Published in final edited form as:
Cancer Prev Res (Phila). 2008 June ; 1(1): 56–64. doi:10.1158/1940-6207.CAPR-08-0011.

A prediction model for lung cancer diagnosis that integrates genomic and clinical features

Jennifer Beane, Ph.D.[1,2], **Paola Sebastiani, Ph.D.**[3], **Theodore H. Whitfield, Sc.D.**[4], **Katrina Steiling, M.D.**[1], **Yves-Martine Dumas, M.P.H.**[1], **Marc E. Lenburg, Ph.D.**[1,2,5], and **Avrum Spira, M.D.**[1,2,*]

[1]The Pulmonary Center, Boston University Medical Center, Boston, MA

[2]Bioinformatics Program, Boston University, Boston, MA

[3]School of Public Health, Boston University, Boston, MA

[4]Biostatistics Solutions Consulting, Boston, MA

[5]Department of Genetics and Genomics, Boston University School of Medicine, Boston, MA

Abstract

Background—Lung cancer is the leading cause of cancer death, in part due to lack of early diagnostic tools. Bronchoscopy represents a relatively noninvasive initial diagnostic test in smokers with suspect disease, but has low sensitivity. We have reported a gene expression profile in cytologically normal large airway epithelium obtained via bronchoscopic brushings that is a sensitive and specific biomarker for lung cancer. Here, we evaluate the independence of the biomarker from other clinical risk factors and determine the performance of a clinicogenomic model that combines clinical factors and gene expression.

Methods—Training (n = 76) and test sets (n = 62) consisted of smokers undergoing bronchoscopy for suspicion of lung cancer at five medical centers. Logistic regression models describing the likelihood of having lung cancer using the biomarker, clinical factors, and these data combined were tested using the independent set of patients with non-diagnostic bronchoscopies. The model predictions were also compared with physicians' clinical assessment.

Results—The gene expression biomarker is associated with cancer status in the combined clinicogenomic model ($p < 0.005$). There is a significant difference in performance of the clinicogenomic relative to the clinical model ($p < 0.05$). In the test set, the clinicogenomic model increases sensitivity and NPV to 100%, and results in higher specificity (91%) and PPV (81%) compared to other models. The clinicogenomic model has high accuracy where physician assessment is most uncertain.

Conclusions—The airway gene expression biomarker provides information about the likelihood of lung cancer not captured by clinical factors, and the clinicogenomic model has the highest prediction accuracy. These findings suggest that use of the clinicogenomic model may expedite

*To whom correspondence should be addressed: Avrum Spira, M.D., The Pulmonary Center, Boston University Medical Center, 715 Albany Street, R304, Boston, MA 02118, Phone: 617-414-6980, Fax: 617-536-8093, aspira@bu.edu.

more invasive testing and definitive therapy for smokers with lung cancer and reduce invasive diagnostic procedures for individuals without lung cancer.

Introduction

Lung cancer is the leading cause of cancer death in the U.S and the world, with over 1 million deaths worldwide annually[1]. Eighty-five to ninety percent of subjects with lung cancer in the US are current or former smokers, with 10–20% of heavy smokers developing this disease [2]. Lack of effective tools to diagnose lung cancer at an early stage before it has spread to regional lymph nodes or metastasized beyond the lung has resulted in a 5-year mortality rate of 80 to 85%[3].

Smokers are often suspected of having lung cancer based on abnormal radiographic findings and/or symptoms that are not specific for lung cancer. Fiberoptic bronchoscopy represents a relatively noninvasive initial diagnostic test to employ in this setting, with cytologic examination of materials obtained via endobronchial brushings, bronchoalveolar lavage and endo- and transbronchial biopsies of the suspect area[4,5]. While cytopathology is 100% specific for lung cancer, the sensitivity of cytologic examination of materials obtained at bronchoscopy ranges from 30% for small peripheral lesions to 80% for centrally located endobronchial tumors[6]. Given the relatively low sensitivity of bronchoscopy, additional and more invasive diagnostic tests are routinely needed which are costly, incur risk, and prolong the diagnostic evaluation of patients with suspect lung cancer. Determining which suspect lung-cancer patients with cancer-negative bronchoscopies should undergo these additional diagnostic tests is currently a matter of clinical judgment. We have recently reported a gene expression profile in cytologically normal large airway epithelial cells obtained via brushing at the time of bronchoscopy that serves as a diagnostic biomarker for lung cancer[7]. This biomarker is an accurate predictor of lung cancer at an early and potentially curable stage, and the sensitivity of the biomarker could substantially reduce the number of individuals requiring additional invasive diagnostic testing following a lung-cancer negative bronchoscopy.

Many groups have developed gene expression profiles that can be used to distinguish between different diagnostic and prognostic subgroups in a variety of cancers. An unexplored issue for many of these biomarkers is whether the gene expression patterns are independent of other clinical risk factors. If so, it presents an opportunity to create clinicogenomic models that incorporate both clinical and gene expression predictors of disease likelihood. There are several examples of such clinicogenomic approaches. Pittman *et al.* have shown improved prediction accuracy for breast cancer recurrence through an integrative clinicogenomic model[8], Similarly, Li combined genomic and clinical data in a survival model to predict the outcome of patients with diffuse large-B-cell lymphoma after chemotherapy[9]. Stephenson *et al* integrated gene expression and clinical data using logistic regression modeling to predict prostate carcinoma reoccurrence after radial prostatectomy, and demonstrated that a combined model had the highest predictive accuracy[10]. In the near future, diverse sources of data such as gene expression, genetic, proteomic, and clinical data will likely be integrated to make accurate diagnoses or prognostic predictions for complex diseases such as cancer[11].

With approximately 90 million former and current smokers in the U.S.[12] and the emergence of sensitive but nonspecific chest imaging technologies, patients increasingly present to clinicians with abnormal radiographic findings concerning for the possibility of lung cancer. While no definitive predictive model for lung cancer exists for use in this setting, numerous clinical and radiographic variables have been associated with the likelihood of lung malignancy: age[13], smoking history[14] (including number of pack-years, age started, intensity of smoking and years since quitting), history of asbestos exposure, clinical symptoms including hemoptysis and weight loss[15], size of the nodule or mass and radiographic appearance on chest imaging[15,16], presence of lymphadenopathy, clinical or radiographic evidence for metastatic disease, evidence of airflow obstruction on spirometry[16], and uptake of FDG on positron emission tomography (PET) scan[17,18]. Several groups have developed predictive models using combinations of the above variables in the setting of solitary pulmonary nodules[15,19,20]. Swensen *et al.* compared such a model for the presence of solitary pulmonary nodules to predictions made by physicians and found that there was no significant difference; though they suggested that the model had potential in the management of patients with benign nodules[21]. In addition, risk prediction models for lung cancer, including a recent large case-control study of never, former, and current smokers, have been reported[22].

In this study, we sought to evaluate whether the lung cancer predictions made by our large airway gene expression biomarker are independent of other clinical risk factors; and if so, to determine the relative performance of a clinicogenomic model that combines clinical risk factors with the biomarker. We show that the biomarker provides information about the likelihood of a patient having lung cancer beyond that which is contained in the available clinical data, despite the clinical model predictions being highly associated with the subjective clinical assessment of patient risk made by pulmonary physicians. Furthermore, we find that the clinicogenomic model has better diagnostic accuracy than either the clinical model or the gene expression biomarker alone. Our data suggest that the clinicogenomic model could be efficacious in predicting the likelihood of lung cancer in those patients where physicians are most uncertain about the likelihood of disease.

Methods

Patient Population

The present study cohort consists of patients who participated in our study to develop the large airway gene expression biomarker[7]. In that study, we recruited current and former smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at four tertiary medical centers between January 2003 and April 2005 as described previously[7]. All subjects were greater than 21 years of age and had no contraindications to flexible bronchoscopy. Never smokers and subjects who only smoked cigars were excluded from the study. All subjects were followed, post-bronchoscopy, until a final diagnosis of lung cancer or an alternative diagnosis was made (mean follow-up time = 52 days). 129 subjects (60 smokers with lung cancer and 69 smokers without lung cancer) who achieved final diagnoses as of May 2005 and had high quality microarray data were included in the primary sample set. Seventy-seven of these samples were randomly assigned to the training

set. The training set for the current study (n = 76) excluded one of these training set samples due to incomplete smoking history (Figure 1). After completion of the primary study, a second set of samples (n=35) was collected prospectively from smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at five medical centers between June 2005 and January 2006. Inclusion and exclusion criteria were identical to the primary sample set. The test set samples in the current study (n=87) combined both the remaining samples from the primary sample set (n=52) and this prospective test set (n=35), but we chose to limit the test set to the subset of patients that did not have a definitive diagnosis following bronchoscopy (n = 62), as is shown in Figure 1 and described in more detail below. Demographic information on all subjects is detailed in Table 1, and information regarding the cell type, stage, and location of the lung tumors (n=78) in the study cohort is shown in Table 2. The study was approved by the Institutional Review Boards of the five medical centers at which patients were recruited (Boston University Medical Center, Boston, MA; Boston Veterans Administration, West Roxbury, MA; Lahey Clinic, Burlington, MA; and St. James's Hospital, Dublin, Ireland, St. Elizabeth's Medical Center, Boston, MA) and all participants provided written informed consent

Large Airway Gene Expression Biomarker for Lung Cancer

Using the Affymetrix HG-U133A microarray, we have previously developed a gene expression biomarker for lung cancer utilizing gene expression profiles in cytologically normal large airway epithelial cells collected from brushing the right mainstem bronchus of smokers undergoing bronchoscopy for suspicion of lung cancer (GEO accession number GSE4115) [7]. The biomarker was developed using the training set of the current study (n = 76) with the addition of one additional sample that did not have a complete smoking history (Figure 1). The biomarker was constructed from the expression levels of 80 probesets (72 unique genes and 7 unknown transcripts) using the weighted-voting algorithm[23] that combines these expression levels into a biomarker score. A positive score is predictive of cancer and a negative score is predictive of no cancer. In this study, we use the biomarker score as a starting point for the following statistical analyses: (1) building three logistic regression models to determine the likelihood of lung cancer using the clinical risk factors alone, the biomarker alone, or the likelihood of lung cancer using the clinical risk factors and biomarkers combined; (2) comparison of these three models using their predictive values on a test set of patients not used in the initial model building phase; (3) comparison of the clinical models with assessments made by expert clinicians.

Construction of Logistic Regression Models

Logistic regression models to quantify the probability of a patient having lung cancer were generated using the training set samples (n = 76). This training set included patients who had cytopathology findings that either confirmed a diagnosis of lung cancer or alternate non-cancer pathology. Patients with diagnostic bronchoscopies were included in the training set to maximize the number of samples and because exclusion of these samples was unnecessary to develop models capable of accurately predicting the lung cancer status of patients with non-diagnostic bronchoscopies (data not shown). For the clinical and clinicogenomic models, the available clinical variables (Table 1) included age and pack years of smoking, and the following dichotomous variables: gender (Male=1, Female=0),

race (Caucasian=1, 0 otherwise), smoking status (former smokers that quit greater than or equal to 10 years ago=1, 0 otherwise), hemoptysis (presence=1, 0 otherwise), lymphadenopathy (mediastinal or hilar lymph nodes greater than 1cm on CT chest scan=1, 0 otherwise), mass size (having a mass size greater than 3cm=1, 0 otherwise). PET scan information was only available for 15 patients and was not included in the model. Backward stepwise model selection using Akaike's Information Criterion (AIC)[24] was used to select the optimal clinical model for the probability of a patient having lung cancer.

To create an integrated clinicogenomic model and determine the independence and magnitude of the contribution of the gene expression biomarker after adjusting for the effects of the clinical variables, we first added the biomarker to the optimal clinical model. The biomarker scores and all of the available clinical variables were then used with backward stepwise model selection by AIC to select the optimal model. Both approaches yielded the same combined model. In order to verify that the biomarker score performs similarly in logistic regression as in the weighted-voting prediction algorithm used in our previous work[7], the accuracy, sensitivity, specificity, positive predictive value, and negative predictive value were compared for the weighted-voting predictions and the predictions made by a logistic regression model that included only the biomarker score across the independent test samples.

Comparison of Model Performance on Independent Patients

The performance of the logistic regression models (clinical, biomarker, and clinicogenomic) was initially evaluated on the subset of patients in the training set (n = 76) in which the cytopathology of materials obtained at bronchoscopy were non-diagnostic (n = 56) (Figure 1). We chose to focus on non-diagnostic bronchoscopies so as to specifically assess the utility of the gene expression biomarker and clinical parameters in the setting of patients that require further diagnostic evaluation for lung cancer. More importantly, we also tested the models in the non-diagnostic bronchoscopy test set (n = 62) (Figure 1). For each of the models, patients that had a probability of lung cancer greater than or equal to 0.5 were classified as having lung cancer, and patients with a probability less than 0.5 were classified as not having lung cancer. Receiver Operating Characteristics (ROC) curves were also used to compare the clinical model to the clinicogenomic model in the training set patients with non-diagnostic bronchoscopies, the independent test set, and combined set of all patients with non-diagnostic bronchoscopies (n=118). In order to assess whether or not two ROC curves based on the same set of samples were significantly different, methods developed for comparing ROC curves derived from the same cases were used[25,26]. To compare ROC curves based on different sample sets, we used a two-sample z-test. The ROC curves serve as a common scale for evaluating the additional merit of variables added to the model as odds ratios for two different variables may not be comparable[27]. The accuracy, sensitivity, specificity, positive predictive value, and negative predicate value were also calculated across the independent test set for the clinical model, the biomarker model, and the clinicogenomic model.

Subjective Clinical Assessment

Three independent pulmonary clinicians practicing at a tertiary medical center, blinded to the final diagnoses, evaluated each patient's clinical history at the time of bronchoscopy. The history included but was not limited to age, smoking status, cumulative tobacco exposure, co-morbidities, symptoms/signs, radiographic findings, and PET scan results if available. Based on this information, the clinicians classified each patient into one of three risk groups: low (<10% assessed probability of lung cancer), medium (10–50% assessed probability of lung cancer) and high (>50% assessed probability of lung cancer). The final subjective assignment for each subject was decided by choosing the median opinion. The interrater reliability for the clinical classification of patients' non-diagnostic bronchoscopies was significant indicating that the level of agreement between the clinicians was greater then would be expected by chance as measured by the kappa statistic ($\kappa = 0.57$; $p < 0.001$)[28].

Comparison of Subjective Clinical Assessment to the Clinicogenomic Model

The sample size for building a comprehensive clinical model to predict the risk of having lung cancer was limited as was the scope of variables that were available for inclusion in the clinical and clinicogenomic model. We therefore sought to determine if the clinical model performs similarly to the subjective clinical assessment made by pulmonary specialists as this assessment is (1) "trained" on the large number of patients seen over each clinician's career and (2) considers all of the information contained within a patient's medical records. A Wilcoxon test was used to assess whether or not the clinical model-derived probability of having lung cancer varied between samples classified as low, medium, or high cancer risk by the clinicians.

Statistical Analysis

All statistical analyses were conducted using R statistical software v 2.2.1.

Results

Evaluating the Gene Expression Biomarker as an Independent Predictor of Lung Cancer

The demographic and clinical characteristics as well as the mean and standard deviation for the biomarker scores stratified by cancer status and membership in the training or test sets are shown in Table 1. Age, race, pack years of smoking, lymphadenopathy, mass size, and the biomarker score were significantly different ($p < 0.001$) between patients with and without lung cancer. The test and training sets, however, were well balanced for the variables used in the analyses (though the incidence of having a mass size greater than 3 cm was somewhat lower in the test set compared to the training set; $p = 0.047$). Information about the cell type, stage, and location of the tumors in the cancer patients, as well as the fraction of diagnostic bronchoscopies for each subgroup is shown in Table 2. Effect estimates and derived odds ratios (OR) for the variables in each of the three logistic regression models are shown in Table 3. We found that the optimal clinical model for this cohort did not include pack years. This is likely due to the strong correlation between age and pack years ($r = 0.56$, $p < 0.001$) in the training set. A clinical model constructed with pack years instead of age yielded similar results when tested on the independent test set

(n=62), with an accuracy of 84% and 85%, and area under ROC curves of 0.94 and 0.86, respectively. The optimal clinical model did not include smoking status (former vs. current smokers) regardless of how time since quitting was dichotomized. In addition, dichotomizing mass size using a threshold value of 2cm (instead of 3cm) produced clinical and clinicogenomic models with similar overall accuracy.

A logistic regression model describing the likelihood of having lung cancer derived from the biomarker score produced equivalent results to the weighted-voting algorithm predictions of lung cancer status described previously(4), resulting in 8 versus 7 incorrect classifications, indicating that the biomarker score is an accurate way to model the original biomarker prediction algorithm in the clinicogenomic model. The biomarker score is a significant predictor of lung cancer likelihood in both the biomarker only model ($p < 0.001$) and in the clinicogenomic model ($p < 0.005$). In the clinicogenomic model, the coefficients of the clinical variables are largely unchanged from the clinical model, and the coefficient of the biomarker is largely unchanged from the biomarker only model. These data suggest that the gene expression biomarker and the clinical variables are independent predictors of lung cancer risk.

Evaluating the Performance of the Clinicogenomic Model

The three models were used to predict the cancer status of a subset of the training samples with non-diagnostic bronchoscopies (n=56), the independent test samples (n=62), and these two sets combined (n=118). ROC curves were used to compare the performance of the clinical model to the clinicogenomic model (Figure 2). The clinicogenomic model had better performance than the clinical model in all three sample sets. While this difference in performance does not reach statistical significance in the test set, when the training and test sets were combined, there was a significant difference in the area under the curve between the clinicogenomic and clinical models ($p < 0.05$). The performance of the models in the training set samples does not appear to be any better than in the test set samples ($p = 0.25$, for the difference in the area under the ROC curves; the AUC difference is 0.065; 95% CI: −0.046–0.174). This suggests that the models do not overfit the training data and that it is therefore reasonable to combine the training and test sets to assess the significance of the difference in the performance of the clinical and clinicogenomic models.

The sensitivity, specificity, and positive and negative predictive values for each of the three models were evaluated across the test set (Figure 3). The combined clinicogenomic model increases the sensitivity and negative predictive value to 100% and results in higher specificity and positive predictive value compared to the other models. Cancer subjects with peripheral lesions were well represented in the test set (70.6%), and the clinicogenomic model was equally accurate among peripheral or central lung tumors. The clinicogenomic model also accurately predicted lesions with a mass size less than 3 cm as well as poorly defined radiographic infiltrates in the test set (Table 4). In addition, the performance of the clinical and clinicogenomic models does not appear to be specific to samples with non-diagnostic bronchoscopies as these models had sensitivities of 90% and 95% on independent samples with diagnostic bronchoscopies (n = 25). Finally, training the clinical and clincogenomic models across only the training samples with non-diagnostic bronchoscopies

(n=56) resulted in similar accuracies in the test set (82% and 91%, respectively) and a significant difference in the area under the ROC curves between the models ($p < 0.05$).

Comparing the Clinicogenomic Model to the Clinical Subjective Assessment

In order to evaluate whether or not the clinical model is comprehensive given the relatively small number of variables it contains, we assessed whether it correlates with the median subjective assessment of three pulmonary physicians. There was an association between the clinical model predictions and the clinical subjective assessment across the test set samples (Figure 4). The clinical model probabilities were significantly different between the three physician-assessed risk groups ($p < 0.01$).

Given the association between the clinical model and subjective clinical assessment, we examined the predictions made by the clinicogenomic model stratified by cancer status and subjective clinical assessment category in the test set samples (Figure 5). The physician's opinion is the most uncertain based on the all the clinical data for the 11 samples in the medium risk category. The clinical model is able to classify 7 out of the 11 samples correctly; however, the clinicogenomic model correctly classifies all 11 samples.

Discussion

A previous study by our group reported a gene expression biomarker capable of distinguishing cytologically normal large airway epithelial cells from smokers with and without lung cancer[7]. These cells can be collected in a relatively noninvasive manner from bronchial airway brushings of patients undergoing bronchoscopy for the suspicion of lung cancer. The cytopathology of cells obtained during bronchoscopy is 100% specific for lung cancer, but has a limited sensitivity of between 30 and 80% depending on the stage and location of the cancer, with early stage disease and peripheral cancers having the lowest sensitivity[6]. As a result, physicians are confronted with a difficult decision as to how to manage the care of patients with potentially early-stage curable disease, when bronchoscopy does not return any cells with aberrant cytopathology. Often the decision about whether to proceed with more sensitive and often more invasive diagnostic procedures or to determine if the initial suspicious radiographic finding resolves in subsequent repeat imaging studies, is based on a subjective assessment of the patient's clinical and radiographic risk factors for lung cancer. As the large airway gene expression biomarker uses material that can be easily collected at the time of bronchoscopy (prolonging the procedure by only 2–3 additional minutes), this test could be a useful component of this decision making process if the biomarker captures information about lung cancer risk that is otherwise occult.

Our results suggest that the pattern of gene expression in large airway epithelial cells reflects information about the presence of lung cancer that is independent of other clinical risk factors. This interpretation results from a comparison of models that contain either clinical variables or the biomarker to a combined clinicogenomic model. The comparison shows that the biomarker is significantly associated with the probability of having lung cancer in both the biomarker and clinicogenomic models and that the importance of each of the variables in the combined clinicogenomic model is similar to their importance in the initial uncombined models.

Given the independence of the biomarker and clinical models, it is not surprising that the clinicogenomic model is a better predictor of lung cancer than either of the initial models in an independent test set. ROC curve analysis shows that the clinicogenomic model performs significantly better than the clinical model. Furthermore, the clinicogenomic model increases the sensitivity, specificity, positive and negative predictive value of the clinical model, and its accuracy does not appear to be influenced by the size or location of the lesion. However, these findings need to be validated in larger patient cohorts. One way to accomplish such validation would be to incorporate gene expression measurements into large epidemiological studies investigating lung cancer risk or lung cancer screening trials involving high-risk smokers.

Despite the limitations of a small sample size and limited clinical parameters, we are encouraged that subjective clinical assessment based on a patient's complete medical record is associated with the clinical model probabilities. This is particularly important given that certain variables, such at PET scan findings, were not included in the clinical model as these studies were performed on only a small number of the subjects in our cohort. All available data, such as PET scan findings, were however considered by the pulmonary physicians as part of their subjective assessment of lung cancer likelihood. Further, the clinicogenomic model appears to correctly classify patients assigned to the medium risk subgroup by the clinical subjective assessment. This subgroup of patients is one that is likely to be especially challenging to manage clinically as almost a third of these patients went on to have a final diagnosis of lung cancer.

Our data suggest that a clinicogenomic model that combines gene expression with clinical risk factors for lung cancer has high diagnostic specificity and positive predictive value among patients with non-diagnostic bronchoscopies, including those with small and/or peripheral lesions on chest imaging. This model might therefore serve to identify those patients who would benefit from further invasive testing (e.g. lung biopsy) to confirm the presumptive lung cancer diagnosis and thereby expedite the diagnosis and treatment for their underlying malignancy. In addition, the clinicogenomic model also results in modest increases in diagnostic sensitivity and negative predictive value. Utilization of this clinicogenomic diagnostic might therefore also result in a reduction in the number of individuals without lung cancer who are subjected to additional and more invasive procedures to rule out a lung cancer diagnosis following a non-diagnostic bronchoscopy. If the ultimate sensitivity and negative predictive value of the clinicogenomic model remains close to 100%, this would allow clinicians to confidently use less invasive and less costly approaches (e.g. repeat CT scan in 3–6 months) to follow patients with a low clinicogenomic lung cancer risk score.

The ability of gene expression profiles within cytologically normal airway epithelium to serve as a biomarker for lung cancer raises questions as to the underlying biology of the cancer-specific molecular changes observed in these cells. The high diagnostic accuracy for the biomarker in the setting of small peripheral lung lesions suggests that changes in airway gene expression between smokers with and without lung cancer are unlikely to be a direct effect of the tumor. The presence of antioxidant and inflammation-related genes in the gene expression biomarker[7] raises the possibility that the biomarker detects an airway-wide

cancer-specific difference in response to tobacco-smoke exposure. Given the hypothesis that this field of injury may provide information about the host-carcinogen interaction, alterations in gene expression could precede the development of lung cancer, and explain the somewhat lower specificity of the biomarker relative to its sensitivity. If this is true, the biomarker might potentially be a useful tool to identify smokers at highest risk for disease who may benefit from chemopreventative strategies.

Conclusion

The gene expression pattern of histologically normal large airway epithelial cells collected at the time of bronchoscopy can be used as a biomarker that provides information that is independent from clinical parameters about lung cancer risk. In the setting of patients with suspect lung cancer that do not have a definitive diagnosis after routine cytology/pathology of materials retrieved by bronchoscopy, a clinicogenomic model that combines both clinical factors and the large airway gene expression biomarker results in improved sensitivity, specificity, positive and negative predictive values over the clinical model alone. This suggests that the integrative clinicogenomic model may help expedite invasive diagnostic testing for those smokers with underlying lung tumors, and decrease the number of individuals without lung cancer requiring further invasive diagnostic testing to rule out suspicion of disease.

Acknowledgments

We thank Gang Liu, Xuemei Yang, Sherry Zhang, Frank Schembri and Norman Gerry for support with collection of samples and performing the microarray experiments. We thank Jerome Brody for his guidance with study design and for his critical review of the manuscript. We thank George O'Connor for his critical review of the manuscript. This work was supported by the Doris Duke Charitable Foundation (AS), NIH/NCI R21CA10650 (A.S.), and NIH/NCI R01CA124640 (A.S., M.E.L, and J.B). A.S. is a founder and has equity in ExProDx Inc. M.E.L is a consultant for and has equity in ExProDx Inc.

Abbreviations

NPV	negative predictive value
PPV	positive predictive value
FDG	fluorodeoxyglucose
PET	positron emission tomography
GEO	Gene Expression Omnibus
AIC	Akaike's Information Criterion
ROC	Receiver Operating Characteristic
AUC	area under the curve

Reference List

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin. 2005; 55:74–108. [PubMed: 15761078]

2. Shields PG. Molecular epidemiology of lung cancer. Ann Oncol. 1999; 10 (Suppl 5):S7–11. [PubMed: 10582132]
3. Hoffman PC, Mauer AM, Vokes EE. Lung cancer. Lancet. 2000; 355:479–85. [PubMed: 10841143]
4. Postmus PE. Bronchoscopy for lung cancer. Chest. 2005; 128:16–8. [PubMed: 16002910]
5. Mazzone P, Jain P, Arroliga AC, Matthay RA. Bronchoscopy and needle biopsy techniques for diagnosis and staging of lung cancer. Clin Chest Med. 2002; 23:137–58. ix. [PubMed: 11901908]
6. Schreiber G, McCrory DC. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. Chest. 2003; 123:115S–28S. [PubMed: 12527571]
7. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med. 2007; 13:361–6. [PubMed: 17334370]
8. Pittman J, Huang E, Dressman H, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci U S A. 2004; 101:8431–6. [PubMed: 15152076]
9. Li L. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. Bioinformatics. 2006; 22:466–71. [PubMed: 16339281]
10. Stephenson AJ, Smith A, Kattan MW, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer. 2005; 104:290–8. [PubMed: 15948174]
11. West M, Ginsburg GS, Huang AT, Nevins JR. Embracing the complexity of genomic data for personalized medicine. Genome Res. 2006; 16:559–66. [PubMed: 16651662]
12. McWilliams A, Lam S. New approaches to lung cancer prevention. Curr Oncol Rep. 2002; 4:487–94. [PubMed: 12354360]
13. Trunk G, Gracey DR, Byrd RB. The management and evaluation of the solitary pulmonary nodule. Chest. 1974; 66:236–9. [PubMed: 4371237]
14. Thurston SW, Liu G, Miller DP, Christiani DC. Modeling lung cancer risk in case-control studies using a new dose metric of smoking. Cancer Epidemiol Biomarkers Prev. 2005; 14:2296–302. [PubMed: 16214908]
15. Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory Radiology. 1993; 186:405–13.
16. Mannino DM, Aguayo SM, Petty TL, Redd SC. Low lung function and incident lung cancer in the United States: data From the First National Health and Nutrition Examination Survey follow-up. Arch Intern Med. 2003; 163:1475–80. [PubMed: 12824098]
17. Wahidi MM, Govert JA, Goudar RK, Gould MK, McCrory DC. Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer? : ACCP evidence-based clinical practice guidelines (2nd edition). Chest. 2007; 132:94S–107S. [PubMed: 17873163]
18. Ung YC, Maziak DE, Vanderveen JA, et al. 18Fluorodeoxyglucose Positron Emission Tomography in the Diagnosis and Staging of Lung Cancer: A Systematic Review. J Natl Cancer Inst. 2007
19. Cummings SR, Lillington GA, Richard RJ. Estimating the probability of malignancy in solitary pulmonary nodules. A Bayesian approach. Am Rev Respir Dis. 1986; 134:449–52. [PubMed: 3752700]
20. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med. 1997; 157:849–55. [PubMed: 9129544]
21. Swensen SJ, Silverstein MD, Edell ES, et al. Solitary pulmonary nodules: clinical prediction model versus physicians. Mayo Clin Proc. 1999; 74:319–29. [PubMed: 10221459]
22. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst. 2007; 99:715–26. [PubMed: 17470739]
23. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–7. [PubMed: 10521349]
24. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–23.

25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]

26. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839–43. [PubMed: 6878708]

27. Sullivan PM, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001; 93:1054–61. [PubMed: 11459866]

28. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.

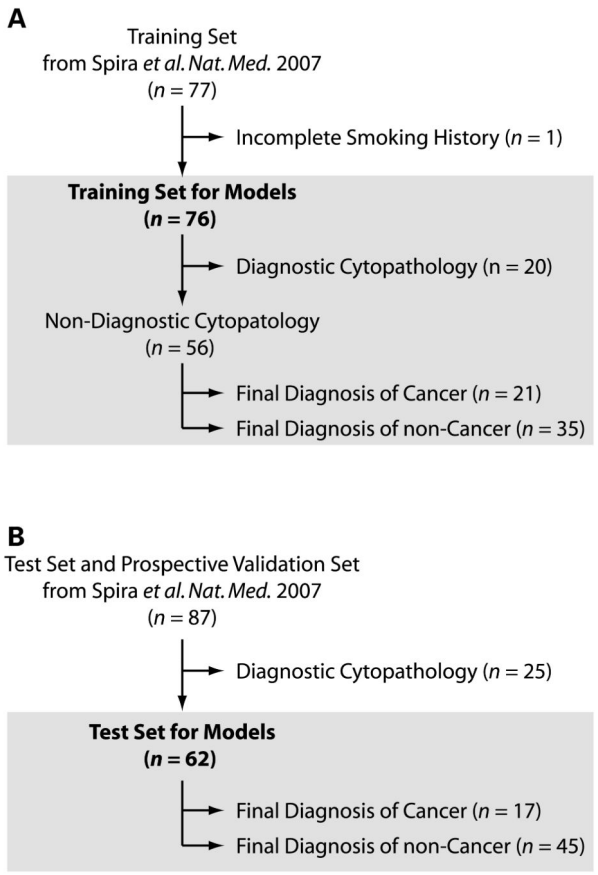

Figure 1.

Training and test sample sets. The training and test samples were derived from a previously published study assaying airway epithelial gene expression from current and former smokers undergoing bronchoscopy for the clinical suspicion of lung cancer[7]. (A.) We previously constructed a gene-expression biomarker that predicts the presence of lung cancer using a training set of 77 patients. For the current study, one of these samples was removed due to incomplete smoking history, resulting in the logistic regression models being trained with data from 76 patients. The models were subsequently tested on the subset of training samples (n=56) that had cytopathology that was non-diagnostic of lung cancer. (B.) The biomarker was also tested on the subset of independent samples with non-diagnostic cytopathology (n =62) from the combined test and prospective validation sample sets (n = 87) used in our previous study.

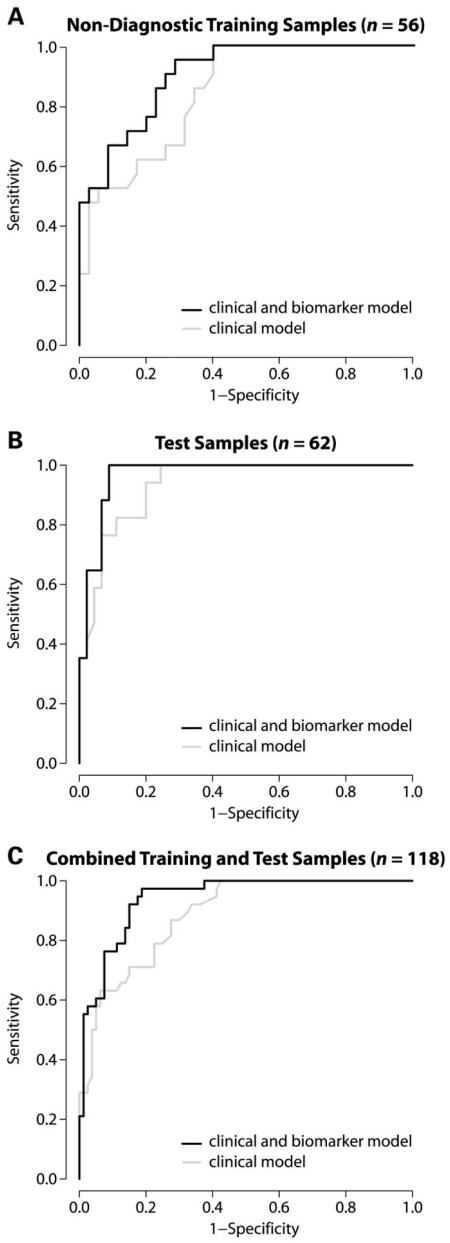

Figure 2.

ROC curves for the clinical model and the clinicogenomic model across the different sample sets. The clinical model (gray line) includes the following variables: age, mass size, and lymphadenopathy, while the clinical and biomarker model includes the above variables and the biomarker score (black line). Both models were derived using the training set samples (n=76). (A.) ROC analysis of the non-diagnostic training set samples (n = 56). The AUC for the clinical and clinicogenomic model is 0.84 and 0.90, respectively. (B.) ROC analysis of the test samples (n = 62). The AUC for the clinical and clinicogenomic model is 0.94 and 0.97, respectively. (C.) ROC analysis of the combined training and test sets (n = 118). The AUC for the clinical and clinicogenomic model is 0.89 and 0.94, respectively, which represents a significant difference between the two curves ($p < 0.05$).

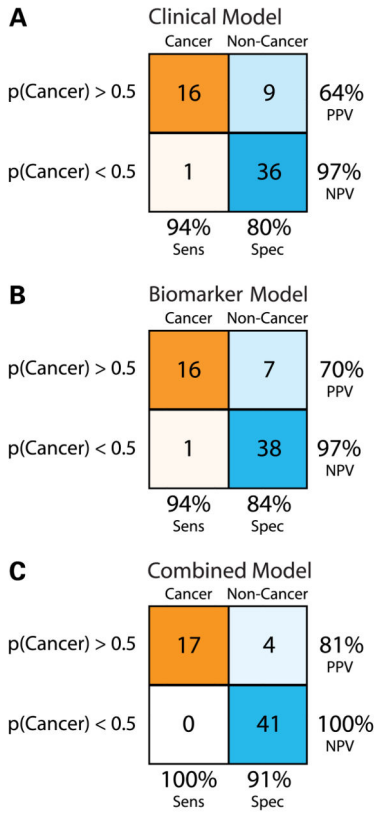

Figure 3.

Performance of three logistic regression models across the test set samples. Samples with model derived probabilities of having lung cancer greater than or equal to 0.5 were classified as cancer, and samples with probabilities less than 0.5 were classified as non-cancer. Samples with a final diagnosis of cancer are indicated in orange while samples with a final diagnosis of no cancer are indicated in blue. The saturation of the colors is representative of the proportion of each final diagnosis group classified as having cancer or no cancer by each of the models. For each model, the sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and the negative predictive value (NPV) are shown. (A.) The Clinical Model (B.) The Biomarker Model (C.) The Clinicogenomic Model. The Clinical Model and the Biomarker Model each perform similarly with accuracies of 84% and 87%, respectively. The Clinicogenomic Model has a greater accuracy (94%), specificity, and PPV than either of the other two models.

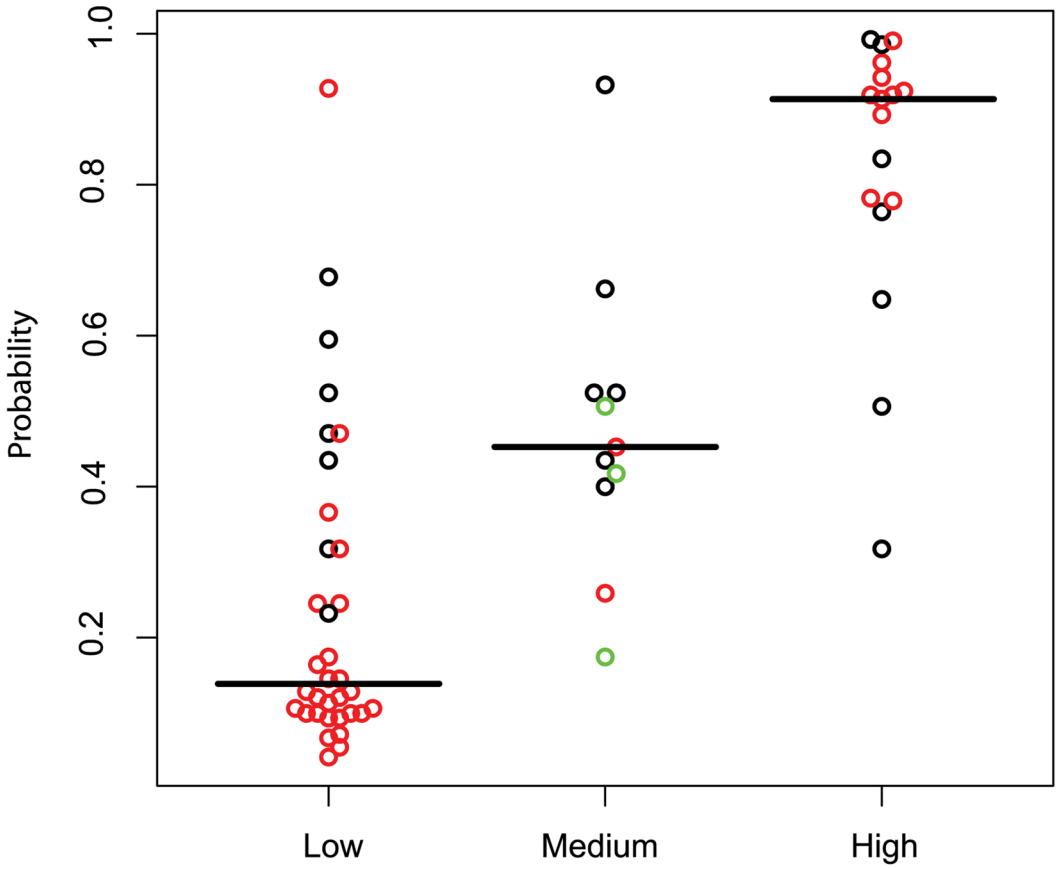

Figure 4.

Association between the probability of having lung cancer as predicted by the clinical model and physician's subjective assessment across the test set samples (n=62). The model derived probabilities are shown on the y-axis and the subjective clinical assessment on the x-axis. Red circles indicate complete agreement among 3 clinicians, black indicates agreement of 2 clinicians, and green indicates no agreement. There are significant differences (Wilcoxon test; $p < 0.01$) between the probabilities in the low versus medium group, the medium versus high group, and the low versus high group.

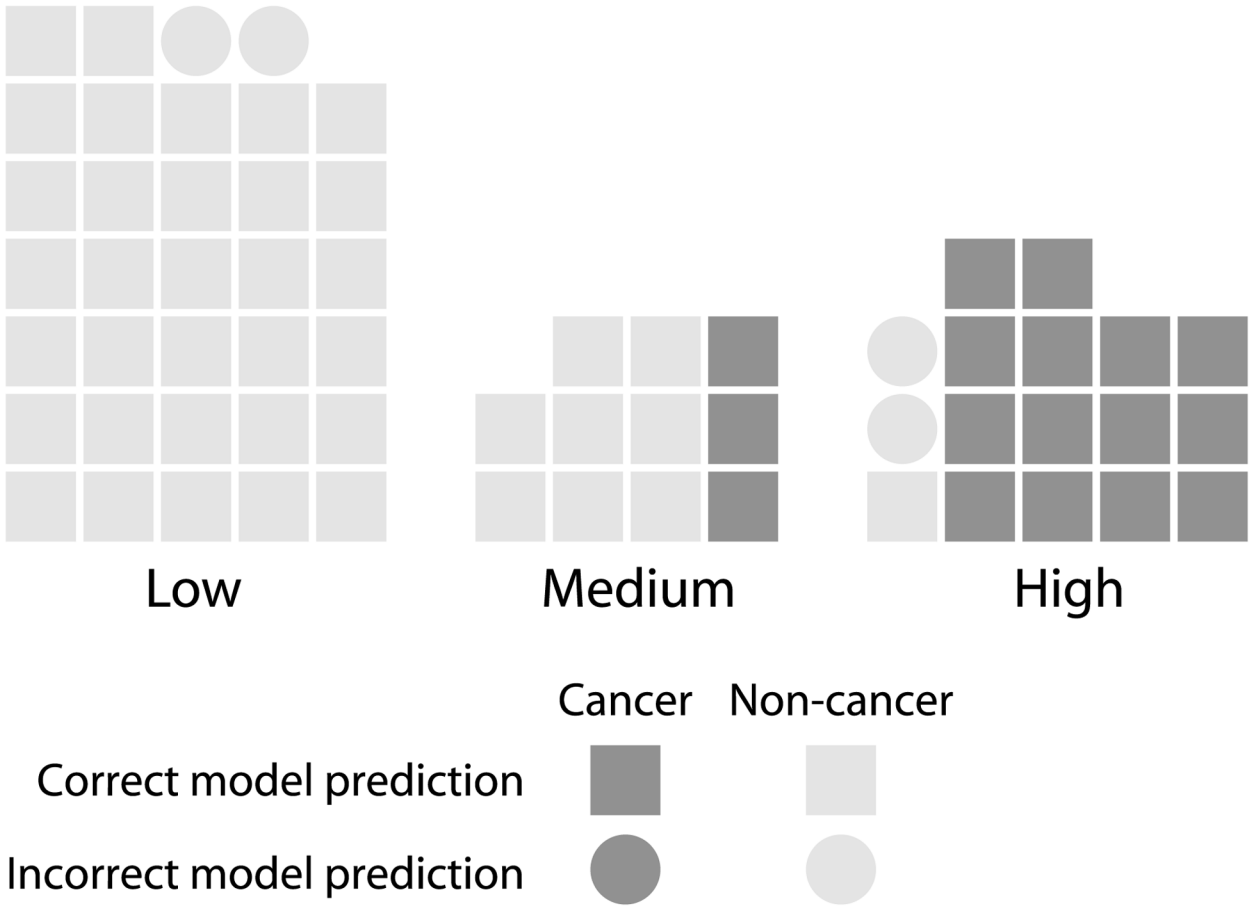

Figure 5.

The clinicogenomic model-derived lung cancer predictions stratified by cancer status and the physician's subjective assessment across the test set samples (n=62). Dark gray represents a final diagnosis of cancer and light gray represents a final diagnosis of non-cancer. Squares represent correct clinicogenomic model predictions and circles represent incorrect model predictions. Each of the samples classified as having a medium risk of lung cancer by physicians was predicted correctly by the clinicogenomic model.

Cancer Prev Res (Phila). Author manuscript; available in PMC 2014 September 18.

Table 1

Demographic, Clinical, and Biomarker Characteristics Stratified by Cancer Status or membership in the training and test sets. Data are means ± standard deviations for continuous variables and proportions with percentages for dichotomous variables.

Factor	Overall (*n* = 163)	Cancer (*n* = 78)	No Cancer (*n* = 85)	*p*†	Train (n=76)	Test (n=62)	*p*†
Age	58.1 ± 14.3	64.5 ± 9.6	52.3 ± 15.4	< 0.001	57.3 ± 14.0	57.5 ± 15.3	0.91
Male	122/163 (74.8)	60/78 (76.9)	62/85 (72.9)	0.59	59/76 (77.6)	42/62 (67.7)	0.25
Caucasian	110/163 (67.5)	67/78 (85.9)	43/85 (50.6)	< 0.001	52/76 (68.4)	36/62 (58.1)	0.22
Smoked Within 10 Years	130/163 (79.8)	60/78 (76.9)	70/85 (82.4)	0.44	62/76 (81.6)	47/62 (75.8)	0.53
Pack Years	44.9 ± 32.0	54.9 ± 26.8	35.7 ± 33.7	< 0.001	45.8 ± 30.2	39.9 ± 35.4	0.3
Diagnostic Bronchoscopy	45/163 (27.6)	40/78 (51.3)	5/85 (5.9)	< 0.001	20/76 (26.3)	0/62 (0)	<0.001
Cancer	78/163 (47.9)	78/78 (100.0)	0/85 (0.0)	< 0.001	40/76 (52.6)	17/62 (27.4)	0.003
Lymphadenopathy	43/163 (26.4)	36/78 (46.2)	7/85 (8.2)	< 0.001	17/76 (22.4)	10/62 (16.1)	0.4
Hemoptysis	15/163 (9.2)	6/78 (7.7)	9/85 (10.6)	0.6	10/76 (13.2)	2/62 (3.2)	0.07
Mass Size > 3 cm	48/163 (29.4)	43/78 (55.1)	5/85 (5.9)	< 0.001	24/76 (31.6)	10/62 (16.1)	0.047
Biomarker	−0.35 ± 8.93	4.65 ± 7.04	−4.94 ± 7.98	< 0.001	0.34 ± 8.97	−2.72 ± 9.12	0.05

† *p*-values are for the comparison of patients with cancer and patients without cancer. Two-sample *t*-tests with unequal variances were used for continuous variables; Fisher's exact test was used for dichotomous variables.

Table 2

Cell type, stage, and location information for lung cancer samples (n=78). The percentage of samples in each grouping where bronchoscopy yielded diagnostic cytopathology for lung cancer is reported. “Other” refers to cases that cannot be characterized as central vs. peripheral.

Cell Type	n	% of Samples with Diagnostic Bronchoscopy
SCLC	14	64.3%
NSCLC (unknown subtype)	15	60.0%
Squamous	27	55.6%
Adenocarcinoma	18	33.3%
Large Cell Carcinoma	4	25.0%
Stage		
Unknown	1	0.0%
1	14	35.7%
2	2	50.0%
3	25	52.0%
4	22	54.5%
Location		
Central	28	71.4%
Peripheral	49	40.8%
Other	1	0.0%

Table 3

Logistic Regression Models Fitted on Training Set Samples. The range, regression coefficients, odds ratio, 95% confidence interval for the odds ratio (95% CI), and the p-value (*p*) of the variables across the training set samples (n=76) is reported.

Model	Range	Coefficient	Odds Ratio	95% CI	*p*
Biomarker Alone					
Intercept	NA	0.07	NA	NA	0.776
Biomarker	(−18.88,16.91)	0.13	1.14	(1.06, 1.21)	0.00017
Clinical Variables Alone					
Intercept	NA	−5.01	NA	NA	0.003
Age	(23, 79)	0.07	1.07	(1.02, 1.13)	0.008
Mass Size	(0,1)	2.19	8.91	(2.08, 38.25)	0.003
Lymphadenopathy	(0,1)	2.09	8.12	(1.45, 45.63)	0.017
Biomarker + Clinical Variables					
Intercept	NA	−4.9	NA	NA	0.014
Biomarker	(−18.88,16.91)	0.13	1.13	(1.04, 1.24)	0.005
Age	(23, 79)	0.07	1.07	(1.00, 1.14)	0.036
Mass Size	(0,1)	1.85	6.38	(1.39, 29.34)	0.017
Lymphadenopathy	(0,1)	2.75	15.69	(2.23, 110.28)	0.006

Table 4

The accuracy of the clinicogenomic model stratified by cancer status and mass size or tumor location in the test set (n=62). “Other” refers to cases that cannot be characterized as central vs. peripheral.

	Cancer		No Cancer	
Mass Size	**n**	**Accuracy**	**n**	**Accuracy**
>3cm	9	100.0%	1	0.0%
<=3cm	5	100.0%	37	91.9%
Poorly Defined Infiltrate	3	100.0%	7	100.0%
Location				
Central	5	100.0%	3	100.0%
Peripheral	12	100.0%	17	76.5%
Other	0	NA	25	100.0%