April 25, 2019

Early diagnosis of non-small cell lung cancer (NSCLC)

Current Project Status and Proposed Next Steps

Summary

This report documents the current status of the non-small cell lung cancer (NSCLC) detection project using OTraces' proprietary, patented detection method. The method uses tumor microenvironment active proteins coupled with Artificial Intelligence (AI) to produce superior predictive power. Also, based upon these results we propose follow-on studies to improve the model by a more inclusive protein and gene survey, including sputum analysis, and increasing the statistics to refine and improve the current model. This current study analyzed stage 1 through 4 samples of non-small cell lung cancer, compared against not cancer, putatively healthy samples. This detect training set model produced 97% predictive power for detection of non-small cell lung cancer (99% sensitivity, 97% specificity) for a data set of about 200 samples.

The focus of the follow on study, discussed herein, is to move the focus to earlier stage NSCLC including stage 0 and pre-cancer lesions. Also we plan to include proteins from sputum samples as well as DNA and more importantly mRNA markers. Recent studies have identified sputum and lung tissue derived proteins and mRNA with pre-cancer lesions. An unresolved question is the availability of lung tissue derived biomarkers in either serum or sputum. We intend to study this. A key part of this study is to develop a source for these early stage cancers and pre-cancer lesions.

1.0 Significance, Background

Early diagnosis of non-small cell lung cancer (NSCLC) is difficult as there is no accepted method for detection prior to presentation with symptoms. Usually symptoms of lung cancer do not appear until the disease is already at an advanced stage, 3 or 4. Even when lung cancer does cause symptoms, many people may mistake them for other problems, such as an infection or long-term effects from smoking. This may delay the diagnosis. Imaging methods such as X-Ray, CT Scan and MRI's have been evaluated for early detection, but have suffered either from high cost for screening or are prone to find nonmalignant abnormalities that cannot be definitively diagnosed as cancer without further invasive medical procedures. As a result, none of these are used for general screening, and are employed only at patient presentation with symptoms.

The primary types of lung cancer are small cell cancer (SCLC) and non-small cell cancer (NSCLC). NSCLC is in turn broken down into three main types: adenocarcinoma, squamous cell carcinoma and large cell carcinoma. NSCLC can often be treated by surgery because it is slower to metastasize than SCLC and is often still localized at the time of diagnosis. SCLC is very aggressive and 60 to 70% of patients have metastases at the time of diagnosis. About 13% of all lung cancers are SCLC and 87% are NSCLC.

Lung carcinomas are typically diagnosed when they are in a late stage, which leads for their poor prognosis. Even the earliest stage (T1N0) patients have disseminated disease between 15 to 30% of the time. Generally, diagnosis is needed since this will offer highest opportunity to increase the survivability. Unfortunately, once the primary lesion has metastasized the prognosis is extremely poor. Therefore, 5 years survivability of lung cancer is only 13%. A significant reason for such a high death rate and low survival rate is that currently cancer is detected after it metastasized, *i.e.* lung cancer mostly spread beyond primary site at the time of detection. The five-year survival rate for early state of peripheral squamosa module of 20 mm or less is as high as 88%. This emphasizes the importance of early detection of lung cancer, which may significantly increase the survival probability.

A noninvasive low-cost method suitable for surveillance patient monitoring or high-volume automated screening would dramatically improve the medical outcome of these patients. OTraces method described herein has these attributes. The method uses tumor microenvironment active proteins combined with Artificial Intelligence (machine learning) to produce a simple low cost protein immunoassay method that achieves predictive power in the 90%+ range (97% for NSCLC). These assays can be deployed on high volume main frame immunoassay systems already widely installed in Clinical Laboratories around the world. High profitability is achievable even under very heavy downward cost pressures endemic in the *in-vitro-diagnostic* market.

2.0 Summary

2.1 Current Status Results

OTraces cancer diagnostic method has shown superior performance in early stage detection of breast cancer as well as NSCLC and aggressive prostate cancer. In blinded third party validation trials the breast cancer test showed 100% accuracy in detecting 26 stage 1 breast cancer samples. Overall the breast cancer validation study showed 96% correct detection of 205 breast cancer positive samples ranging from stage 0 to 4. Overall predicative power of the method for detection of breast cancer versus not cancer, i.e. putatively healthy controls, included 410 samples, was 97%.

The NSCLC proof of concept project showed 100% accuracy in detection of 25 stage 1 samples and 99% accuracy in detection of all 103 cancer samples with stage range from 1 to 4. The overall study of 199 samples; ½ NSCLC positive and the other ½ not cancer controls showed 97% predictive power. It should be emphasized that the method was shown to be able to detect stage 1 NSCLC with very high accuracy. Also, the biomarkers showed strong action in the transition from not cancer to stage one NSCLC which indicates that stage 0 NSCLC may be detectable.

The lung cancer test was internally validated by OTraces using samples derived from the team of **The Department of Tumor markers, M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Roentgena 5, PL-02781 Warszawa, Poland.** This study did not have sufficient samples to develop the predictive model and process a separate blind validation sample set. The training set requires a minimum of about 100 positive and 100 negative samples. In order to validate this smaller data set, the samples were subjected to a statistical process that OTraces calls boot strapping. In this case, one sample is removed from the training set group and the model is

rebuilt without it. This removed sample is then scored as "blinded". This is then done for each sample in the data set. The resulting predictive power results from this removed samples "blinded" scoring.

OTraces method involves a mathematical processing technique involving artificial intelligence and machine learning. The process involves anchoring each biomarker by the mean value for the biomarker for both the positive and negative samples. These mean values are also determined by the patient's age. The mathematical process then normalized the drift in mean values by age and compresses the data set anchored at the determined mean values. This new normalized and compressed data set comprises the independent variables in the correlation that is done by what OTraces terms spatial proximity scoring. The data set is plotted in multidimensional orthogonal space (in this case, 5 dimensions) and unknown samples are scored by proximity to the training set.[1]

2.2 Proposed Next Steps, Project Summary

The proposed next step project goals are to greatly improve the biomarker survey to include additional biomarkers that may be probative based upon results of the data from the current project described wherein, and those that may be of interest from new scientific literature. **The focus of this, more inclusive biomarker survey, is to focus on stage 0 and pre-cancer lesions**. **We would include proteins and genomic markers (RNA) in the survey. Also samples from sputum would be considered as candidates. Biomarkers found in lung tissue would also be considered to the extent they are available for analysis in non-invasive fluids (serum sputum).**

Additionally we propose to increase the training dataset to 400 samples, 200 cancer and 200 non cancer. This larger training data set will be used to refine the training set model. Finally, this model will be validated with a separate blind validation data set independent of the training set.

3.0 Innovative, Cancer-relevant Technology

The OTraces cancer diagnostic method reduces false negative and positive rates dramatically (> 96% predictive power shown in third party validations). The method is a simple blood test (proteomics panel) that represents a cancer detection platform that will work on any solid tumor cancer. The method uses simple, inexpensive, ELISA assays suitable for high volume screen testing in the clinical lab. These tests measure the concentrations of several, 5 or 6, very low level signaling proteins. These proteins fall within several cytokine functional groups --- pro-and anti-inflammatory, angiogenesis, anti-tumor genesis, vascularization and a cancer tumor marker. The biomarkers used for NSCLC were, pro-inflammatory IL 6, Anti-tumor TNF-RI, angiogenesis IL 8, vascularization VEGF, and a colony stimulating factor GCSF. A tumor marker was not used in this study. These functional groups are important in that they have separate distinct actions that when combined are a clear indication of the presence or absence of cancer. The measurement of these concentrations does not in and of itself yield high predictive power. The method also uses other physiological parameters (e.g. age or body mass index) to produce a cancer score that is highly predictive.

[1] For more information see "Diagnostic Proteomics – A New Approach- White Paper" Available from OTraces.

3.1 Substantial Improvement and/or New Capabilities

Table 1 below shows cancer predictive power performance for OTraces' method for various cancers. Note that different tumor markers were used for each cancer and the selection of the cytokines varied, but the functional groups noted above were used in all cases. Breast and Prostate cancer are an exception in that the tumor markers for both are the same.

Condition	Status	Cohort Size	% Correct	Test Location
Prostate Cancer Screening	Cancer	60	95.0%	Johns Hopkins Lab[1]
	Not Cancer	180	87.0%	
	Cancer	111	96.4%	OTraces Lab
	Not Cancer	148	96.6%	
Prostate Cancer Aggressive Vs. Non Aggressive Surveilance	High Gleason Score	160	96.0%	OTraces Lab
	Low Gleason score	111	89.0%	
Breast Cancer	Cancer	200	97.0%	Gertsen Inst. Moscow[2]
	Not Cancer	207	96.6%	
	Cancer	651	96.9%	OTraces Lab
	Not Cancer	529	97.5%	
Ovarian Cancer	Cancer	101	96.0%	OTraces Lab
	Not Cancer	111	99.1%	
Melanoma	Cancer	172	98.3%	OTraces Lab
	Not Cancer	172	97.7%	
Lung Cancer	Cancer	96	100.0%	OTraces Lab
	Not Cancer	96	97.9%	

1. Third party validation trial at JHU under Dr. Kenneth Pienta

2. Third party validation at gertsen Institute, Moscow

Table 1, Comparison of Cancer Detection Predictive Power

3.1.1 Historical Perspective

Note that previous research has progressed from the search for the single protein marker and DNA markers without success for screening applications, as predictive power is insufficient. PSA for prostate cancer screening remains the only protein biomarker approved albeit with poor predictive power. Proteomics was next pursued, in the form of mass spectrometry without success, the primary problems being lack of sensitivity for low-level proteins, oversampling and the fact that the mass spectrometry is too complex for very high volume clinical lab production. Limited success has been found with high level proteins (HAPs) measured with ELISA methods in the form of helping make treatment decisions (ovarian cancer). OTraces method focuses on very low level (difficult to measure) signaling proteins. Also, the correlation methodology will not yield adequate predictive power without including physiological parameters such as patient age. OTraces method uses a transforming method that embeds the age effect as well up down regulation non-linearities inherent in the low level signaling proteins in its transformation of concentrations into new variables used in the spatial proximity correlation analysis (see Discussion of the Technical Method reference 1).

The experimental data shown in Table 1 above was run in OTraces laboratories, except for the Russian market clearance breast cancer validation study, conducted at the Gertsen Institute,

Moscow, and the prostate cancer screening test conducted the Johns Hopkins Medical Center, Brady Urology Institute. In the breast cancer study, the Not-Cancer cohort was comprised of anonymous donors with the population characteristic of women presenting for annual physical exams without screening out any non-cancer abnormalities, aged 35 to 75. The cancer cohort was positive in mammography and biopsy for breast cancer and serum was drawn prospectively for this study.

Likewise, for prostate cancer the Not-Cancer cohort was comprised of anonymous donors with the population characteristic of men presenting for annual physical exams without screening out any non-cancer abnormities, aged 50 to 75. The cancer cohort was drawn retrospective, from the Johns Hopkins prostate cancer serum bank

In both cases, the blinded samples were run at the clinical laboratory of the medical facility by medical technicians. OTraces supplied equipment, reagents and training for the operators. The sample concentration measurements and patient age were blinded and scored by OTraces. Final results were un-blinded by the medical institution.

4.0 NSCLC OTraces/Sklodowska-Curie Memorial Study

4.1 Patients

The study comprised 103 previously untreated patients with NSCLC (77 males and 26 females) with age between 34 and 81 years (median = 63 years). The patients were referred to the Institute of Oncology, Warsaw. The samples are from the Institute serum bank. The cohort consisted of 26 stage 1, 10 stage 2, 48 stage 3 and 19 stage 4 lung cancers. There were 35 adenocarcinoma and the remainder were either squamous cell carcinoma or large cell carcinoma. About half of the patients had undergone surgical resection but after blood draw. Clinical staging and histologic typing were performed on post-operative material. In the other patients (51%), the staging included clinical examination, chest radiography, bronchoscopy, chest and brain CT, abdominal Ultrasound and radio-isotopic bone scans. The patients were followed for up to 6 years (median 580 days) after blood sample was drawn.

The not lung cancer cohort enrolled in the study were, patients who presented no clinical manifestation of infection and no radiological or clinical features of obstructive pneumonia (N=96) cohort. Almost all of the patients were former cigarette smokers.

No data were available for age for the not LC cohort in this study. This will degrade the correlation somewhat. Age was supplied for the LC cohort, however, age, used as a meta-variable to adjust the concentration values in the correlation is focused on the difference between the mean value for not cancer versus the mean value for cancer as it changes with age. The noise suppression also is focused on the "signal", again this is the difference between the not cancer and cancer mean values for each year of age. Since the data did not have age values for the not cancer cohort, we assumed all samples were the same age, set at 60 years for age the software, though the actual value used is not important.

4.2 Measurement methods

Samples of blood were drawn before treatment. The sera were separated within 1 h after blood collection; the coagulation was in room temperature. The level of biomarkers were measured by ELISA with the use of commercially available kits from R&D Systems, Minneapolis, Minn. USA. Measurement equipment used was the semi-automated OTraces LHS Immunochemistry System

4.1 Protein Survey

Eleven proteins were surveyed as candidate biomarkers for this study, see table 2 below. Due to project limitations, the survey was limited to the eleven shown in the table. The best biomarkers surveyed for up regulation were: IL – 6, pro-inflammatory; IL – 8, Angiogenesis; VEGF, vascularization; TNF – Ri, Anti-tumor genesis; and G-CSF a colony-stimulating factor. Also due to reagent limitations, TNF alpha was not included and thus TNF Receptors were surveyed. The role of colony stimulating factors, though somewhat uncertain, seems to be related to stimulating angiogenesis or immunosuppression. Thus, secretion of such proteins as GCSF, bFGF and M-CFS may be of interest as biomarkers for NSCLC and were surveyed.

Tumor-markers (markers not functional in immune response or tumor vascularization) were not included in the panel of biomarkers surveyed, again due to reagent and project limitations. Various molecules detectable in the serum, useful as putative markers of the disease may include chromogranin A (CgA), pro-gastrin releasing peptide (ProGRP) and neuron-specific enolase (NSE; an γ-γ isoform of the ubiquitous enolase enzyme), cytokeratin 19 marker CYFRA 21-1 etc. Also tumor markers such as Chromogranin A (CgA), CEA, SCC, and CA 19-9 may be useful tumor markers.

Protein	Function	NSCLC patients			Controls		Ratio
		median	range		median	range	
bFGF	Fibroblast Growth Factor	3.5	0.5 - 44.2		2	0.5-4.4	1.75
G-CSF	Granulocyte Colony Stimulating factor	30.4	7.0 - 144		21.2	7.0-34.0	1.43
IL-1O	Anti or Inhibiting Inflammatory Factor	5.8	2.0 - 620		3	2.0-7.3	1.93
IL-6	Pro-inflammatory Factor	12.75	0.7-111		0.98	0.7-2.4	13.01
IL-8	Angeogenesis	31.52	10.0-752		10.15	10.0-12.5	3.11
IL-IRA	Anti or Inhibiting Inflammatory Factor	13	98 - 2,397		248	98-473	0.05
M-CSF	macrophage colony-stimulating factor	11	232 -3,034		457	100-600	0.02
sIL-2R	Inhibiting Inflammatory Factor Self Regulation	3,488	758-8,148		1,112	758-1,753	3.14
TNF RI	Anti-Tumor Receptor	1,500	749-4,124		850	600-1,200	1.76
TNF RII	Anti-Tumor Receptor	3,535	1,650- 8,206		2,265	1,577-3,500	1.56
VEGF	Vascularization	633	56-2,000		125	9.0-325	5.06

Table 2, Biomarker Survey

The choice of active biomarkers is not related to just the degree of up-regulation in the transition from not cancer to early stage cancer, the functionality of the protein is very important. Note that using two proteins that denote the same function or action of the tumor with both having a high degree of up-regulation may be not useful as the second protein indicative of that function will plot as a 45 degree line against the first protein in multiple dimensional space. Thus the

second protein adds little or no predictive power. Thus we test biomarkers in multiple dimensional space and only use one tumor marker for improved predictive power.

As noted a key goal of the proposed next step research project is to expand on the surveyed biomarker list based upon these results to complete a thorough list.

4.2 Biomarker Action by Stage

Figure 1 below shows all eleven surveyed biomarkers action by NSCLC stage. The five chosen are noted at the top of the chart legend. Also their actions are plotted as solid lines for clarity. Note that three biomarkers IL 10, bFGF, and MCFS all showed strong response but only in later stage cancer. These biomarkers to more or less degree are related to immune suppression.

Note that the number of samples within each stage were not equal with 26 stage 1, 11 stage 2, 48 stage 3 and 18 stage 4 in the supplied cancer positive cohort. Thus, the degree of overall up-regulation shown in table 1 cannot be used alone for selection. The action at early stage is far more important.

BioMarker	Healthy	Stage 1	Ratio
IL 6	1	3.7	3.70
IL 8	8	16.2	2.03
Gcsf	10	25.8	2.58
vegf	125	500	4.00
TNF Ri	850	1362	1.60

Table 3, Up Regulation from Healthy to Stage 1

Up-regulation from Healthy (-1) to Stage 1 is shown in Table 3, to the left. Note that all showed marked increases by a factor of 2 or more except TNF Ri. OTraces experience with TNFα rather than its receptor indicates that it will likely show a high degree of up-regulation at early stage onset of the disease. **Though no stage 0 samples were available in this data set, the data strongly indicated that stage 0 may be detectable**. It should be noted the method has shown capability to detect stage 0 breast cancer on true stage zero samples detected and verified by imaging and biopsy, in blinded validation (Gertsen Institute Moscow).

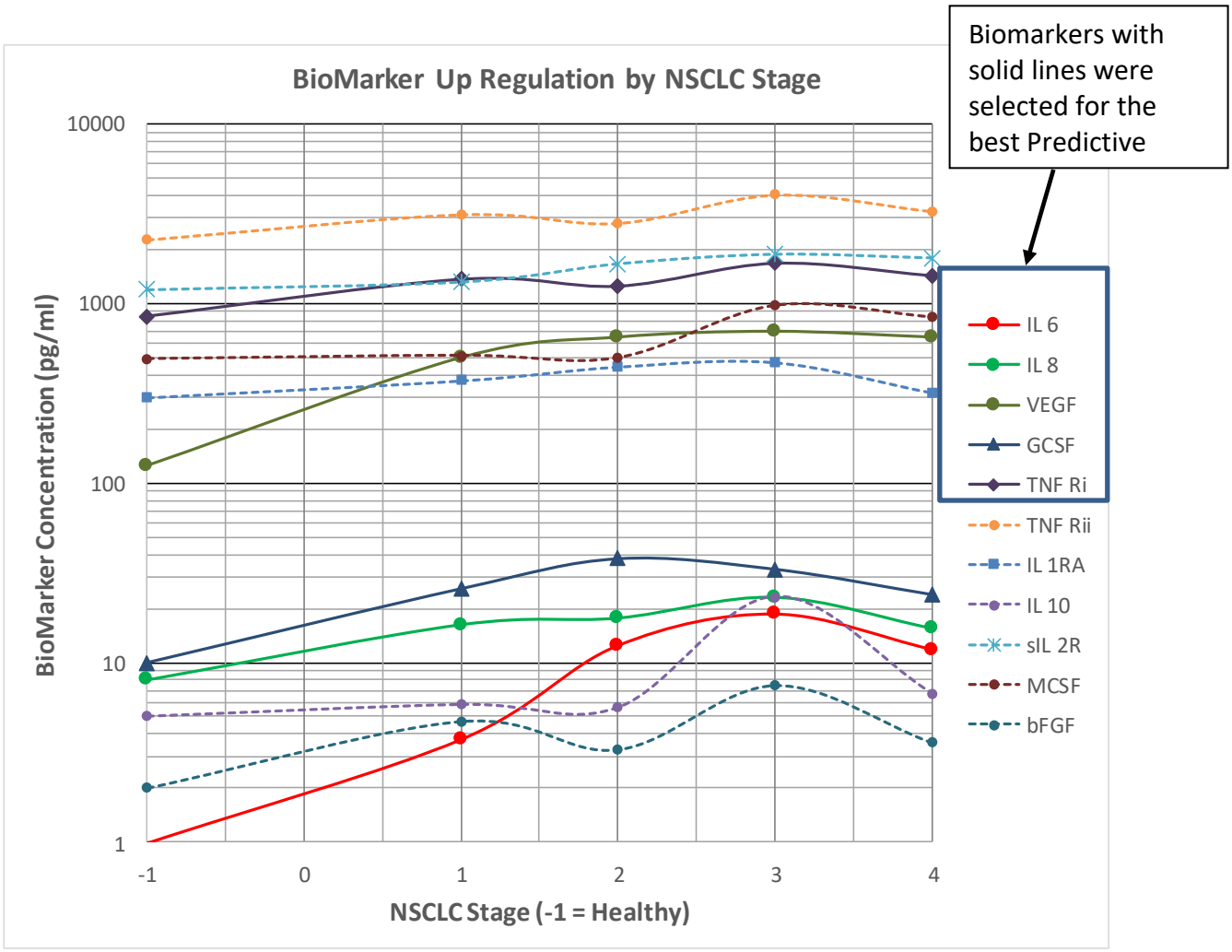


Figure 1 Biomarker up Regulation by Stage

These results suggest that IL-6, IL-8, IL-10, VEGF and GCSF are especially interesting biomarkers. IL-6, (pro-inflammatory) and TNF RI TNF RII (anti-tumor) represent immune system reactions to the tumor. IL 10 also known as human cytokine synthesis inhibitory factor (CSIF), is an anti-inflammatory cytokine. IL 8, though its functionality is complex is associated with angiogenesis in cancer. VEGF is associated with vascularization of the bulk tumor in cancer. Note that choosing biomarkers based simply on the single dimensional up-regulation, though helpful, may not produce the best correlation. OTraces proteomic noise suppression tends to bring out separation of the biomarkers as they transition from not cancer, to cancer. When plotted in the multidimensional spatial proximity correlation method, this separation information is retained. As noted above, two biomarkers that measure the same tumor or immune system action will plot

at as a 45 degree line in a two dimensional projection of the space. Thus choosing biomarkers with orthogonal, or unrelated functionality is preferred.

For this first-pass analysis, IL 6, TNF RI and RII, IL 8 and VEGF were chosen as the functional biomarkers. The colony stimulation factor chosen was GCSF. This colony-stimulating factor seems to be active even in the early stages of tumor development unlike the others. The supposed anti-inflammatory actions of the IL 10 cytokine seems to only be secreted by the tumor in to the tumor microenvironment in late stage NSCLC, as was shown by the other colony stimulating factors. This same response is seen in breast and prostate cancer. Thus, for early detection this biomarker in not suitable. OTraces has seen this immune suppression in prostate and breast cancer in later stages of the disease. In these cases the most active biomarker is the anti-inflammatory cytokine IL 10. This late stage action is not useful for early stage detection. In the case of aggressive prostate cancer, this later stage action is very useful for detection of the transition from quiescent low Gleason Score cancer to the aggressive form.

The best biomarkers for early stage detection are IL 6, IL 8, VEGF, TNF Ri and GCSF. This five were used for NSCLC Model development.

4.3 The Non-Small Cell Lung Cancer Model

The training set model is derived from the data set after the raw concentrations are mathematically processed. The mean values for each biomarker for the lung cancer cohort and the not cancer cohorts are determined first. These mean values are determined as they drift with age. The raw concentration values for each sample in each cohort is then compressed using several different compression algorithms with the mean values and the derived mid-point between the cancer and not cancer mean as the anchor points. The two mean values and the mid-point (not cancer to cancer transition point forms four zones of compression.

The compression algorithms may be different for each of the zones. Raw concentration data points at these means are transitioned to the new compressed independent variable and are adjusted so that all points at the means will be converted to the new variable at the same value. Thus mean value drift with age is zeroed out when the conversion to the new compressed independent variable (age drift is normalized by a meta-variable method) is done. These new independent variables are then plotted in multidimensional space and the unknown samples are scored by proximity to the training set in this multi-dimensional orthogonal plot. The scoring is an arbitrary range of 0 to 200 with 0 to 100 being not cancer and 101 to 200 being cancer. Unknown samples that only “see” not cancer in proximity will be scored 0 and those that only see cancer training set samples in proximity are scored 200. In between scoring is simply a count of each side (see reference 1 for more information).

Figures 1 and 2 shows the results of the correlation. The graphs show two cancer score parameters, so called Cancer Score Linear, CSL, and Cancer Score Quadratic, CSQ, These are different calculations of the same thing, the cancer score. The 5 dimensional grid (5 biomarkers) is cut into two-dimensional planes for the computation, bi-marker planes. This produces 15 bi-marker planes and the plane axes are segregated into several thousand (typically 2000) grid points. The bi-marker plane (2 D Plane) thus has 2000 x 2000 (4 million) grid points to score. This is done to facilitate

computer computations. Each bi-marker plane has the training set plotted on it then each point in the remaining two dimensional grid is scored by proximity to several, 3 to 5, nearby training set points. If the shortest distance from the grid point of interest is to not cancer training set points it is scored as not cancer, 0 and vice versa as +1. A blind sample is scored by this method for each bi-marker plane grid point. The internal predictive power of the training set itself is scored in the same way except the actual diagnosis of the training set point being scored is ignored.

Note that the values plotted on the individual orthogonal axes are not concentration, but rather concentration after noise suppression and age normalization. The new independent variable is called proximity score.

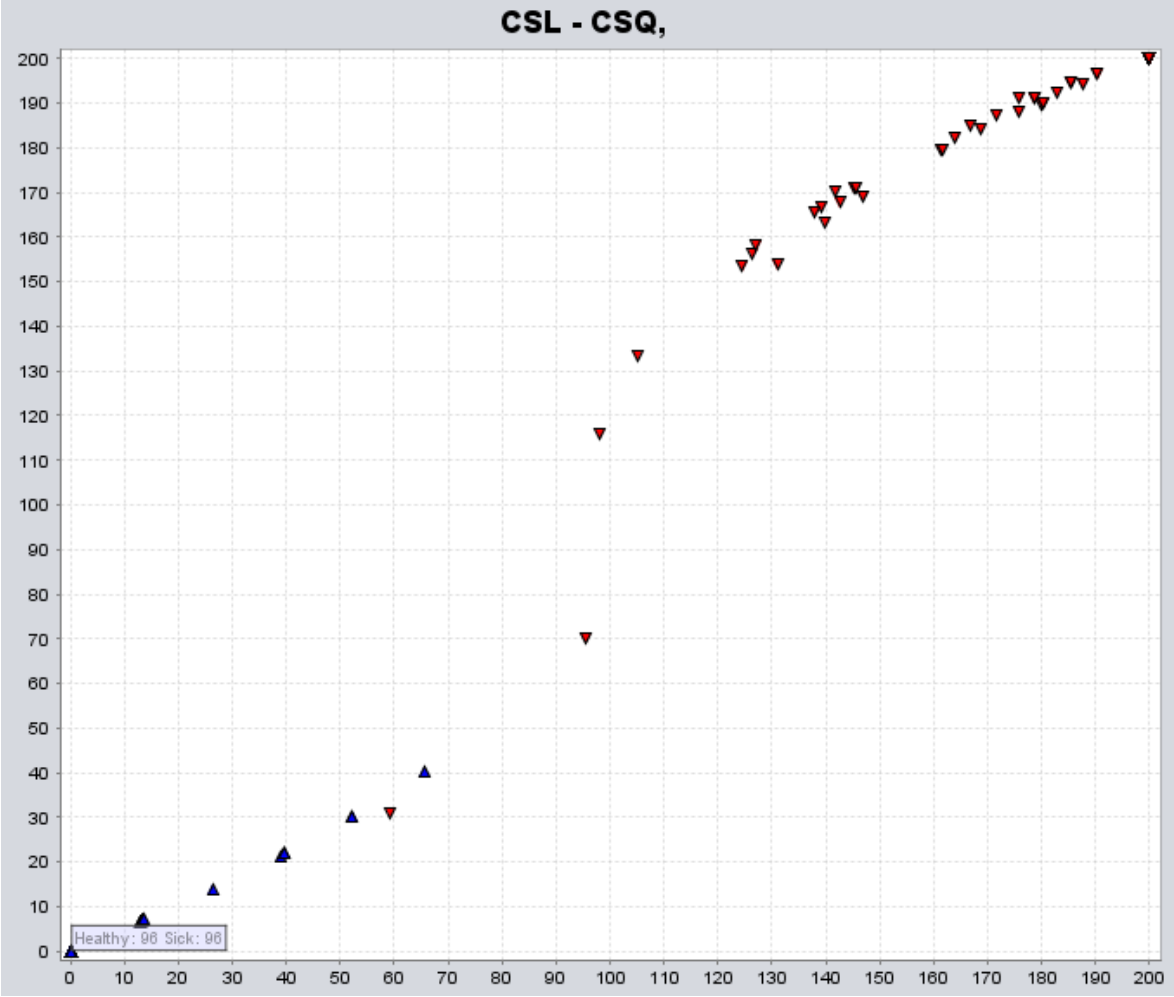
These multiple bi-marker planes are then added up for a composite score. **Cancer Score Linear CSL** is simply the sum of each bi-marker plane result (0 or +1) multiplied by the predictive power of that bi-maker plane. The **Cancer Score Quadratic CSQ** is simply the square root of the sum of the squares of these individual bi-marker plane scores. The CSL value is the reported value for diagnosis. The CSQ value is used for testing individual blind samples for topology instability. OTraces has patented a method testing individual blind sample points for topology instability and correcting them. The CSL/CSQ plot are also an easy, eyeball, way of seeing the resulting predictive power.

This scored set of grid points across all bi-marker planes constitutes the training set model.

Figures 2 and 3 below shows the results for algorithm I and II respectively. Algorithm I uses each biomarker's proximity score as a single orthogonal dimension in the spatial proximity correlation, 6 in this case. Algorithm II used the ratio of the biomarker proximity scores, 15 orthogonal dimensions in this case. The final result is scaled such that the range goes from 0 to 200.

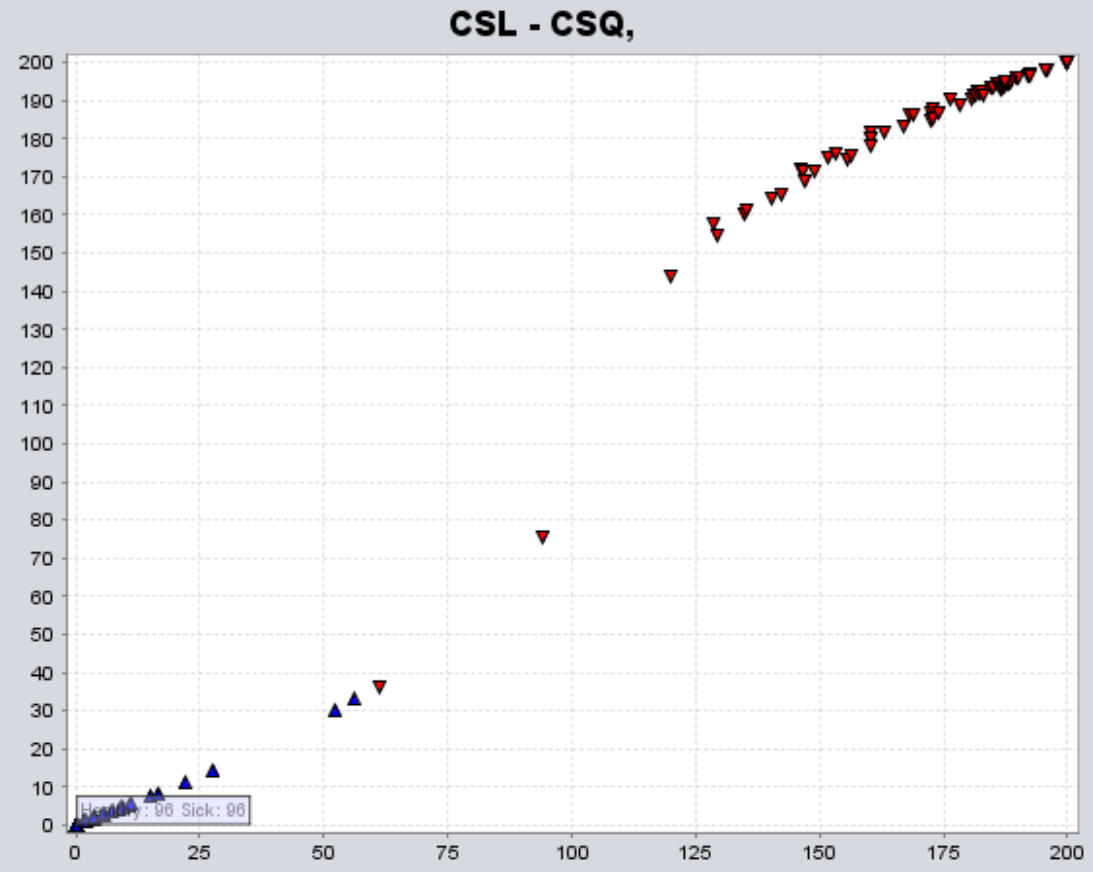
Note that this process of cutting the 5-dimensional gird into orthogonal bi-marker planes cuts computation time. The 5-dimensional grid would require 32×10^{15} grid points to be scored if this bi-marker planes projection were not used. The bi-marker plane simplification requires 20 million grid points to be scored for the 5 dimensions.

Figure2
Lung cancer samples from Poland N = 103
Not Cancer samples, anonymous N = 96
Algorithm 1
(Note that multiple samples are piled up on CSL, CSQ = to 0 and CSL, CQS= 200
Red Inverted Triangles are lung cancer Blue Triangles are not Lung Cancer)



Note that 3 samples out of the total of 103 NSCLC samples scored less than 100 and none of the NOT NSCLC samples scored over 100. Note that this is the internal predictive power of the training set.

Figure 3
Lung cancer samples from Poland
Algorithm 2
(Note that multiple samples are piled up on CSL, CSQ = to 0 and CSL, CQS= 200
Red Inverted Triangles are lung cancer, Blue Triangles are not Lung Cancer)



Note that 2 samples out of the total of 103 LC samples scored less than 100 and none of the NOT NSCLC samples scored over 100. Note that this is the internal predictive power of the training set.

We would expect some reduction in predictive power with a blind validation set. Note that a 100/100 training set cohort for breast cancer produced substantially the same predicative power for the training set (internal) and the blind validation set. Also the addition of age values for the complete cohorts will improve the predictive power. And the training set is unbalanced. Ideally the training set should have equal numbers of each cohort being separated. This assures the training set model will not be biased. Correcting this should also improve predictive power slightly.

This model produced 97% accuracy for the cancer samples and 99% accuracy for the not a cancer samples. The separation was very good. The cancer samples ranged from a score of 200, 30

samples to 60 for one sample. This sample was a Stage 2 cancer but, displayed a low pro-inflammatory and VEGF response. The not cancers scored 0, 72 samples, and the highest incorrectly called not cancer scored 120.

The cancer/not-cancer cohorts were too small to develop both a training set and a separate validation set. At least 100/100 cancer/not cancer samples (preferably 200/200) are needed to produce a training set that statistically represents to population in general. The minimum training set size is based upon the industry rule one thumb of about 25 samples per biomarker to reduce oversampling problems. At 25 samples per biomarker 125/125 is indicated. Thus the full sample cohort was used for the training set without a traditional validation set.

In order to validate the model a method called boot strapping was used. In this method one sample is removed from the training set and the model is re-built without it. This single sample is then scored "as blind" with the reduced model. This is then done for all samples in the data set. The result was very consistent as the in training set samples scoring did not change in the majority of cases and if change only by a few counts. All of the missed called samples remained as miss calls. The sensitivity and specificity published above is from the boot strapping results.

5.0 Proposed Research Project.

5.1 Goals

As noted the key goal of this study is to focus on early stage detection including stage 0 and pre-cancer lesions. Thus sample from sputum and lung tissue will be considered. Also genomic biomarkers will be included.

5.2 Biomarker Survey

The Protein survey will involve a thorough literature search, several months, to build the candidate biomarker list. The evaluation will also include testing for availability of the candidate biomarkers in convenient non-invasive measurement fluids such as sputum and serum, as a simple easily deployed test cannot involve tissue sampling using navigational bronchoscopy (R-EBUS, and TTNA) or other more invasive sampling methods.

Focal adhesion kinase and C-terminal Src kinase have been associated with pre-cancerous lesions[2]. The proteins were found in a mouse model tissue, the question is the levels in human serum or sputum. Also mRNA biomarkers have been associated with precancerous lesions in humans[3]. There are a number of additional recent publications on biomarkers associated these pre-cancer lesions and early stage NSCLC.

[2] Early Candidate Biomarkers of Non-Small Cell Lung Cancer Are Screened and Identified in Premalignant Lung Lesions.;Nan Y1, Du J2, Ma L3, Jiang H3, Jin F3, Yang S1.;Technol Cancer Res Treat. 2017 Feb;16(1):66-74. doi: 10.1177/1533034615627391. Epub 2016 Jul 8.

[3] Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions; Jennifer E. Beane, Sarah A. Mazzilli, Joshua D. Campbell, Grant Duclos, Kostyantyn Krysan, Christopher Moy, Catalina Perdomo, Michael Schaffer, Gang Liu, Sherry Zhang, Hanqiao Liu, Jessica Vick, Samjot S. Dhillon, Suso J. Platero, Steven M. Dubinett, Christopher Stevenson, Mary E. Reid, Marc E. Lenburg & Avrum E. Spira; Nature Communicationsvolume 10, Article number: 1856 (2019)

5.3 Sample Collection

Once the candidate biomarkers are selected about 800 samples will need to be collected, 400 for the training set. About one half being cancer positive and one half cancer negative. The training set needs to be balanced with ½ 200 being positive and ½ being negative. A key aspect of this study is to develop sources for early stage 0 and 1 samples and samples from pre-cancer lesions. If the selected biomarkers include a sputum source this will also be needed for both positive and negative cohorts. These samples will tend to be difficult to collect and may involve invasive collection methods such as lung tissue sampling.

5.4 Refine Training Set Model

This will require running about 400 ELISA assays. The project will require about 2 months. A Mathematician and Software Engineer will be needed for this phase. In addition these will be needed throughout the project to evaluate various different compression/ normalization algorithms using current data and finally the new data set compiled on this project.

Inclusion of continuous (e.g. concentration levels) biomarkers derived from sputum will be handled seamlessly by the analytical method. The samples source is immaterial. Inclusion of quantitative genomic biomarkers is also seamless. Handling a multiplicity of binary genomic biomarkers (Yes or No) is not easily handled by the spatial proximity correlation method as they plot in multi-dimensional space as two piles of data points on each respective axis. No discernable pattern for the machine learning algorithm to operate on, can be developed in the spatial coordinates. Augmenting the method by the use of statistical techniques, such as Bayesian Statistics for these biomarkers though will improve predictive power. The current cloud deployed analytical software has the hooks for inclusion of such methods.

5.5 Run Validation Set

This will require running about 300 ELISA assays. The project will require about 2 months.

5.6 Goal Summary

The goals of the follow on Non-small cell lung Cancer research project are:

1. Improve the size and detail of the biomarker protein data base. We propose to review the literature and add possible biomarker candidates in the following functional groups:
    a. Pro-inflammatory, only one, IL 6, has been surveyed at this point. Other candidates would be from the cytokine group such as IL 1 as well as other inflammatory markers such as CRP
    b. Colony Simulating Factors, G-CSF, and M-CSF were surveyed in this study. The literature indicates others may be of value.
    c. TNF receptors versus TNFα, Only TNF receptors were surveyed in this study. TNFα needs to be included.
    d. A number of tumor markers, proteins not implicated in functional aspects of cancer tumor growth or suppression (immune system), are implicated in NSCLC. These should be surveyed and at least one be considered for inclusion in the test panel.
    e. Most importantly the protein Survey should include biomarkers associated with pre-cancerous lesions.

f. **A number of gene mutations have been associated with pre-cancer lesions, these have been found in sputum. Proteins associated with these mutations should be included in the survey**.

2. Increase the training set to 200/200.
3. Refine the training set model.
4. Run a true validation set separate from the training set.
5. Include sputum samples in the training set models.

Keith Lingenfelter
OTraces, Inc.
Chief Executive Officer
301 529 3824
Keith.lingenfelter@otraces.com