LungCAD: A Clinically Approved, Machine Learning System for Lung Cancer Detection

R Bharat Rao, Jinbo Bi, Glenn Fung, Marcos Salganicoff
Siemens Medical Solutions
51 Valley Stream Parkway, Malvern, PA 19355

Nancy Obuchowski
Quantitative Health Sciences
The Cleveland Clinic Foundation
9500 Euclid Ave., Cleveland, OH 44195

David Naidich
Department of Radiology
New York University Medical Center
400 East 34 Street, New York, NY 10016

ABSTRACT

We present LungCAD, a computer aided diagnosis (CAD) system that employs a classification algorithm for detecting solid pulmonary nodules from CT thorax studies. We briefly describe some of the machine learning techniques developed to overcome the real world challenges in this medical domain. The most significant hurdle in transitioning from a machine learning research prototype that performs well on an in-house dataset into a clinically deployable system, is the requirement that the CAD system be tested in a clinical trial. We describe the clinical trial in which LungCAD was tested: a large scale multi-reader, multi-case (MRMC) retrospective observational study to evaluate the effect of CAD in clinical practice for detecting solid pulmonary nodules from CT thorax studies. The clinical trial demonstrates that *every* radiologist that participated in the trial had a significantly greater accuracy with LungCAD, both for detecting nodules and identifying potentially actionable nodules; this, along with other findings from the trial, has resulted in FDA approval for LungCAD in late 2006.

Categories and Subject Descriptors

I.5.m [**Pattern Recognition**]: Miscellaneous

General Terms

Algorithms

Keywords

computer aided detection, lung cancer prognosis, classification, clinical trial

1. INTRODUCTION

Lung cancer is the most commonly diagnosed cancer worldwide, accounting for 1.2 million new cases annually. Lung cancer is an exceptionally deadly disease: 6 out of 10 people will die within one year of being diagnosed. The expected 5-year survival rate for all patients with a diagnosis of lung cancer is only 15%, compared to 65% for colon, 89% for breast and 99.9% for prostate cancer. In the United States, lung cancer is the leading cause of cancer death for both men and women, causing more deaths than the next three most common cancers combined, and costs $9.6 Billion to treat annually. However, lung cancer prognosis varies greatly depending on how early the disease is diagnosed; as with all cancers, *early detection* provides the best prognosis. At one extreme are the patients diagnosed with metastatic tumors (that have spread far from the lung), for whom the 5-year survival rate is just 2%. On the other hand, when diagnosed at an early stage, when the disease is still localized within the lung, the 5-year survival rate is 49%, and many treatment options (surgery, radiotherapy, chemotherapy) are viable. Today, only 24% of lung cancer cases are diagnosed at an early stage. [1, 10].

The recent development of multidetector computed tomography (MDCT) scanners has made it feasible to detect lung cancer at very early stages in priciple. Despite these advances in technology, many potentially clinically significant lesions still remain undetected [13]. One contributing factor is the explosion of generated data: The state-of-the-art 64-slice dual-source CT acquires up to 3,687 axial images in 30 seconds for each patient (each image must then be carefully examined by a radiologist). There is a growing consensus among clinical experts that the use of computer-aided diagnosis (CAD) software when used as a second reader (i.e., in conjunction with the radiologist) not only offers the potential to improve the detection accuracy of a radiologist, but also to reduce mistakes related to misinterpretation [2, 11]. In order for a CAD system to be used in clinical practice in the United States, it must first receive approval from the the Food and Drug Administration (FDA). All CAD systems must go through a rigorous clinical trial to receive approval (in much the same way as a new drug). A handful of CAD systems have received approval for detecting breast cancer lesions in the past 8 years. To be approved CAD systems must show satisfactory performance in two areas. The principal value of CAD is determined *not* by its stand-alone performance, but rather by carefully measuring the *incremental value* of using Computer-Aided Diagnosis in normal clinical practice with the radiologist in-the-loop. Secondly, CAD systems must not have a negative impact on patient

KDD'07, August 12-15, 2007, San Jose, California, USA.

management (for instance, false positives which cause the radiologist to recommend unnecessary, and potentially dangerous, follow-ups). Additionally, designing a trial for lung cancer detection is considerably more challenging than for breast cancer. One factor is the relative difficulty in obtaining *ground truth (correct labeling) for lung cancer related lesions*. Whereas, in breast cancer virtually all suspicious lesions are routinely biopsied (providing definitive histological ground truth), a lung biopsy is a dangerous procedure, with a 2% risk of serious complications (including death); this makes obtaining definitive ground truth infeasible, particularly for patients being evaluated for early signs of lung cancer.

Section 2 describes some of the machine learning challenges involved in learning a classifier for detecting lung cancer. We review some of our previous solutions. Section 3 describes the clinical trial design for our LungCAD system, which includes a fairly complex mechanism for determining ground truth and measuring incremental improvement. Section 4 summarizes the experimental results of the clinical trial that has resulted in granting clinical approval for LungCAD. We conclude in Section 5 with some discussion about CAD in general and future challenges.

2. MACHINE LEARNING CHALLENGES

LungCAD system consists of 5 stages: 1. *lung segmentation* to identify the lung area within the chest; 2. *candidate generation* which identifies suspicious unhealthy candidate regions of interest (ROI) from a medical image; 3. *feature extraction* that computes descriptive features for each candidate so that each candidate is represented by a vector $\mathbf{x}$ of numerical values or attributes [15]; 4. *classification* that differentiates candidates based on candidate feature vectors; 5. *visual presentation of CAD findings* to the radiologist in order for him to accept or reject the CAD findings. In this section, we focus on learning the classifier in Step 4.

Automatic learning technologies greatly reduce the time required to develop algorithms that act as "second readers" besides improving the diagnostic accuracy. Many standard algorithms (such as support vector machines (SVM), backpropagation neural nets, kernel Fisher discriminants) have been used to learn classifiers for detecting malignant structures [2, 11]. However, these general-purpose learning methods either make implicit assumptions that are commonly violated in CAD applications, or cannot effectively address the difficulties arisen when learning a CAD system.

Non-IID Data Traditional learning methods almost universally assume that the training samples are independently drawn from an identical albeit unobservable underlying distribution (the IID assumption), which is often not the case in CAD systems. Due to spatial adjacency of the regions identified by a candidate generator, both the features and the class labels of several adjacent candidates are highly correlated. This is true both in the training set and in the testing data. A batch-classification algorithm in [14] derives a probabilistic classification model by specifying a priori guess on the candidate labels with a covariance matrix Σ that encodes the spatial-proximity-based correlations within an image. Multiple-instance learning methods [9, 3] optimize the classifier design by taking into account the fact that multiple candidates can exist to associate with a single malignant structure. Random effects may exist in patient images from the same hospital, or exist in different candidates extracted from the same patient. The approach in [7] proposes to use additional mix-effect parameters, each for one hospital, or for one patient. All these algorithms improve the classification accuracy significantly.

Unbalanced Data and Speed In the candidate identification stage, high sensitivity (ideally close to 100%) is essential, because any cancers missed at this stage can never be found by the CAD system, which potentially produces many false positives (less than 1% of the candidates are positive), making the classification problem highly unbalanced. Moreover, a CAD system has to satisfy real-time requirements that it finishes running during the radiologists first read. These issues were addressed by employing effective cascaded classification frameworks as shown in [4, 5]. The method in [4] investigates a cascaded classification approach that solves a sequence of linear programs, each constructing a sparse hyperplane (linear) classifier. It incorporates the computational complexity of various features into the cascade design for time efficiency. A more recent work [5] does not follow standard cascade procedure where individual classifiers are optimized towards one specific stage given the candidates survived from early stages. Instead, it uses a novel AND-OR cascade training strategy which optimizes all of the classifiers in the cascade in parallel by minimizing the regularized risk of the entire system and providing implicit mutual feedback to individual classifiers to adjust parameter design. These cascaded approaches have been compared with the well-known cascade AdaBoost, and are superior with many additional advantages.

Irrelevant and Redundant Features When searching for descriptive features, researchers often deploy a large amount of experimental image features to describe the identified candidates, which consequently introduces irrelevant and redundant features. Feature selection is essential in CAD systems. A previous LungCAD system [15] utilizes a greedy forward selection approach to select one feature at one time from the feature set according to certain discriminant score ranking. Recent research has focused more on general sparsity treatments to construct sparse estimates of classifier parameters, such as in [6, 4]. These models control the classifier complexity by sparse-favoring regularization terms, such as the ℓ_1-norm regularization $||\mathbf{w}||_1 = \sum |w_i|$ for a linear classifier of the form $sign(\mathbf{w}^T\mathbf{x})$.

3. LUNGCAD TRIAL DESIGN

The clinical trial design is illustrated in Figure 2. The principal challenges we faced in designing the clinical trial are described below:

Measure incremental improvement: The principal value of CAD is determined *not* by its stand-alone performance, but rather by carefully measuring the *incremental value* of Computer-Aided Diagnosis in normal clinical practice; as reflected in incremental improvement in accuracy as objective evaluation by the radiologist.

Patient management impact: It is not enough that LungCAD improves the detection of lung cancer. It must result in a net improvement in patient management since unnecessary false positive findings lead to unnecessary follow-ups.

Ground truth: As discussed earlier, due to the unavailability of lung biopsies , an alternative method had to be devised for determining ground truth.

We retrospectively collected MDCT studies from 200 consecutive patients (mean age: 61.5y, 56% male) who had been

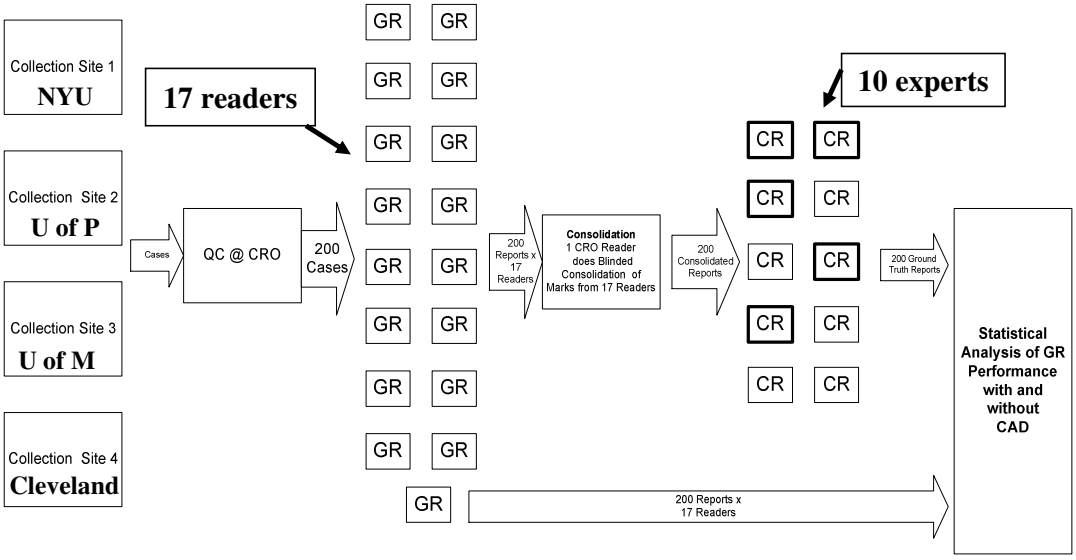

Figure 1: A multicenter, Multi-Reader Multi-Case (MRMC) retrospective clinical study to assess the incremental value of LungCAD in the identification of pulmonary nodules on thoracic CT examinations (CRO= contract research organization, GR=general radiologist, CR=chest radiologist).

referred for evaluation of potential pulmonary nodules from 4 clinical sites: NYU, Univ. of Pennsylvania, Univ. of Maryland and the Cleveland Clinic; These studies were processed by an independent Contract Research Organization (CRO), BioImaging, Inc, Yardley PA. 4 studies were excluded due to respiratory or cardiac motion, or image artifacts.

All 196 studies were initially evaluated by 17 board-certified general radiologists (GR) in active community practice, each using a predetermined randomized order, to detect potential nodules of diameter $\geq 3mm$. The GR's were required to score potential nodules on a "*nodule*" scale, from 1 ("unlikely") to 10 ("definite"). GR's were also required to determine if each nodule could be identified as "*actionable*" again on a 10 point scale ($0-2$ denoting "no followup needed", $3-6$ "indeterminate", >6 "definite need for followup"). To illustrate, a benign calcified granuloma would be represented as true (10), non-actionable (<3) nodule.

Then CAD-identified potential nodules were presented to the GR's (after eliminating nodules that had already been found by the GR), and were assessed using the same two scales. These blinded, independent reviews were re-sent to the CRO, where findings were examined by an independent fellowship trained chest radiologist to consolidate any nodules independently found by more than one GR.

The results were then reviewed separately by 5 fellowship-trained expert chest radiologists (CR) randomly chosen from a panel of 10, each interpreting 100 studies. Expert CR's were required to evaluate each nodule separately without knowledge of whether these had been identified by radiologists or by CAD, and to assess them on both a "nodule" and "actionability" binary decision and its rating. Further, the nodule size and lung lobe in which each nodule was seen was also recorded. For nodule candidates to be considered true nodules (ground truth) a minimum consensus of 3 out of 5 experts was necessary.

A note on sample size: Based on pilot studies, we assumed that at least 60% of patients would have a nodule in an average of 3 lobes, that the CR's would have average ROC area without CAD of 0.80 with moderate inter-reader variability, and that CAD would improve the ROC area by 0.025. To yield 80% power in the trial, we estimated that 17 readers and 200 patients would suffice.

4. CLINICAL TRIAL RESULTS

Ground truth was defined as having at least 3 of 5 expert chest radiologists identifying at least one nodule in a lobe (*affected* lobe); otherwise, lobes were labeled *normal*. Similarly, an *actionable* lobe was one in which 3 or more CR's identified one or more actionable nodules.

A total of 1320 ($\geq 3mm$) nodules were identified in 196 patients of which 863 (65.4%) were interpreted by expert CR's as *actionable*. (Unless specified otherwise, from here on all nodules will be assumed to be in the clinically relevant range of $\geq 3mm$ in diameter.) 181 patients had at least 1 nodule (prevalence rate of 92.3%): only 15 patients were interpreted as *normal* (all lobes were normal). 1320 nodules were detected in 525 (53.6%) of 980 ($= 196 \times 5$) potentially evaluable lobes of which 397 (40.5%) had at least one actionable nodules.

The primary measurement for the diagnostic accuracy of the 17 general radiologists (GR), both with and without CAD, for detecting solid pulmonary nodules, is the area under ROC curve, using lobes as the unit of analysis. A nonparametric estimator was used to adjust for the clustered data as described in [12]. Sensitivity was defined as the probability that a GR identified at least one nodule in an *affected* lobe; specificity was defined as the probability that a GR did not identify a nodule in a *normal* lobe (i.e., correctly identified it as nodule-free).

Figure 2 shows that the 17 GR's accuracy for identifying nodules ranged from 0.704 to 0.853 without CAD to

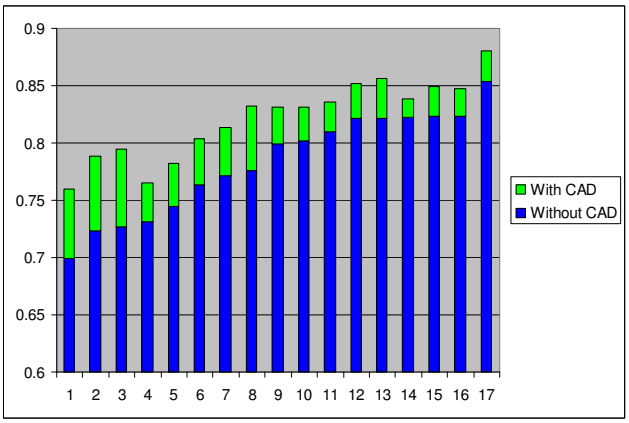

Figure 2: Area under receiver operating curve with and without CAD, for actionable solid nodules.

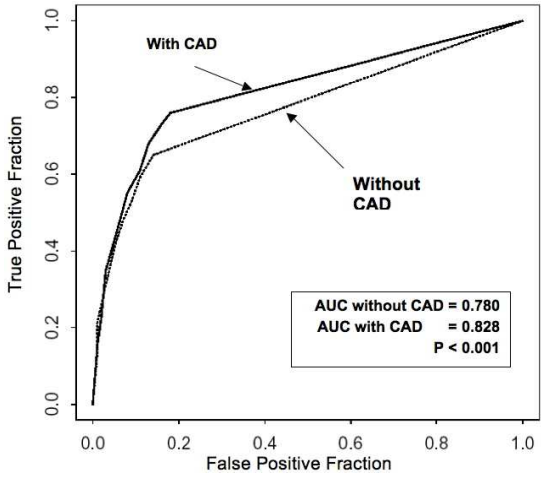

Figure 3: Average nonparametric ROC curve of all 17 readers for detecting nodules without and with CAD.

0.738 to 0.883 with CAD. The most important result was that *every one of the 17 GR's had statistically significantly greater accuracy with CAD for detecting lung nodules.* Assessed collectively, the GR's mean accuracies were 0.780 and 0.828, without and with CAD, respectively ($p < 0.001$; 95% CI of 0.036 to 0.059), as shown in Figure 3.

Similar results were achieved for the clinically significant *actionable* nodules: the 17 GR's accuracy for ranged from 0.699 to 0.854 without CAD to 0.760 to 0.880 with CAD. Again, *every one of the 17 GR's had statistically significantly greater accuracy with CAD for identifying actionable lung nodules.*We stress these findings because most CAD trials demonstrate a statistically significant increase for the readers considered as a group, with only some of the readers individually having statistically significantly greater accuracy. These results are particularly significant because every GR showed statistically significant improvement for both tasks - detecting nodules, and identifying actionable nodules. Figure 4 shows the ROC performance for all 17 readers without and with CAD for identifying actionable nodules.

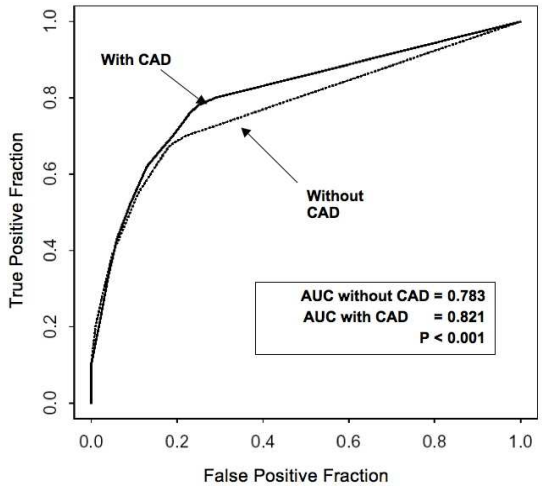

Figure 4: Average nonparametric ROC curve of all 17 readers without and with CAD for identifying actionable nodules.

We varied the definition of expert truth by changing the number of expert confirmations required for acceptance from any 1 CR to 2, 3, 4, 5 expert CR's for both nodules and actionable nodules. With one exception, every one of the 17 GR's showed statistically significant improvement both for detection and identification of actionable nodules with CAD. (The sole exception was the case that all 5 expert CR's must agree about an actionable nodule - which tended to happen for fewer and more obvious actionable nodules, thus making it harder to shown statistically significant improvement, but the trend was towards improvement with CAD.) In another analysis, statistical improvement in GR's accuracy was achieved for all nodules regardless of size ($\geq 3mm$).

To determine the patient management impact we estimated the number of patients, where CAD lead to a positive management change: i.e., a recommendation for additional imaging studies and/or biopsy in an *actionable* lobe which was missed without CAD); and estimated the number of patients where CAD lead to a negative management change: a recommendation for additional studies and/or biopsy in a normal lobe which was correctly diagnosed without CAD. As this is a patient-level analysis, patients with both positive and negative management changes were labelled as a positive change, under the assumption that detecting a missed nodule is more beneficial to a patient than the risk of an unnecessary follow-up (typically another imaging exam).

The average number of patients with a positive management change resulting from using CAD was 24.8 (averaged across the CR's), meaning that 7.9 patients (= 196/24.8) must be evaluated for a positive management change, on average. On the other hand, 12 patients had negative management changes (averages across the 17 CR's), meaning that 16.3 patients must be evaluated with CAD for a negative management change to result. As the positive management changes exceeded the negative management changes on average, this was sufficient, even without considering that on

average positive management changes are more beneficial than negative managment changes are harmful.

Additional details on the multi-reader, multi-case (MRMC) statistical methodology used, are provided in our submission and in [16]. The LungCAD clinical trial summary of safety and effectiveness [8] (which is available on the FDA's web site) contains many more results and analyses, including: patient-level analysis of GR's increase in accuracy with CAD, bootstrap sampling to estimate variability of expert CR's.

5. DISCUSSION

To summarize our clinical results, CAD is an effective second reader, both for detecting nodules and for identifying potentially actionable nodules. The false positive rate is acceptably low given the increased rate of positive management changes. These findings have resulted in LungCAD being granted clinical approval by the FDA for detecting solid pulmonary nodules from CT thorax studies. Although some debate remains about the precise value of screening (for breast cancer, and now for lung cancer), all experts agree that early detection is key for improvement of cancer cure rates. Many efforts are ongoing to pave the way for MDCT to be used for identifying lung cancer at early stages. However, much remains to be done in this area. First, our study focused on solid pulmonary nodules; in high risk patients, part-solid and ground-glass nodules (GGN) are also seen on chest MDCTs. GGNs are defined as nodules with hazy attenuation without obscuration of underlying vascular markings, and will necessitate the development of improved machine learning and image processing methods to detect.

Our focus in this study been to *detect* pulmonary nodules. However, the eventual goal is not just to detect nodules, or even to detect actionable nodules, but to *detect lung cancer* in early stages, and thereby, intervene and treat the patient and improve survival. Therefore, CAD needs to move in the next few years, from detecting nodules to *classifying* nodules as benign or malignant. A first step could be to report the probability of malignancy, although the clinical and regulatory challenges to design a trial to prove the efficacy of such a system would be daunting (larger sample size is not the answer - our study took nearly two years to complete - and the FDA is already taking steps to reduce the regulatory burden, while ensuring the safety and efficacy of CAD). An even more intriguing notion would be to identify lesions that are currently benign, but would have a high probability of turning malignant - pre-cancerous lesions - to move from a reactive paradigm of treating cancer to a more proactive paradigm of prevention.

We have described some machine learning challenges in the lung CAD domain and reviewed some of our previous machine learning work. Our methods are not specific to lung cancer only, and have shown equivalent or superior performance on other data sets. For instance, the PECAD (Pulmonary Embolism) problem (that formed the basis of the 2006 KDDCup) is very different in its evaluation criteria; treatment of PE is systemic (as opposed to localized in lung cancer) and the goal is to identify patients as having one of more PE's or being PE-free. In the ColonCAD problem, the goal is to detect all pre-cancerous polyps; the cost of a false positive is not very high, and the treatment of choice is to remove all potentially suspicious lesions. Yet, despite the very different optimization criteria and the vastly different medical domain knowledge, many of these machine learning methods described here, also translate to these and other CAD problems.

6. REFERENCES

[1] American Lung Association. Trends in lung cancer morbidity and mortality report. 2006.

[2] S. G. Armato-III, M. L. Giger, and H. MacMahon. Automated detection of lung nodules in CT scans: preliminary results. *Medical Physics*, 28(8):1552 – 1561, 2001.

[3] J. Bi and J. Liang. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[4] J. Bi, S. Periaswamy, K. Okada, T. Kubota, G. Fung, M. Salganicoff, and R. B. Rao. Computer aided detection via asymmetric cascade of sparse hyperplane classifiers. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[5] M. Dundar and J. Bi. Joint optimization of cascaded classifiers for computer aided detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] M. Dundar, G. Fung, J. Bi, S. Sandilya, and R. B. Rao. Sparse fisher discriminant analysis for computer aided detection. In *Proceedings of SIAM International Conference on Data Mining*, 2005.

[7] M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao. Learning classifiers when the training data is not IID. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.

[8] Food and Drug Administiration. Siemens Syngo lung CAD summary of safety and effectiveness, PMA No.0500022. October 2006.

[9] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance algorithms for computer aided diagnosis. In *Advances in Neural Information Processing Systems*, 2006.

[10] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun. Cancer statistics. *CA Cancer J. Clin.*, 57:43–66, 2007.

[11] D. P. Naidich, J. P. Ko, and J. Stoechek. Computer aided diagnosis: Impact on nodule detection amongst community level radiologist. A multi-reader study. In *Proceedings of CARS 2004 Computer Assisted Radiology and Surgery*, pages 902 – 907, 2004.

[12] N. A. Obuchowski. Nonparametric analysis of clustered roc curve data. *Biometrics*, 53:170–180, 1997.

[13] S. J. Swensen, J. R. Jett, T. E. Hartman, D. E. Midthun, S. J. Mandrekar, S. L. Hillman, A.-M. Sykes, G. L. Aughenbaugh, A. O. Bungum, and K. L. Allen. CT screening for lung cancer: five-year prospective experience. *Radiology*, 235(1):259–265, 2005.

[14] V. Vural, G. Fung, B. Krishnapuram, J. Dy, and R. B. Rao. Batch-wise classification with applications to computer aided diagnosis. In *Proceedings of European Conference on Machine Learning*, 2006.

[15] M. Wolf, A. Krishnan, M. Salganicoff, J. Bi, M. Dundar, G. Fung, J. Stoeckel, S. Periaswamy, H. Shen, P. Herzog, and D. P. Baidich. CAD performance analysis for pulmonary nodule detection on thin-slice MDCT scans. In H. Lemke, K. Inamura, K. Doi, M. Vannier, and A. Farman, editors, *Proceedings of CARS 2005 Computer Assisted Radiology and Surgery*, pages 1104–1108, 2005.

[16] X.-H. Zhou, N. A. Obuchowski, and D. K. McClish. *Statistical Methods in Diagnostic Medicine*. Wiley, New York, NY, 2002.