many more suspected nodules are being identified than on either projection chest radiography or conventional chest CT—with some as small as 2 or 3mm in size--there is no established gold standard for verifying that any of them are abnormal or that any are indeed nodules. And for those findings below the biopsy threshold, there remains no reliable standard for determining clinical significance other than waiting to see if they grow, remain stable, or regress over time.

In the absence of a viable gold standard, a common work-around employed in diagnostic imaging is to have two or more qualified radiologists provide a "consensus" interpretation of a set of images.

Though there are numerous ways of doing this, perhaps the most common is to have two radiologists read together and render a single opinion, or read independently and convene only to resolve areas of disagreement. To avoid a situation in which one reader's opinion (e.g. the more senior reader's) prevails a disproportionate number of times, a third reader may be asked to serve as arbiter, where needed.

Alternatively, a panel of three or more readers may be utilized, with each reader offering his or her independent judgment, and a collective decision as to "truth" being made at various levels of consensus (e.g. 2 of 3, 3 of 3, etc). This type of arrangement prevents one reader from dominating the "truth" determining process, mitigates the influence of any "outlier" that might be present, and allows for a higher level of "truth" when unanimity is achieved.

Though less definitive than a biopsy standard for determining clinical significance, a unanimous consensus panel truth can provide a reasonable alternative "truth" for assessing the accuracy of an imaging modality. And it has the advantage of doing so in a manner that best simulates the use of that modality at the actual "point of care" in clinical practice.

Without a unanimous consensus (e.g. 3 of 3) threshold for truth, however, determinations such as abnormal vs. normal, nodule vs. non-nodule, and actionable vs. non-actionable can be bedeviled by inter-observer variability and semantic disagreements. Furthermore, such determinations can overlap significantly such that it can be difficult to separate one from the other without explicit and uniform instructions to each reader.

A major weakness of panel truths was identified by Revesz, Kundel, and Bonitatibus[1]. They examined three competing imaging modalities and evaluated them against several different reference truths including a majority-rule decision panel truth, a unanimous consensus panel truth, and an expert-review panel truth. In doing so, they uncovered the disturbing result that almost any of the modalities could be shown to be statistically significantly better than the other if measured against the most favorable reference truth.

This result is discussed further in a recent report by Dodd and Wagner[2] from the Lung Image Database Consortium (LIDC). They acknowledge that verification bias makes it difficult, if not impossible, to depend on a clinical outcome alone, as it is not ethical to submit patients who screen negative to further radiographic follow-up, much less biopsy. On the other hand, they state unequivocally that panel truths "will inevitably introduce additional uncertainty or noise".

In the absence of biopsy truth, one may wish to choose a reference truth from a set of possible panel truths based on which one has the greatest face validity when compared to whatever limited clinical data are available. Alternatively, a simple solution to the problem identified by Revesz *et al.* is to always evaluate competing modalities against several alternative reference truths. If one technology is better than the other with respect to a range of logical reference truths, then one would like to assume that it is truly better. Nonetheless, this does not capture the additional uncertainty. That is, evaluating multiple reference truths provides a more robust answer with regard to the directionality of the modality effect, but it may still yield an incorrect confidence interval for any given comparison because the potentially random effect of the panelists selected to participate in the panel is not captured.

In short, comparing two modalities based on a panel-based reference truth is a dicey business, requiring a wealth of extra validation analysis if the study is to be considered compelling. In spite of these hurdles, the panel approach is the only viable option for demonstrating improved reader performance for some new technologies. For example, a new computer-aided detection (CAD) algorithm for detection of lung nodules with thin-slice MDCT (ImageChecker CT, R2 Technology, Sunnyvale, CA) was evaluated by conducting a multi-reader multi-case (MRMC) receiver operating characteristic (ROC) study using a consensus panel as the reference truth.

This clinical study provided an opportunity to examine potential or stated flaws in the panel-based approach for establishing a reference truth. Can the results of a study based on a panel truth be shown to be robust after (a) considering the available clinical data, (b) re-evaluating the results against a range of fixed reference truths, and (c) re-computing confidence intervals after adding in the random noise attributable to panel variability? If so, the common wisdom about panel truths is due for another correction. Prior to the Revesz *et al* publication, most researchers saw little reason to doubt the panel truth method. After the publication, researchers saw every reason to doubt the method. The truth, almost certainly, is in the middle. With appropriate attention to robustness analysis and appropriate study design choices, truth panels are a viable research method.

2. DEFINITIONS

The following terms, some of which have already introduced, will be used throughout this paper.

Gold standard – an objective invariant standard of truth, generally used with objective constructs such as malignancy
Actionability – a subjective judgment by a physician indicating that a case warrants further follow up or intervention
Reference truth – a standard for truth that falls short of being a gold standard
Panel – two or more physicians assigned the task of rendering a subjective opinion about a set of cases
Panel truth – a reference truth based on a panel
Nodule-present – a case or region of interest with at least one actionable nodule based on the panel truth
Nodule-absent – a case or region of interest that is not defined as nodule-present
Unanimous Consensus – a panel truth requiring that all panelists agree that an actionable nodule is present
Majority – a panel truth requiring that a majority of panelists agree that an actionable nodule is present
Minority – a panel truth requiring that only a minority of panelists agree that an actionable nodule is present
ROC - Receiver Operating Characteristic
Reader – a radiologist whose rating may be compared against a reference truth to compute an ROC curve
Case – a patient exam
Quadrant – the upper or lower portion of the left or right lung
Modality – a method of reading cases (e.g. CT with CAD and CT without CAD)
MRMC – mutli-reader multi-case, describes a study design in which many readers read many cases
ANOVA – Analysis of variance

3. CLINICAL STUDY BACKGROUND

Five (5) regionally diverse sites contributed 90 cases to the study; 2 sites in the Northeast, and 1 site each from the South, the Midwest, and the West. Of these sites, 3 were private imaging centers and 2 were academic medical centers. All centers contributed cases identified on the site radiology report as nodule-present (41) and nodule-absent (49).

All cases were culled consecutively from the sites' digital archives according to the inclusion and exclusion criteria identified in a case collection protocol. The consecutive set of nodule-present cases covered a greater period than the consecutive set of nodule-absent cases so that a roughly equal sample of nodule-present and nodule-absent cases would be available for analysis.

The nodule-present cases collected included only those in which a diagnosis of cancer, either primary lung cancer or an extrathoracic neoplasm, had been documented. Other co-existing disease processes resulting in the formation of nodules (e.g. TB, histoplasmosis, rheumatoid lung) were allowed, as were cases containing other "background" pathology such as lobar pneumonia, emphysema, and heart failure. The radiology reports are insufficiently specific to localize all suspect nodules with certainty, so cases, rather than regions of interest, were designated as nodule-present or nodule-absent during the case collection.

The nodule-absent cases collected were those in which no nodules were deemed to be present by the principal investigator at each site. Other disease processes could be present, including the presence of masses (>3cm). Histories of cancer, radiation therapy, or even previous thoracotomy, were allowed.

Although the nodule-present cases were collected in a way that maximized their chance of being clinically important, biopsy proof of lung cancer on specific nodules was rarely available. Additionally, the nodule-absent cases were identified as such based on a single exam by a single radiologist. Peldschus et al[3] demonstrated that a second review of such cases often contain overlooked nodules.

Thus, to create a reference truth, all 90 cases were reviewed by a panel of 3 experienced radiologists to localize and adjudicate the actionability of all nodular abnormalities they identified in the case. Although the 3 panelists reviewed the cases independently, all findings that were identified by 1 or 2 panelists (but not all 3) were automatically localized and presented to all 3 panelists for a second-pass adjudication of whether or not the finding was an actionable nodule.

The 90 cases were reviewed over the course of 8 different sessions, each of which involved exactly 3 panel radiologists independently reading the cases. A total of eleven different panelists participated in 1, 2, or 3 sessions.

The panels had the challenge of simultaneously making the decisions 1) that a nodule was present and, if so (2) that the detected nodule was actionable. Because nodules, as opposed to a tissue diagnosis of cancer, are a subjective construct, a second panel of five radiologists was asked to review all of the unanimous actionable findings to identify those that fit the "classic" text book description of a lung nodule. This subset of the unanimous nodules were categorized as "classic" nodules.

As a result of this process, nodules could be classified into five categories:

1. Classic unanimous nodules (ie, those detected by at least one panelist on the first pass and determined to be an actionable nodule by all 3 on the second pass, and judged to be a "classic" nodule by a super-majority of the 5-member expert panel).
2. Non-classic unanimous nodules (detected by at least one panelist on the first pass and determined to be an actionable nodule by all 3 on the second pass, and not judged to be a "classic" nodule by a super-majority of the 5-member expert panel).
3. Majority actionable nodules (detected by at least one panelist on the first pass and rated actionable by exactly 2 out of 3 on the second pass).
4. Minority actionable nodules (detected by at least one panelist on the first pass and rated actionable by exactly 1 out of 3 on the second pass.
5. All other panel findings (includes findings rated as non-actionable by 3 out of 3 panelists and findings that did not fall within the size or density criteria specified in the protocol).

Similarly, cases or quadrants within cases could be classified as a unanimous nodule-present case (or quadrant) if they contained at least one unanimous actionable nodule or a majority nodule-present case if they contained at least one majority actionable nodule.

For the truth panel, radiologists were specifically instructed to take extra time reading the cases to try to identify every possible nodule. Additionally, they were given the opportunity in a second pass to rate the findings identified by another reader in the first pass as regards whether the finding represented a nodule and, if so, whether it was "actionable". The additional 5-member expert panel was similar to the second pass of the 3-member panels in that radiologists were only asked to perform the classification task, not the detection task.

Finally, a new set of 15 radiologists, who participated in none of the prior panels, was enlisted to evaluate cases reading at a usual pace, consistent with their typical clinical practice, for an ROC study. These 15 radiologists will be referred to as readers rather than panelists. The 90 cases were divided into 4 quadrants according to a region of interest (ROI) approach[4], yielding a total of 360 ROI for evaluation. Each of 15 radiologists sequentially reviewed the 360 quadrants , first without computer-aided detection (CAD) and then with CAD. Using a 0-100 scale, they were asked to evaluate their confidence that the quadrant did or did not contain an actionable nodule. The ratings were provided before and after CAD marks were placed on the case. For a given reference truth, an area under the receiver operator curve (A_z) could then be computed for each of the 15 radiologists before and after CAD.

4. STATISTICAL METHODS

For each of the 15 radiologists, A_{ZB} (the area under the curve before seeing the CAD output), A_{ZA} (the area under the curve after seeing the CAD output), and $A_{Z\Delta}$ was computed. All A_Z were computed by the trapezoidal rule (also referred to as the empirical curve as discussed by Hanley and McNeil[5].

The primary analysis for the clinical study was based on the Dorfman-Berbaum-Metz (DBM) ANOVA-after-jackknife approach[6]. Using this approach, 90 leave-one-out samples were generated, each time leaving out one case (i.e., leaving out all 4 quadrants in the case), and the A_Z values computed from the leave-one-out samples were used to compute the A_Z pseudo-values that form the ANOVA-after-jackknife analysis dataset. The primary conclusions of the study are based on the model proposed by DBM that includes modality as a fixed effect, case and reader as random effects, and all three of the possible two-way interaction terms as random effects. Unanimous consensus is the reference truth used for the primary analysis.

As an alternative to the semi-parametric DBM approach, the fully non-parametric bootstrap method has been proposed by Rutter[7] as being especially helpful when there are multiple ROI per case. Every reader in a given bootstrap sample reads every case in the bootstrap sample -- in each of the two competing modalities (reading without CAD and reading with CAD). The confidence intervals based on the bootstrap samples should approximate the parametric confidence interval based on the ANOVA-after-jackknife model. For each of 1000 bootstrap samples, the 2.5th and 97.5th percentile is computed for all measures of interest (e.g. $A_{Z\Delta}$) and these percentiles form the approximate 95% confidence interval. Additionally, the proportion of samples for which the measure of interest is less than zero is computed. Multiplying this proportion by 2 yields the approximate two-sided p-value.

In addition to sampling cases and readers at random, it is possible to select a random reference truth by comparing the ratings of the ROC readers to a single panelist. A random truth of this type is generated such that each case may have a different random panelist, but all 15 ROC readers are judged against the same randomly generated reference truth for a given bootstrap sample.

It is easy to extend this concept to randomly select a pair of readers (with replacement) or a random 3-member panel (necessarily with replacement, since only one fixed 3-member panel is available).

Both the DBM approach and the bootstrap approach use the A_Z values as their starting point, and computing these values requires a reference truth for each ROI in each case. What may not be obvious is that the same set of ratings gathered from the same set of readers may be evaluated against any number of reference truths. That is, the ratings are made completely independently from the panel process, so the statistical evaluation of the experiment may be repeated many times (in fact thousands of times in the bootstrap approach) with a variety of different reference truths without having to actually repeat any portion of the clinical experiment.

In a similar manner to that noted previously, quadrants may also be placed into several buckets depending on the presence or absence of panel findings in those quadrants.

1. Quadrants containing at least one classic unanimous (3/3) actionable nodule.
2. Quadrants containing at least one non-classic unanimous (3/3) actionable nodule.
3. Quadrants containing at least one majority (2/3) actionable nodule that do not contain any unanimous (3/3) actionable nodules.
4. Quadrants containing at least one minority (1/3) actionable nodule that do not contain any unanimous (3/3) or majority (2/3) actionable nodules.
5. Quadrants containing at least one non-actionable panel finding that do not contain any findings rated as actionable by any panelist.
6. Quadrants that do not contain any panel findings.

For the primary reference truth, quadrants in categories (1) and (2) are considered “positive” or nodule-present and quadrants in categories (3)-(6) are considered “negative” or nodule-absent. Based on this definition, there were 75 nodule-present quadrants in the study vs. 285 nodule-absent quadrants.

5. RESULTS

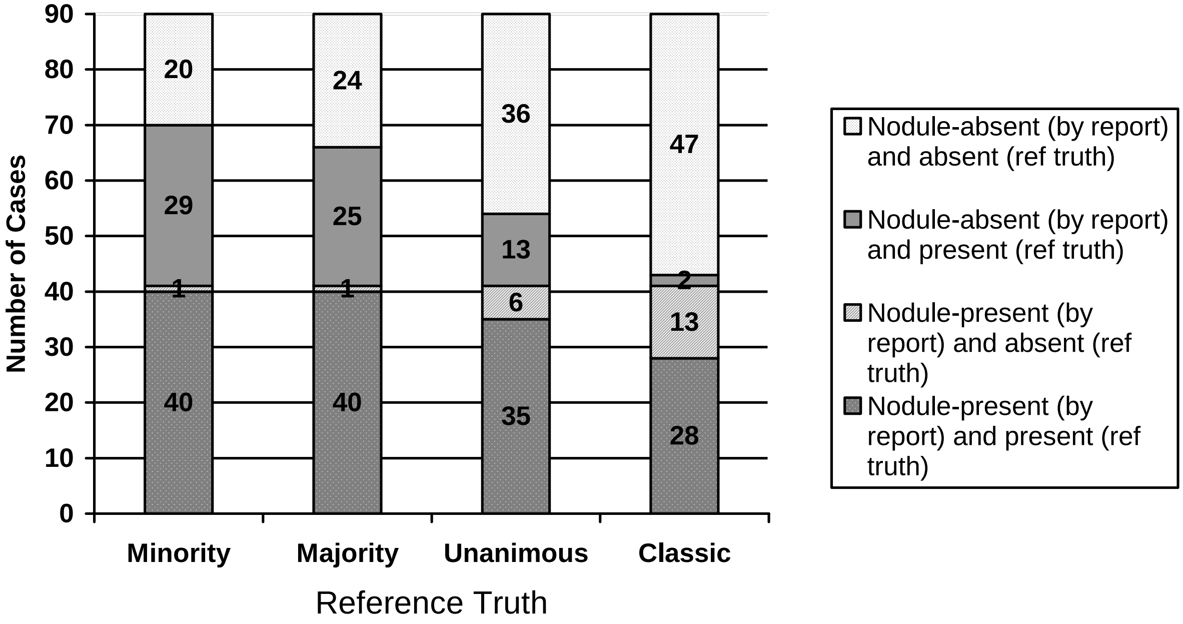
There are three sets of results. The first is a comparison of the reference truths to the available clinical data. The second is a repetition of the primary ROC analysis using the DBM methodology for different reference truths. The third is an analysis using bootstrap methodology and a random reference truth.

5.1 Clinical data

Each of the different possibilities for fixed (non-random) reference truths was compared to the data from the clinical sites that contributed the cases. Obtaining an exact match of consensus nodules with the images was beyond the scope of this study, so the comparison was done at the case level. For every truth considered, all cases (and all quadrants) are counted as either nodule-present or nodule-absent. For instance, when the consensus reference truth is used, this means that a case is nodule-present if a consensus actionable nodule is present, but it is counted as nodule-absent even if two-thirds or one-third of the panelists believe the case is actionable. This method of establishing a strict binary truth for every case, without making any exclusions, avoids the common problem of biasing the results by studying the sickest of the sick and the wellest of the well.

Of the 90 cases, 41 were collected as nodule-present cases and 49 were collected as nodule-absent cases. However, using a reference truth of unanimous consensus (3 out of 3 panelists agree the case has at least one actionable nodule), 75 quadrants from 48 cases are adjudicated as nodule-present. This increase from 41 to 48 nodule-present cases occurred because 14% of the cases that were collected from the sites as nodule-present (6/41) are counted instead as nodule-absent in the ROC analysis, and 27% of the cases that were collected from the sites as nodule-absent (13/49) are counted instead as nodule-present in the ROC analysis. Therefore, for this sample of 90 cases, the unanimous reference truth represents a 17% increase in nodule-present cases compared to the classifications on the original radiology clinical reports.

Figure 1: Site Radiology Report versus Reference Truths



In Figure 1, the uppermost and lowermost shaded regions represent the number of cases for which the original radiology report and the reference truth are concordant, defining the case as nodule-absent or nodule-present respectively. The middle two shaded regions represent discordance between the report and the reference truth.

If the unanimous reference truth is replaced with a majority (2 out of 3) reference truth, almost all of the cases collected as nodule-present (40/41, 98%) are counted as nodule-present in the ROC analysis, and more than half of the cases collected as nodule-absent (25/49, 51%) are also counted as nodule-present in the ROC analysis. Using this “majority” reference truth, 44 quadrants that would be counted as nodule-absent based on the consensus standard are counted as nodule-present, so a total of 119 (75+44) quadrants are considered nodule-present by the majority reference truth. An additional 24 quadrants would be counted as nodule-present if the loosest possible threshold (at least one panelist out of 3 rates actionable) were used for the reference truth. This loose panel truth will be referred to as a minority reference truth to indicate that a quadrant is defined as nodule-present even if only a minority of panelists indicate an actionable nodule is present.

If the unanimous reference truth is replaced with a tighter truth based on the classic nodule definition from the second 5-member expert panel, 68% (28/41) of the cases collected as nodule-present are counted as nodule-present in the ROC analysis and 4% (2/49) of the cases collected as nodule-absent are counted as nodule-present in the ROC analysis. Using this tighter truth, only 38 quadrants are classified as nodule-present.

5.2 DBM analysis for different fixed reference truths

For each of four candidate reference truths, an ROC analysis was conducted using DBM’s ANOVA-after-jackknife approach. These four reference truths (minority, majority, consensus, and classic) classify 143, 119, 75, and 38 quadrants as nodule-present respectively, as detailed in the previous section. The “classic” reference truth is approximately twice as tight as the unanimous reference truth and the minority reference truth is approximately twice as loose as the unanimous reference truth, so the four reference truths cover a broad spectrum of potentially aggressive or conservative definitions of the subjective classification of actionability.

Figure 2: Upper Left Corner of Average-Reader ROC Curves for Four Different Reference Truths

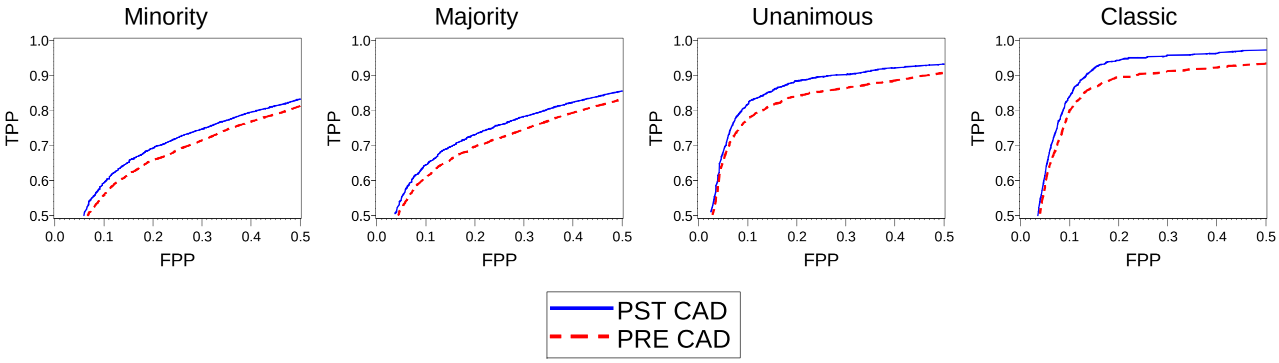


Figure 2 above shows the pre-CAD versus post-CAD reader performance assessed according to each of four reference truths; minority, majority, consensus, and the “classic” nodule reference truths. Three trends should be noted in these graphs. First, both the pre-CAD and post-CAD performance improve as the definition of the reference truth becomes tighter. Second, the post-CAD performance is better than the pre-CAD performance in every case. Third, the gap between the pre-CAD and post-CAD performance widens with the tighter definitions.

These three observations about the graphs are further supported by the results from the DBM analysis, shown in Table 1 below.

Table 1: ANOVA-after-jackknife results for four different reference truths

Reference Truth	Average Reader Pre-CAD Az	Average Reader Post-CAD Az	Average Reader Improvement With CAD	95% CI	p-value
Minority (1/3)	0.7811	0.8013	0.0202	(.0097, .0306)	0.0003
Majority (2/3)	0.8055	0.8269	0.0213	(.0097, .0329)	0.0006
Unanimous (3/3)	0.8805	0.9046	0.0240	(.0084, .0395)	0.0033
“Classic” nodule	0.8987	0.9299	0.0313	(.0112, .0513)	0.0040

The most noteworthy statistical artifact in this analysis is that the confidence bounds using the two-thirds majority reference truth are actually tighter than the confidence bounds using the unanimous three-thirds reference truth. This is a sample size issue. Using the unanimous reference truth, the sample is less evenly divided into nodule-present and nodule-absent quadrants than it is in the analysis using the majority reference truth. Including the minority (one-out-of-three panelists rate actionable) quadrants as nodule-present quadrants splits the sample even more evenly. This sample size effect will be important to recall in the later discussion of the random truths.

5.3 Bootstrap analysis of random truths

The bootstrap is a computationally-intensive method for using the data from a single experiment to simulate the universe of all possible similar experiments. This is done by randomly sampling observations (with replacement) from the original study dataset. Unlike the ANOVA-after-jackknife, which is partly parametric and partly non-parametric, the bootstrap is fully non-parametric, so it involves fewer assumptions.

The bootstrap method of computing confidence intervals is much easier to generalize to incorporate variability in the truth. As a first step, though, and to ground the later bootstrap analyses, bootstrap confidence intervals are computed with the same fixed reference truth used for the primary ANOVA-after-jackknife analysis with the consensus truth. Table 2 below shows estimates of the same quantities using the ANOVA-after-jackknife and the bootstrap.

Table 2: ANOVA-after-jackknife compared to the Bootstrap using the same reference truth

Analysis Using Consensus Truth	Estimated Improvement in Az	95% CI	p-value
ANOVA (DBM)	.0240	(.0084, .0395)	.003
Bootstrap (non-random panel)	.0246	(.0089, .0446)	<.001

The two methods of computing the 95% confidence intervals are very similar; however, the bootstrap confidence intervals are slightly broader on the higher end. Both methods shown in the table above use a fixed 3-out-of-3 unanimous consensus as the reference truth.

A random 3-member panel could also be obtained by sampling with replacement. For this method, each of the reference truths based on the one-member panel truths have a probability of 1/27, each of the reference truths based on the 3 possible two-member panels have a probability of 6/27, and the reference truth from the original 3-member has a probability of 6/27.

A 2-member random truth could be also be assigned for each case. Unlike the 3-member panels, the random 2-member panels may be assigned with or without replacement; however, sampling with replacement is most true to the bootstrapping framework. Sampling with replacement, there is a 1/9 probability assigned to the reference truths based on each of the 3 possible one-member panels and a 2/9 probability assigned to the reference truth based on each of the 3 possible two-member panels.

Finally, we also considered a random single panelist. This is probably the most straightforward way of obtaining a random truth, but it yields a reference truth that is difficult to defend as being any better than the original radiology report.

For all three of these random truths, the panelists randomly selected were assumed to be the panelists in both pass 1 and pass 2 of the consensus panel. Therefore, to be included as a nodule-present quadrant, the finding had to be identified by at least one of the randomly chosen panelists in the first pass and all of the randomly chosen panelists (same random set in pass 1 and pass 2) had to agree the finding was actionable in pass 2.

The results of the three random panels are contrasted with the results for the non-random consensus panel results in Table 3 below.

Table 3: Comparative estimates and confidence intervals using fixed (non-random) and random reference truths

Bootstrap reference truth	Average Reader Pre-CAD Az	Average Reader Post-CAD Az	Estimated Improvement in Az	95% CI	p-value
Fixed (non-random) 3-panelist unanimous consensus	.8786	.9031	.0246	(.0089, .0446)	<.001
Random 3-panelist unanimous consensus	.8454	.8678	.0224	(.0076, .0403)	<.001
Random 2-panelist unanimous consensus	.8324	.8540	.0216	(.0077, .0387)	.002
Random single panelist	.8170	.8380	.0209	(.0080, .0370)	<.001
Fixed (non-random) 3-panelist majority	.8032	.8248	.0216	(.0100, .0363)	<.001
Random 3-panelist majority	.8089	.8302	.0213	(.0087, .0363)	<.001

There is much that is noteworthy about these results, but three particular findings stand out. The first, and most important, is that the results are robust to the variation in the reference truth in that all of the variations yielded a statistically significant increase in Az. The second is that the confidence bands are not substantially broadened when this source of variability is accounted for. The third is that the estimated improvement in Az for the random 3-panelist consensus falls between the estimated improvement from the fixed 3-panelist consensus and the fixed 3-panelist majority. Each of these findings will be discussed further.

6. DISCUSSION

6.1 Clinical Data

A 3-panelist consensus of actionability is quite distant from biopsy proof of cancer in the extended family of all possible reference standards, gold or otherwise. Some would argue that consensus actionability is a poor substitute for documented malignancy; however, such an argument ignores the realities of clinical practice. Radiologists are not generally in a position of making a definitive diagnosis based on a single CT examination. Instead, their role is to first determine whether the image contains a finding that is not normally present in that set of images (detection); and, then to determine its potential clinical significance (interpretation) and suggest if further assessment is necessary (recommendations). It is well known that most lung nodules that are detected and followed for interval growth are not cancer[8,9], but the risk of not having at least a minimal follow-up (e.g., repeat CT in 3-6 months) is generally considered to be much greater than the risks associated with having that follow-up. Thus, the class of actionable nodules is much greater than the class of malignant nodules, and it is a subjective construct of current medical practice.

The decision to collect nodule-present cases from cancer patients provides a suggestion, in the absence of documented assurance, that the nodules detected in them may be clinically important. Using the unanimous consensus reference truth, a small number of these cases were adjudicated as containing no actionable nodules (nodule-absent case), and a somewhat larger number of cases that were collected as nodule-absent cases were reclassified as nodule-present cases. A truth of this sort is consistent with the basic gestalt that detection errors and overcalling of nodules both occur, but that the former occurs more frequently, and with more serious consequences, than the latter.

To some extent, selecting criteria for the best reference truth would depend on one's concern about detection errors (sensitivity) compared to overcalling (specificity). Thus, two-thirds majority would be a better reference truth if one believed that overcalling in clinical practice is rare and unimportant, and that detection errors are more common and of greater clinical significance. On the other hand, if one believed that overcalling is more common than detection errors, the "classic" nodule standard would be the most appropriate for judging a CAD algorithm. In fact, it is likely that the appropriateness of the different reference truths would depend on the different operating point and preferences of the user. This issue has been previously addressed in screening mammography by offering multiple operating points to the user.

Depending on what one expects to achieve with a new technology, any of these reference truths may be appropriate, but the unanimous consensus reference truth appears to come closest to current clinical practice. The difference between the site radiology reports and the majority reference truth merits further investigation in future studies, or it may simply reflect a greater tendency to recommend follow up exams when the full patient history and past exams are not available to assist in differential diagnosis or when there is no clinical consequence of the decision.

6.2 DBM analysis for different fixed reference truths

In this study, regardless of which reference truth was used, reading with CAD was always the superior modality compared to reading without CAD; however, the magnitude of the modality effect differed. Recall that every analysis was based on a binary truth, in which every quadrant was either nodule-present or nodule-absent. A more textured definition would involve classifying each quadrant as positive, equivocal, or negative.

If both minority and majority actionable nodules are considered equivocal, there is nearly the same number of equivocal quadrants as there are positive quadrants, but negative quadrants are far and away the most common. When the equivocal quadrants are included with the negative quadrants, they dilute the assessment of specificity, but the truly negative quadrants greatly outnumber the equivocal quadrants so the dilution is not overwhelming. That is, there is little doubt that most of the nodule-absent quadrants really are nodule-absent, so the choice of the reference truth does not substantially affect the specificity. When the equivocal quadrants are included with the positive quadrants, they dilute the assessment of sensitivity. Because there are an almost equal number of positive and equivocal quadrants, this dilution degrades the Az values due to the loss of sensitivity associated with readers providing low ratings to equivocal quadrants.

However, even though the equivocal quadrants dilute the magnitude of the estimated modality (CAD vs. no CAD) effect, they more evenly split the sample into nodule-present and nodule-absent quadrants. As a result, the slightly smaller modality effect is estimated with greater precision (i.e., lower variance) increasing the statistical power for demonstrating that reading with CAD is superior to reading without CAD.

6.3 Bootstrap analysis of random truths

Perhaps due to its complexity and perhaps due to the lack of deliberate design, it may have been understated that the three-out-of-three unanimous consensus reference truth actually involved the participation of 11 panelists over the course of 8 panel sessions. A somewhat surprising result of the ANOVA analysis (not shown) is that the case effect is very large and the reader effect is fairly small. This may be because the case effect already incorporates the bulk of the random variability associated with the imprecise reference truth. Further, each ROC reader is not compared to a small set of 3 panelists (with whom they might be more likely or less likely to agree simply due to the random selection of a small set of panelists), but rather they are compared to a set of 11 panelists (with whom there are undoubtedly some that share their reading and diagnostic habits and others that do not).

Lastly, within the discussion of the variable reference truths, there is the matter of the confidence intervals. Which confidence intervals are most appropriate? As a guide, we may wish to consider the pre-CAD A_Z. For the random 3-member panel, random 2-member panel, or single random panelists, the pre-CAD A_Z values are .8454, .8324, and .8170 respectively. By comparison, the pre-CAD A_Z based on the fixed reference truth of unanimous consensus is .8786 and the pre-CAD A_Z based on the fixed reference truth of a two-thirds majority is .8032. That is, without considering any modality effect, all three of the methods for addressing the truth variability using a random panel may be estimating a quantity that lies between the fixed-panel unanimous consensus reference truth and the fixed-panel majority reference truth.

On the other hand, the estimates based on a majority reference truth for the random 3-member panel are very similar to those for the fixed 3-member panel majority reference truth. In this case, the confidence band is slightly broader for the random reference truth than the fixed reference truth, but this is only true of the majority reference truths. For the unanimous consensus reference truth, the attempt to model the variation due to the panelists may bias the estimated effect size towards zero. In spite of this bias, the method provides a way of demonstrating that the overall ROC result is robust to the variability in panels. Because the bias favors the null hypothesis of no difference in modalities, a result favoring the alternative (one modality better than the other) is not a result of bias and in fact was in spite of the bias.

7.CONCLUSION

Previous research has demonstrated that results based on consensus panels may be flawed. This case study demonstrates that reference truths based on panels may also yield robust results. More importantly, a three-prong approach is provided for determining whether or not a panel-based result is robust using (1) clinical data (2) a range of definitions of reference truths, and (3) bootstrap samples of random truths.

Panels are ideal for evaluating subjective reader decisions. While gold standards may exist for cancer diagnosis (e.g., is this cancer?), gold standards do not always exist for detection-oriented questions (e.g., does this case have a nodule? is this nodule actionable?). Clinical practice, as indicated by site radiology reports, can provide a helpful metric for how well a panelist, or a consensus of panelists, in the research setting, mimics performance in the clinical setting.

Simple variation of the underlying reference truth can be accomplished with minimal effort compared to the effort of conducting an experiment. Any study for which results are based on a single panel-based reference truth should be viewed with some skepticism, and one may reasonably ask if the most favorable reference truth was cherry-picked from the set of possible reference truths. In the clinical study described here, the results were robust against the variations in reference truths.

Finally, bootstrapping provides a method for adding panel variability into the model, while still remaining rooted in the well-established DBM random-effect paradigm. In this study, all of the fixed and random truths produced a common result with respect to both the direction of the effect and the statistical significance of the effect. It is not clear what a study conclusion would be if analysis variations demonstrate a common directionality, but a mix of marginally significant and marginally non-significant results. If we require that multiple p-values fall below .05, the alpha level (type I error) for the experiment is much lower than .05. It may be that a lower, more challenging, overall alpha is simply the price one must pay to convince ones peers that a consensus panel result is robust.

8. REFERENCES

1. Revesz G, Kundel HL, Bonitatibus M. “The effect of verification on the assessment of imaging techniques.” Invest. Radiol., **18**, 194-198, 1983.

2. Dodd L, Wagner RF, Armato SG, McNitt-Gray MF, Beiden S, Chan H, Gur D, McLennan G, Metz CE, Petrick N, Sahiner B, Sayre J, and the LIDC Research Group “Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in CT: Contemporary research topics relevant to the Lung Image Database Consortium.” submitted to Academic Radiology, 2004.

3. Peldschus K, Herzog P, Wood SA, Cheema JI, Costello P, Schoepf UJ. “Computer Aided Diagnosis as a Second Reader: Spectrum of Findings in Computed Tomography Studies of the Chest Interpreted as Normal.” Radiology, **229(S)**, 291, 2003.

4. Obuchowski NA, Lieber ML, Powell KA. “Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography.” Acad Radiol, **7**, 516-525, 2000.

5. Hanley, JA, McNeil, BJ. “The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve.” Radiology, **143**, 29-36, 1982.

6. Dorfman DD, Berbaum KS, Metz, CE. “Receiver Operating Characteristic Rating Analysis.” Invest Radiol, **27**, 723-731, 1992.

7. Rutter, C. “Bootstrap Estimation of Diagnostic Accuracy with Patient-clustered Data.” Acad Radiol, **7**, 413-419, 2000.

8. Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby D, Pasmantier M, Koizumi J, Altorki N, Smith JP. “Early lung cancer action project: a summary of the findings on baseline screening.” Oncologist., **6(2)**, 147-52, 2001.

9. Swensen SJ, Jett JR, Hartman TE, Midthun DE, Sloan JA, Sykes AM, Aughenbaugh GL, Clemens MA. “Lung cancer screening with CT: Mayo Clinic experience.” Radiology, **226(3)**, 756-61, 2003.

10. Metz CE “Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies.” Invest Radiology, **24**, 234-245, 1989.