CENTRO DE INVESTIGACIÓN Y ASISTENCIA EN TECNOLOGÍA Y DISEÑO DEL ESTADO DE JALISCO, A. C.

“Obtención de un panel de proteínas séricas optimizado para la discriminación de pacientes con cáncer de pulmón e individuos control mediante métodos de análisis multivariado”

TESIS

que para obtener el grado de

MAESTRO EN CIENCIA Y TECNOLOGÍA

en la especialidad de

BIOTECNOLOGÍA PRODUCTIVA

PRESENTA:

IBQ. GISELA LEAL PACHECO

Guadalajara, Jalisco, diciembre 2011.

CIENCIA Y TECNOLOGIA

Guadalajara, Jalisco a 25 de Noviembre de 2011

Dr. Guillermo Rodríguez Vilomara
Director de Posgrado
PICYT-CIDESI
Querétaro

Los abajo firmantes miembros del comité tutorial de la estudiante **Gisela Leal Pacheco**, una vez leída y revisada la tesis titulada **"Obtención de un panel de proteínas séricas optimizado para la discriminación de pacientes con cáncer de pulmón e individuos control mediante métodos de análisis multivariado"** aceptamos que la referida tesis revisada y corregida sea presentada por la alumna para aspirar al grado de Maestro en Ciencia y Tecnología en la opción terminal de Biotecnología productiva durante el examen correspondiente.

Y para que así conste firmamos la presente a los veinticinco días del mes de noviembre de 2011:

Dr. Moisés Martínez Velázquez

Tutor académico

Dr. Rodolfo Hernández Gutiérrez

Tutor en planta

CIENCIA Y TECNOLOGIA

Guadalajara, Jalisco a 6 de Diciembre de 2011

Dr. Guillermo Rodríguez Vilomara
Director de Posgrado
PICYT-CIDESI
Querétaro

Los abajo firmantes miembros del Jurado de Examen de la estudiante **Gisela Leal Pacheco**, una vez leída y revisada la tesis titulada "**Obtención de un panel de proteínas séricas optimizado para la discriminación de pacientes con cáncer de pulmón e individuos control mediante métodos de análisis multivariado**" aceptamos que la referida tesis revisada y corregida sea presentada por la alumna para aspirar al grado de Maestro en Ciencia y Tecnología en la opción terminal de Biotecnología productiva durante el examen correspondiente.

Y para que así conste firmamos la presente a los seis días del mes de diciembre de 2011:

Dr. Humberto Gutiérrez Pulido
Presidente

Dr. Mario Alberto Flores Valdez
Secretario

Dr. Enrique Jaime Herrera López
Vocal

Dra. Alicia Del Toro Arreola
Vocal

Dr. Moisés Martínez Velázquez
Vocal

Dedicatorias

A mis padres, por ser la inspiración que saca lo mejor de mí.

A mi estimado Lic. Ignacio Carrillo, quien me ha enseñado que la lucha no tiene límite cuando se trata de sobrevivir.

AGRADECIMIENTOS

Al leer la versión final de este escrito me llevó a analizar que la magnitud del presente estudio hubiese sido imposible de alcanzar sin la participación de personas e instituciones los cuales aportaron los medios para llegar a resultados tan satisfactorios. Por ello, es para mí un placer expresar mi más sincero agradecimiento a todos y cada uno de los que a continuación, sin orden de importancia, quiero mencionar:

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico durante los dos años de posgrado. Así como al Fondo Sectorial de Investigación en Salud y Seguridad Social SSA/IMSS/ISSSTE-CONACYT 2008-1-87628, por el financiamiento con el cual fue posible la realización del presente estudio.

Debo agradecer de manera especial al grupo de médicos neumólogos Francisco Sánchez Llamas, Luz Audina Mendoza Topete, Marco Antonio Padilla Navarro, Juan Pablo Juárez Contreras, Antonio Rojas Calvillo, Paula Anel Cabrera Galeana, por su colaboración en la captación de pacientes. De igual manera, a Rosario y Socorro, muchas gracias por ayudarme en la toma de muestras y sobre todo por brindarme tanta hospitalidad y siempre una sonrisa. A todos los pacientes y donadores, porque sin su participación hubiera sido imposible la realización del estudio.

Es un gran placer poder agradecerle, Dr. Moisés Martínez Velázquez, no sólo el haberme aceptado para realizar esta tesis de maestría, sino por todo el apoyo y confianza en mi trabajo; por todas aquellas discusiones e ideas que han sido un aporte invaluable tanto para el desarrollo de este trabajo, como en mi formación en el área de la investigación. Me llevo una sólida amistad, grandes enseñanzas y sobre todo creo que ha sembrado en mí esas ganas de seguir por el camino de la ciencia. Muchas gracias Doc.

Al Dr. Mario Flores Valdez, quien además de brindarme sus conocimientos y experiencias profesionales, me ha permitido considerarlo un amigo con el que, a pesar de los regaños y diferencias, siempre se puede contar. También quiero agradecer al Dr. Jesús Cervantes por sus sabios consejos que seguro me harán llegar a viejo.

Para mis compañeros y amigos, tengo sólo palabras de agradecimiento, por tantos momentos memorables dentro y fuera del laboratorio, llenos de risa y algarabía. En especial a Saira que ha

compartido conmigo desde largas jornadas en el laboratorio hasta lo más insignificante en el plano personal, gracias güerita por hacerme el día a día. Chuck, con quien aún la conversación más ligera te deja algo provechoso. Pita porque a pesar de nuestras diferencias, tuvimos la oportunidad de compartir muchas experiencias, y sobre todo vivir de cerca la problemática de nuestros pacientes.

Para todos los miembros de la Unidad de Biotecnología Médica y Farmacéutica, van también mis agradecimientos. He de agradecer de manera especial a Fátima Ordoñez y Jesús Fuentes por su amabilidad y disposición.

Y por supuesto, el agradecimiento más profundo va para mi familia. A mis padres, René y Esther por su apoyo incondicional; a mi hermano Renee por su instinto paternal el cual siempre lo ha llevado a estar al pendiente de mí; a mi hermano Beto por ese gran don de hacerme reír incluso en los momentos difíciles; a Isela por todos aquellos consejos que te convierten en la hermana que no tuve. A todos ustedes por su capacidad de superación que los convierte en mi ejemplo a seguir. Por último, pero no menos importante, a Chinito sencillamente por llegar en el momento indicado =).

RESUMEN

Antecedentes: Actualmente no existe una prueba en sangre o suero para diagnóstico de cáncer de pulmón. Numerosos biomarcadores séricos han sido propuestos para la detección y manejo de dicha patología. En su mayoría, los biomarcadores carecen de suficiente sensibilidad y especificidad, limitando con ello su utilidad clínica. Recientemente, diversos autores han sugerido la aplicación de métodos estadísticos multivariados en la identificación de paneles de biomarcadores con el propósito de incrementar la precisión diagnóstica.

Métodos: Se evaluó la concentración sérica de 14 proteínas con valor diagnóstico previamente reportado, en un total de 153 individuos (64 pacientes con cáncer de pulmón, 60 pacientes con enfermedad pulmonar obstructiva crónica y 29 fumadores), utilizando kits de ELISA comerciales. Las concentraciones obtenidas para cada proteína fueron analizadas por estadística descriptiva y curvas ROC. Diversos métodos estadísticos multivariados como análisis discriminante lineal (ADL), árbol de clasificación (AC) y random forest (RF) fueron usados para determinar el panel de proteínas óptimo.

Resultados: Del total de proteínas, nueve presentaron concentraciones elevadas en el grupo de cáncer de pulmón (CP) respecto a los controles, con un nivel de significancia del 99% (prueba de Kruskall Wallis). En contraparte, sólo una se observó disminuida en pacientes con CP comparado con los controles, al mismo nivel de significancia. Del análisis de curvas ROC se obtuvo que el mejor biomarcador es CA125, dado que presentó mayor área bajo la curva (0.84) y valores de sensibilidad y especificidad de 83.15% y 78.13% respectivamente. Finalmente, mediante random forest se obtuvo un modelo de clasificación conformado por 3 proteínas, CYFRA 21.1, CA125 y CRP, capaz de clasificar correctamente 91.5% del total de observaciones (sensibilidad, 86%; especificidad, 94.4% y área bajo la curva de 0.93). Con el panel de proteínas se logró aumentar 11.8% (91.5%) de observaciones correctamente clasificadas comparado con CA125 (79.5%).

Conclusiones: En el presente estudio fue posible identificar un panel óptimo que consta de tres proteínas, el cual mostró mayor capacidad para discernir entre pacientes con cáncer de pulmón e individuos control, comparado con el rendimiento de cada proteína por separado. Por otra parte, nuestro análisis preliminar sugiere que el algoritmo RF presenta mayor utilidad para dilucidar un conjunto de biomarcadores que permita incrementar la capacidad diagnóstica.

ABSTRACT

Background: Currently, a blood or serum test for lung cancer detection does not exist. Many serum biomarkers have been proposed as proficient for management and diagnosis of lung cancer. However, they often lack of sufficient diagnostic specificity and sensitivity, thus, limiting their clinical utility. Recently, many studies suggest multivariate statistic methods as useful tools for identifying biomarker panels, in attempting to enhance diagnostic tests precision.

Methods: serum levels of 14 proteins with diagnostic value for lung cancer, according to currently available reports, were evaluated in 153 individuals consisting of 64 lung cancer patients, 60 chronic obstructive pulmonary disease patients and 29 smokers using commercial ELISA assays. Descriptive statistics and ROC curves for concentrations of each protein were obtained. Multivariate statistical methods (linear discriminant analysis, classification tree and random forest) were used on all proteins to define an optimized protein diagnostic panel.

Results: A total of 10 proteins achieved statistical relevance upon evaluation, at 99% significance level. The results of ROC curves analysis were obtained, showing CA125 as the best biomarker; it showed the higher area under the curve (0.84) giving 83.15% sensibility and 78.13% specificity. Finally, we obtained a classification model composed of 3 proteins (CYFRA 21.1, CA125 and CRP). Test performance characteristics for this panel increased the area under the curve to 0.93, 86% sensibility and 94.5% specificity. We observed classified correctly cases gain of 11.8% (91.5%) when using the model over CA125 (79.5% classified correctly cases).

Conclusions: In this study was possible to identify an optimal panel of 3 serum proteins, which showed greater accuracy to discriminate between lung cancer patients and controls, which predicted better than any individual biomarker. On the other hand, our preliminary analysis suggest that RF may be more useful than other multivariate methods to elucidate a biomarker set that allows increase sensibility and specificity parameters.

ABREVIATURAS

2-DE	Electroforesis de 2 dimensiones
A1ATTP	Alfa-1-antitripsina
AA	Aminoácidos
ABC	Área bajo la curva
AdenoCa	Adenocarcinoma
ADL	Análisis discriminante lineal
ApoA1	Apolipoproteina A-I
ATBA	Aspiración transbronquial con aguja
BAAF	Biopsia por aspiración con aguja fina
BATTA	Biopsia por aspiración transtorácica con aguja
CA125	Antígeno de cáncer 125
CART	Árbol de clasificación y regresión
CB	Cáncer broncogénico
cADN	Ácido desoxiribonucleico complementario
CEA	Antígeno carcinoembrionario
CHM	Complejo mayor de histocompatibilidad
CIATEJ	Centro de investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco
CPCG	Cáncer de pulmón de células grandes
CPCNP	Cáncer de pulmón de células no pequeñas
CPCP	Cáncer de pulmón de células pequeñas
CRP	Proteína C reactiva
CYFRA 21.1	Fragmento de citoqueratina 19
DNA	Ácido desoxirribonucleico
EGF	Factor de crecimiento epidermal
EGFR	Receptor de factor de crecimiento epidermal
ELISA	Ensayo por inmunoabsorción ligado a enzimas
EPOC	Enfermedad pulmonar obstructivo crónica
FBO	Fibrobroncoscopía
FD	Función discriminante
FEV1	Volumen de expiración forzado al primer segundo
FFP	Fracción de falsos positivos
FISH	Hibridación *in situ* fluorescente
FVC	Capacidad vital forzada
FVP	Fracción de verdaderos positivos
GPI	Glucosilfosfatidilinositol
HPT	Haptoglobina
IARC	International Agency for Research on Cancer
IL-6	Interleucina 6

IMSS	Instituto Mexicano del Seguro social
ISSEMyM	Instituto de Seguridad Social del Estado de México y Municipios
KLKB1	Kalicreina B1
LBA	Lavado broncoalveolar
MMP-1	Metaloproteinasa de matriz 1
MMP-9	Metaloproteinasa de matriz 9
MS	Espectrometría de masas
MUC 1	Mucina unida a membrana
NNK	Nitrosamina derivada de nicotina
NSE	Enolasa específica de neuronas
OOB	Out of bag
PCR	Reacción en cadena de polimerasa
PLCO	Prostate, lung, colorectal and ovarian cancer screening trial
ProGRP	Péptido liberador de progastrina
RBP	Proteína de unión a retinol
RF	Random forest
ROC	Receiver operating characteristics
RT PCR	Transcripción reversa y reacción en cadena de la polimerasa
SAA	Proteína amiloide de suero A
SCC	Carcinoma de células escamosas
SVM	Vectores de soporte mecánico
TC	Tomografía computarizada
TEP	Tomografía de emisión de positrones
TF	Transferrina
TNM	Tumor, nódulos linfáticos, metástasis
TPA	Antígeno polipéptido de tejido
uPA	Activador de plasminógeno tipo urocinasa
VPH	Virus del papiloma humano

ÍNDICE

ÍNDICE DE TABLAS Y FIGURAS

1. INTRODUCCIÓN

El cáncer es la principal causa de muerte en hombres y mujeres menores a 85 años, en países desarrollados, y la segunda causa de muerte en países en vías de desarrollo. Alrededor de 12.7 millones de casos y 7.6 millones de muertes por cáncer fueron estimadas a nivel mundial en el 2008, siendo el cáncer de mama en mujeres y el de pulmón en hombres, los cánceres diagnosticados con mayor frecuencia y los responsables del mayor número de muertes a nivel mundial (Jermal, 2011). La incidencia de cáncer continúa en aumento, sobre todo en países en vías de desarrollo, como resultado del crecimiento y envejecimiento de la población, incremento en el consumo del tabaco, el alcoholismo, inactividad física, y una dieta poco balanceada, entre otros factores. La relación entre tabaco y cáncer de pulmón es evidente y ha sido bien establecida, ya que el riesgo relativo se incrementa de 10 a 20 veces en fumadores, comparado con no fumadores; además alrededor del 85% de casos de cáncer de pulmón, bronquio y de tráquea son atribuibles al tabaco (Sun, 2007).

El cáncer de pulmón presenta tasas de incidencia y mortalidad similares, debido en gran medida al diagnóstico tardío. Desafortunadamente, más del 75% de los pacientes son diagnosticados en etapas avanzadas III y IV, limitando con ello las opciones terapéuticas y, dando como consecuencia un mal pronóstico a corto plazo, con una mediana de sobrevida entre 6-12 meses al tiempo del diagnóstico (Spiro, 2002). La tasa de sobrevivencia a los 5 años permanece aproximadamente en 15%, en el mejor de los casos (Rossi, 2005); por lo tanto, una detección temprana podría disminuir potencialmente la mortalidad del cáncer de pulmón.

Actualmente las técnicas de imagenología, como la tomografía computarizada (TC), son ampliamente utilizadas para la detección de neoplasias malignas en pulmón, debido a la capacidad de detectar lesiones incluso menores a 1 cm de diámetro. Estas características han llevado a proponer a la TC como un método de tamizaje en poblaciones de alto riesgo, tomando en cuenta principalmente el índice tabáquico, la edad, el sexo y la obstrucción pulmonar crónica (Lam, 2001). Sin embargo, esta técnica presenta baja especificidad debido a la alta prevalencia de nódulos pulmonares benignos, además, son necesarias las tomografías repetidas para determinar la tasa de crecimiento a través del tiempo, lo que involucra una exposición del paciente a radiaciones potencialmente nocivas, así como a costos elevados (Greenberg, 2007). Como un enfoque alternativo, se ha propuesto desde hace tiempo la utilización de biomarcadores séricos,

los cuales pudieran representar una vía rentable y mínimamente invasiva para identificar individuos con cáncer de pulmón, en etapas tempranas. De manera alternativa, los biomarcadores podrían diferenciar entre nódulos no calcificados benignos y neoplasias malignas. Sin embargo, hasta el momento ningún biomarcador ha mostrado tener la adecuada sensibilidad, especificidad y reproducibilidad para ser validado como biomarcador de detección temprana de cáncer pulmonar. Lo anterior ha conducido a la evaluación de paneles de biomarcadores, con el fin de aumentar la sensibilidad/especificidad en la detección de dicha patología.

2 ANTECEDENTES

2.1 EPIDEMIOLOGÍA

A nivel mundial el cáncer de pulmón ha sido la neoplasia maligna mayormente diagnosticada, así como la primera causa de muerte por cáncer en hombres y la segunda en mujeres, según lo reportado en el 2008. El 13% (1.6 millones) del total de casos y el 18% (1.4 millones) de muertes en el 2008, se le atribuyeron al cáncer de pulmón. En hombres, la mayor tasa de incidencia se encuentra en el Este y Sur de Europa, Norte América y Este de Asia, mientras que en mujeres se concentra en Norte América, Norte de Europa y Australia/Nueva Zelanda (Jermal, 2011). En México, durante el periodo de 1998-2004 se estimaron 45,578 muertes por cáncer de pulmón, con una edad promedio de muerte de 68 años y una relación hombre: mujer de 2:1 (Ruiz, 2007).

Existe evidencia suficiente del efecto carcinogénico del humo del tabaco. De manera global, se estima que el 80% de los casos de cáncer de pulmón en hombres, y al menos el 50% en mujeres es atribuible al tabaco (Ezzati, 2005). Comparado con los no fumadores, el fumador persistente incrementa 20 veces el riesgo de padecer cáncer de pulmón; el riesgo aumenta según el número de cigarros fumados al día, los años que lleva con este hábito, el grado de inhalación, el contenido de nicotina y alquitrán y el uso de cigarros sin filtro. El patrón geográfico de incidencia y mortalidad por cáncer de pulmón está altamente relacionado con el hábito tabáquico. En México, se ha calculado una prevalencia de fumadores en la población de 12 a 65 años, de alrededor de 26.4% (28, 526,833 fumadores aproximadamente; Ruiz, 2007).

El humo del cigarro contiene más de 3,000 compuestos, de los cuales 40 se reconocen como potentes carcinógenos. Se ha demostrado que las nitrosaminas y cetonas derivadas de la nicotina (NNK, por sus siglas en inglés), así como los hidrocarburos poliaromáticos (benzopireno, por ejemplo) inducen carcinomas pulmonares en roedores (Malkinson, 1992). Los NNK se unen a receptores colinérgicos nicotínicos, en células del epitelio respiratorio, activando la vía de señalización AKT, además de la activación de k-ras y la sobreexpresión de la DNA metiltransferasa, y la subsecuente hipermetilación de genes supresores de tumores, en neumocitos *in vitro* e *in vivo* (Lin, 2010). En cambio, los hidrocarburos poliaromáticos forman aductos de ADN que inducen mutaciones en genes supresores de tumores como p53, y con ello interrumpen la regulación del ciclo celular, la reparación del ADN y la apoptosis (Ruiz, 2004).

Aunque el tabaquismo es el principal factor de riesgo, se han identificado otras causas, tales como: factores ocupacionales (exposición a asbestos, arsénico, níquel, cromo, radón) y factores ambientales como el tabaquismo pasivo, la contaminación ambiental y los hábitos alimenticios.

En China, a pesar de la baja prevalencia del tabaquismo (menor al 4%), las mujeres presentan tasas elevadas de cáncer de pulmón (21.3 casos por 100,000 mujeres) en contraste con países europeos como Alemania (16.4) e Italia (11.4) donde la prevalencia del tabaquismo es alrededor del 20%. Esta carga relativamente elevada de cáncer de pulmón en las mujeres Chinas, se considera un reflejo de la exposición a los contaminantes aéreos provenientes de estufas de carbón sin ventilación, así como al humo de leña (Jermal, 2011).

2.1.1 EPIDEMIOLOGÍA MOLECULAR

La exposición constante y a largo plazo a factores contaminantes, como el humo del tabaco, y la susceptibilidad genética, interactúan para conducir a múltiples eventos genéticos necesarios para el desarrollo de neoplasias malignas. Así también, se han sugerido factores no relacionados con el tabaquismo, como factores genéticos, hormonales y virales (VPH, por ejemplo). El proceso de carcinogénesis involucra la iniciación, promoción y progresión del tumor:

Iniciación del tumor. Se refiere a efectos directos y cambios irreversibles en el ADN celular, como la formación de aductos, mutaciones y alteración en la expresión de genes (en particular este último involucra la activación de proto-oncogenes y/o la inactivación de genes supresores de tumor). Cuando un carcinógeno produce un evento genético que activa al proto-oncogen, éste se convierte en oncogen, el cual va desencadenar una desregulación en las vías de diferenciación y crecimiento celular, promoviendo el desarrollo neoplásico y aumentando el crecimiento del tumor. Los eventos genéticos que permiten la activación de oncogenes incluyen mutaciones por sustitución de base, translocación cromosómica y amplificación de genes. Por otro lado, un evento genético a su vez puede desactivar genes supresores de tumor, mediante la metilación de promotores por ejemplo, y como resultado se tiene una señal continua o anormal de proliferación celular y el posible crecimiento de la neoplasia. Ambos alelos del gen supresor de tumor deben mantener mutaciones para inactivar la función del producto de ese gen. El gen p53 es el más estudiado. La tabla 1 presenta anormalidades en genes involucrados en el desarrollo de diferentes tipos histológicos de cáncer de pulmón (Roy, 2008).

Tabla 1. Anormalidades genéticas específicas en cáncer de pulmón de células no pequeñas (CPCNP) y de células pequeñas (CPCP).

Anormalidad	CPCNP		CPCP
	Carcinoma de células escamosas	Adenocarcinoma	
Precursor			
Lesión	Displasia	Hiperplasia atípica adenomatosa	
Cambio genético	mutación p53	mutación KRAS (en fumadores), mutación dominio cinasa de EGFR (no fumadores)	
Cáncer			
mutación KRAS	muy rara	10-30% ł	muy rara
mutación BRAF	3%	2%	muy rara
EGFR			
mutación dominio cinasa	muy rara	10-40% ł	muy rara
Amplificación	30%	15%	muy rara
HER2			muy rara
mutación dominio cinasa	muy rara	4%	no conocida
Amplificación	2%	6%	no conocida
MET			
mutación	12%	14%	13%
Amplificación	21%	20%	no conocida
mutación p53	60-70%	50-70% ł	75%

ł Variación atribuible al perfil de tabaquismo

Promoción del tumor. Es un proceso gradual que requiere la exposición prolongada al agente promotor (carcinógeno), ocupa la mayor parte del periodo latente de la carcinogénesis y es parcialmente reversible. La promoción del tumor ocurre cuando las células iniciadoras son reproducidas selectivamente a través de factores que influyen en la proliferación celular, tales como crecimiento alterado y resistencia a citotoxicidad, dando lugar a la expansión clonal de células neoplásicas localizadas.

Progresión del tumor. Una vez transformada la lesión preneoplásica a neoplasia, la progresión del tumor requiere proliferación clonal continua de células alteradas fenotípica y genotípicamente; este estadio tardío incluye angiogénesis, invasión y metástasis.

A pesar de que la exposición al tabaco se presenta en el 90% de los casos de cáncer de pulmón, la mayoría de los fumadores crónicos no desarrolla cáncer. Esta observación sugiere que los

individuos pueden presentar variaciones en la susceptibilidad genética, ya sea polimorfismos en los genes encargados del metabolismo de carcinógenos, en genes de reparación de ADN, y otras vías homeostáticas que determinan el riesgo a desarrollar cáncer de pulmón. En Estados Unidos, estudios prospectivos de casos y controles en fumadores crónicos demuestran una mayor susceptibilidad en las mujeres a desarrollar cáncer de pulmón (Bain, 2004).

2.2 CLÍNICA Y DIAGNÓSTICO

Los signos, síntomas y el método diagnóstico empleado en individuos con sospecha de cáncer de pulmón, dependen del tipo histológico, localización y tamaño del tumor, así como del número de metástasis. Los tumores centrales o de crecimiento endobronquial causan pérdida de peso, tos, hemoptisis, sibilancias y estridor, disnea, e incluso atelectasia con o sin neumonía y absceso; por su parte, los tumores de crecimiento periférico comprenden dolor en la pared torácica o pleural en caso de derrame pleural.

La obstrucción de tráquea, compresión del esófago, parálisis del nervio laríngeo, síndrome de la vena cava superior, arritmias, obstrucción linfática, entre otros, son síntomas relacionados con la expansión regional del tumor en el tórax.

La distinción de los tipos histológicos constituye un parámetro importante en el tratamiento, evolución y pronóstico del cáncer pulmonar. Para efectos prácticos los tumores pulmonares se dividen en carcinomas de células pequeñas (CPCP) y carcinomas de células no pequeñas (CPCNP). El CPCNP se refiere al conjunto de tres histologías distintas que por sus características pueden ser tratadas de manera homogénea: carcinoma de células escamosas, adenocarcinoma y carcinoma de células grandes. Publicaciones de la IARC (International Agency for Research on Cancer) estiman a nivel mundial, una tasa de incidencia del 20% para CPCP y cerca del 9% para carcinomas de células grandes. Sin embargo, la tasa de incidencia para los demás tipos histológicos difiere según el sexo y la población; el carcinoma de células escamosas comprende el 44% de los cánceres de pulmón en hombres, y el 25% en mujeres, mientras que el adenocarcinoma comprende el 28% de los casos en hombres y el 42% en mujeres. En México, se ha reportado al adenocarcinoma como el tipo histológico más común, seguido del carcinoma de células escamosas y carcinoma de células pequeñas, siendo el carcinoma de células grandes el menos frecuente (Medina, 2000; Gurrola, 2009).

Una vez realizada la confirmación histopatológica del cáncer pulmonar, debe determinarse la extensión de la enfermedad, con el fin de identificar y separar a los pacientes candidatos a resección quirúrgica. La estadificación del CPCNP se basa en el sistema internacional TNM (véase Anexo 1). En cambio, el CPCP es usualmente estadificado como enfermedad limitada, es decir, restringida a un hemitórax con metástasis a nódulos linfáticos regionales (equivalente al estadio I-III del sistema TNM) y enfermedad extendida (estadio IV en el sistema TNM; IARC, 2004).

El procedimiento diagnóstico en pacientes con sospecha de cáncer de pulmón parte de la identificación del sitio de lesión por medio de radiografía o tomografía computarizada (TC) de tórax, para posteriormente determinar el método óptimo para diagnóstico histopatológico (figura 1). La fibrobroncoscopía (FBO), es la técnica invasiva mayormente utilizada; permite el examen macroscópico del árbol respiratorio hasta la mayoría de los bronquios subsegmentales y permite la toma de biopsia por aspiración y cepillado bronquial en lesiones centrales. Debido a la frecuente ubicación central del carcinoma de células escamosas, éste es fácilmente diagnosticado por biopsia broncoscópica, cepillado y/o citología de esputo. Para lesiones periféricas, generalmente se utiliza la biopsia por aspiración con aguja fina (BAAF) guiada por TC transtorácica.

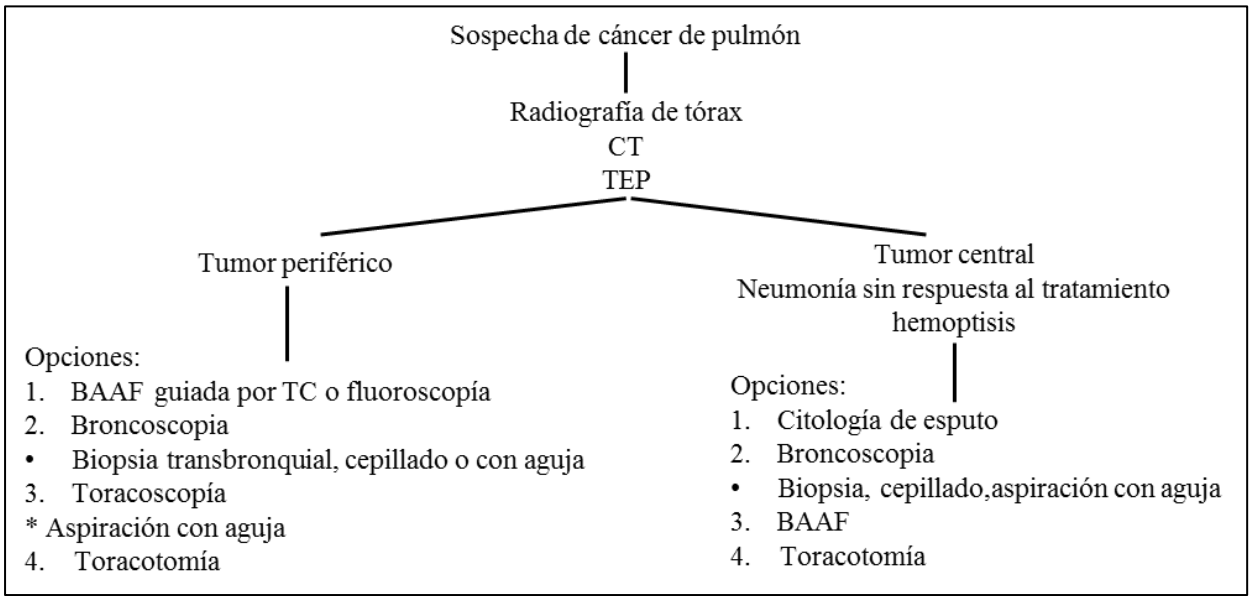


Figura 1. Esquema que indica el procedimiento diagnóstico de cáncer de pulmón, dependiendo del sitio de la lesión. CT= tomografía computarizada; TEP= tomografía por emisión de positrones; BAAF= Biopsia por aspiración con aguja fina.

Las características de rendimiento de las técnicas para diagnóstico histológico y citológico en lesiones de pulmón, dependen en gran medida del tamaño de la lesión y la localización. No obstante, ningún método diagnóstico presenta sensibilidad y especificidad de 1 (véase tabla 2; Schreiber, 2003; Rivera, 2003).

Tabla 2. Sensibilidad y especificidad de los métodos diagnósticos utilizados en la clínica para carcinoma broncogénico*.

Método diagnóstico	**Sensibilidad**		**Especificidad**	
	CB central	**CB Periférico**	**CB Central**	**CB Periférico**
Citología de esputo	0.71	0.49	0.99	
FBO				
Biopsia endobronquial	0.74			
Biopsia transbronquial		0.46		
Lavado	0.48			
Cepillado	0.59	0.52		
ATBA	0.56	0.67		
LBA		0.43		
Todos	0.88	0.69		
FBO por tamaño de la lesión				
< 2 cm	0.33			
> 2 cm	0.62			
BATTA dirigida por TC		0.88-0.92		0.96-0.98

*CB= Carcinoma broncogénico; ATBA = Aspiración transbronquial con aguja; LBA= Lavado broncoalveolar; FBO= Fibrobroncoscopía; BATTA= Biopsia por aspiración transtorácica con aguja; TC= Tomografía computarizada.

2.2.1 SCREENING (TAMIZAJE) DE CÁNCER DE PULMÓN

El screening se realiza con el fin de detectar una enfermedad en una etapa donde ésta es curable o su control es posible. Se presume que una prueba o una serie de pruebas identificarán a personas asintomáticas con la enfermedad en etapas tempranas. Idealmente, una vez establecido el diagnóstico la intervención temprana deberá cambiar el curso de la enfermedad, resultando en una disminución en la mortalidad (número de muertes específicas por la enfermedad en relación con el total de personas evaluadas). Este efecto en la mortalidad más que en la sobrevivencia es necesario para calificar a los métodos de screening como efectivos.

Las estrategias ideales de screening comparten ciertas características, incluyendo: alta sensibilidad, suficiente especificidad, aceptabilidad (método poco invasivo, con bajo riesgo a

presentar complicaciones) y rentabilidad. En los años 50, el primer método de screening utilizado para cáncer de pulmón fue la radiografía de tórax de dos dimensiones, sin embargo, los resultados no mostraron una reducción en la mortalidad. A principios de los 70s, cuatro estudios de screening integraron la radiografía de tórax con citología de esputo, dirigido a hombres fumadores mayores a 45 años. El análisis extensivo de los cuatro estudios reportaron un incremento en la incidencia de estadios tempranos de cáncer pulmonar (mayormente resecables) y aumento en la tasa de sobrevivencia a los 5 años en el grupo de screening (35%) comparado con el grupo control (15%), no obstante, no hubo diferencia estadísticamente significativa en la mortalidad atribuible a cáncer de pulmón entre los dos grupos (Patz, 2000).

Los resultados previos han conducido a la implementación de nuevas técnicas de screening, como la tomografía computarizada de baja dosis (LDCT, por sus siglas en inglés). Recientemente, seis estudios cohortes evaluaron dos grupos de screening, comparando la LDCT contra la radiografía de tórax. Los estudios concluyeron que la LDCT es significativamente más sensible en la detección de nódulos no calcificados, sin embargo, presentó una tasa elevada de falsos positivos, razón por la cual muchos pacientes fueron sometidos a procedimientos invasivos con el fin de discriminar entre nódulos benignos y malignos. A pesar del alto costo económico y la elevada cantidad de radiación a la que se expone al paciente, la LDCT es actualmente la técnica más prometedora para futuros estudios de screening (Ashton, 2005; Rossi, 2005). Lo más reciente en el seguimiento de screening de CP ha sido publicado por la NCCN (National Comprehensive Cancer Network, 2011) la cual recomienda el uso de LDCT en el screening de individuos de alto riesgo: de 55 a 74 años, índice tabáquico ≥ 30 paquetes/año, incluso si han dejado de fumar entre los últimos 15 años (ver anexo 2).

Algunas estrategias alternativas podrían contribuir en la detección temprana del cáncer de pulmón y potencialmente en el screening de dicha patología. Actualmente se investigan: el análisis inmunohistoquímico de esputo con anticuerpos monoclonales, identificación de mutaciones genéticas, metilación anormal del ADN, así como proteínas y ácidos nucleicos circulantes en suero/plasma, entre otros.

2.2.2 USO POTENCIAL DE BIOMARCADORES EN CÁNCER

El término “biomarcadores” se refiere a medidas cuantificables de homeostasis biológica que definen lo que es “normal”, proporcionando un marco de referencia para la predicción o

detección de lo “anormal” (Dalton, 2006). Los biomarcadores de cáncer no sólo son útiles en el diagnóstico temprano, también proveen información en el pronóstico de la enfermedad y respuesta a la terapia. Existen distintos tipos de biomarcadores para cáncer: genéticos, epigenéticos, proteómicos y metabolómicos.

Las técnicas comúnmente utilizadas para la detección de biomarcadores genéticos de cáncer incluyen microarreglos de ADN, reacción en cadena de la polimerasa (PCR, por sus siglas en inglés), RT-PCR, hibridación *in situ* fluorescente (FISH), etc. A pesar de las múltiples ventajas de los biomarcadores basados en genómica, éstos presentan algunas limitaciones:

a) La accesibilidad al ADN de tejido, esputo o sangre se limita a la cantidad y calidad del ADN contenido en la muestra.
b) Los niveles de ARN evaluados por RT-PCR no siempre indican el nivel de proteína presente debido a los diversos eventos que ocurren durante el proceso de traducción y modificación post-transcripcional.

Por otro lado, modificaciones epigenéticas como cambios cromosomales (pérdida de heterocigocidad, por ejemplo) e hipermetilación de regiones promotoras (silenciamiento subsecuente de genes supresores de tumor) se han propuesto como biomarcadores de cáncer. Así también, el uso de técnicas proteómicas incluyendo espectrometría de masas (MS), inmunoensayos (ELISA), inmunohistoquímica, electroforesis de dos dimensiones (2-DE) entre otras, han conducido al descubrimiento y validación de nuevos biomarcadores, incluyendo proteínas derivadas y/o asociadas a tumor principalmente. Además de proteínas y ADN, recientemente, el estudio de moléculas de bajo peso molecular o metabolitos (metabolómica) como aminoácidos, péptidos, lípidos y carbohidratos, ha constituido otra área por explorar.

Estudios recientes han demostrado la activación del sistema inmune debido a la presencia de células malignas, induciendo autoinmunidad a antígenos celulares autólogos. Muchos de los antígenos blanco son proteínas celulares, cuya expresión aberrante o desregulación puede dar lugar a la tumorigénesis. Esta respuesta sistémica ante la presencia del tumor provee la oportunidad de detectar cáncer en etapa temprana. Numerosos anticuerpos contra antígenos inmunogénicos asociados a tumor (es decir, autoanticuerpos) se han identificado usando análisis serológico de bibliotecas de expresión de cADN recombinante de tumores humanos con sueros autólogos.

2.2.2.1 BIOMARCADORES PROTEICOS EN CÁNCER DE PULMÓN

Los biomarcadores proteicos de cáncer de pulmón pueden ser clasificados, según la fuente de la proteína, en tres categorías principales: biomarcadores de suero, tejido y esputo. Sin embargo, en suero/plasma se encuentran la mayoría de biomarcadores potenciales, incluyendo los encontrados en tejido de biopsia y muchos fragmentos de proteínas circulantes generados del microambiente del tumor. Debido a que el objetivo final de los biomarcadores es el diagnóstico específico, temprano y no invasivo, así como el monitoreo post-terapia, la sangre se considera un material biológico adecuado. Por lo tanto, muchos de los biomarcadores propuestos para cáncer de pulmón son proteínas séricas. No obstante, hasta el momento ningún biomarcador ha mostrado la sensibilidad, especificidad y reproducibilidad adecuada para su validación como método de detección temprana de cáncer de pulmón (Sung, 2008). La tabla 3 muestra algunos de los biomarcadores séricos recomendados por la NACB (National Academy of Clinical Biochemistry) para diagnóstico diferencial cuando no es posible la toma de biopsia, así como en el pronóstico, monitoreo de tratamiento, entre otros. Sin embargo, ninguno ha sido validado por la FDA para diagnóstico clínico.

Tabla 3. Biomarcadores proteicos potenciales para la detección de cáncer pulmonar*

Proteína	**Diagnóstico**	**Monitoreo terapéutico**	**Pronóstico**	**Ontología**
CEA	AdenoCa, CPCG (>10ug/l)	AdenoCa, CPCNP avanzado	AdenoCa, CPCNP	Glucoproteína de membrana de la superfamilia de inmunoglobulinas
CYFRA 21.1	CPCNP, SCC	CPCNP avanzado	CPCNP, SCC	Constituyente estructural del citoesqueleto
TPA	CPCNP,SCC	-	CPCNP	
ProGRP	CPCP (>200 ng/l= alta sospecha)	CPCP	-	Actividad hormonal neuropéptido
NSE	CPCP (>100 ng/l= alta probabilidad)	CPCP	CPCP	Actividad fosfoglicerato deshidrogenasa; se localiza en citoplasma
Piruvato cinasa tumor M2	AdenoCa	-	AdenoCa	Actividad piruvato cinasa; glicólisis; citoplasma

SAA	Cáncer de pulmón	Transporte de lípidos; respuesta a fase aguda; quimiotaxis de células inmunes; región extracelular
Haptoglobina-α-2	AdenoCa	Actividad endopeptidasa tipo serina; proteólisis; región extracelular
APOA1	AdenoCa	Transporte de lípidos; inhibidor de lipasa; homeostasis, metabolismo y transporte de colesterol; región extracelular
KLKB1	AdenoCa	Actividad peptidasa; proteólisis
Anticuerpo anticiclina B1	Probable valor diagnóstico y pronóstico en estudios de screening	Expresión aberrante de ciclina B1, niveles elevados en citoplasma durante el ciclo celular donde puede ser proteolisada y presentada al CMH

*Tomada y modificada de Sung y colaboradores (2008). CEA= Antígeno carcinoembrionario, CYFRA 21.1= Fragmento de citoqueratina 19, TPA= Antígeno de polipéptido de tejido, ProGRP= Péptido liberador de progastrina, NSE= Enolasa específica de neuronas, SAA= Proteína amiloide de suero A, APOA1= Apolipoproteína A1, KLKB1= Kalicreína B1, AdenoCa=Adenocarcinoma, CCE= carcinoma de células escamosas, CPCP= cáncer de pulmón de células pequeñas, CPCNP= cáncer de pulmón de células no pequeñas, CPCG= Cáncer de pulmón de células grandes, CMH= Complejo mayor de histocompatibilidad

2.3 MÉTODOS ESTADÍSTICOS EN LA IDENTIFICACIÓN DE BIOMARCADORES

En el caso particular de cáncer de pulmón, las proteínas biomarcadoras individuales tienen baja utilidad diagnóstica debido a la limitada sensibilidad/especificidad que presentan, lo cual es atribuible en parte a la heterogeneidad de la enfermedad y los múltiples factores de riesgo asociados al desarrollo de la misma. Lo anterior ha conducido a la búsqueda de perfiles proteicos característicos de los individuos portadores de dicha patología, utilizando técnicas anteriormente mencionadas como 2-DE, espectrometría de masas (MS), arreglos de proteínas, etc. Tomando como ejemplo el uso de MS, ésta permite medir la expresión de cientos de proteínas o péptidos en muestras de sangre, tejido o esputo de pacientes con cáncer de pulmón e individuos control. Estos espectros, con 200-300 picos por muestra analizada son, sin embargo, difíciles de comparar entre muestras.

Uno de los pasos más importantes es entonces reducir las grandes dimensiones de los espectros, con el fin de extraer la mejor combinación de proteínas capaz de discriminar entre los grupos de interés (cáncer versus control). Para este fin, los espectros se procesan mediante algoritmos computarizados, basados en análisis estadísticos multivariados. Recientemente, los algoritmos más utilizados han sido los conocidos como "métodos de clasificación supervisados"; algunos ejemplos son la regresión logística binaria, el análisis discriminante lineal y cuadrático, los vectores de soporte mecánico (SVM por sus siglas en inglés), las redes neuronales y los árboles de clasificación, entre otros (Dossat, 2007). Se les llama métodos supervisados ya que se conocen de antemano los grupos a los cuales pertenecen las observaciones a clasificar, y lo que se pretende es ubicarlas en uno de los grupos. Ben-Dor y colaboradores (2000), después de comparar diversos algoritmos de aprendizaje[1] en la clasificación de cáncer contra grupo control, concluyeron que los métodos basados en árboles y SVM tienen mejor desempeño en términos de sensibilidad y especificidad. Por su parte, Chen y colaboradores (2009) construyeron un modelo de árbol de clasificación con 5 genes para predecir el resultado del tratamiento en pacientes con cáncer de pulmón de células no pequeñas, mientras que Wu y colaboradores (2003), compararon el rendimiento de algunos métodos para clasificar pacientes con cáncer ovárico, encontrando que el random forest (RF) supera a otros métodos en el análisis de datos de espectrometría de masas.

2.3.1 FUNDAMENTO TEÓRICO DE LOS MÉTODOS DE CLASIFICACIÓN UTILIZADOS EN EL PRESENTE ESTUDIO

Análisis Discriminante

El análisis discriminante es una técnica que permite analizar las diferencias entre grupos de objetos a partir de variables medidas sobre los mismos. Lo anterior se logra construyendo combinaciones lineales de las variables (funciones discriminantes). En particular el análisis discriminante lineal (ADL) se basa en el supuesto de normalidad multivariada e igualdad de las matrices de varianza y covarianza de los grupos a clasificar. En la ecuación 1 se observan las

[1]En un escenario donde se tiene una variable categórica (como sanos/enfermos), la cual se desea predecir basándose en un conjunto de variables de entrada (datos clínicos, por ejemplo). Al mismo tiempo, se tiene un conjunto de datos de entrenamiento, obtenido a partir de medir las variables de entrada en un grupo de objetos (individuos); usando estos datos, se puede construir un modelo de predicción, el cual pudiera ser capaz de predecir el resultado de nuevos objetos. El ejemplo anterior es conocido como algoritmo de aprendizaje supervisado.

relaciones lineales entre las variables x_i observadas, donde q es el número de grupos, p el número de variables medidas y $m = min\ (q\text{-}1, p)$, número de relaciones lineales.

$$y_1 = a_{11}x_1 + \ldots + a_{1p}x_p + a_{10}$$

$$y_m = a_{m1}x_1 + \ldots + a_{mp}x_p + a_{m0} \quad (1)$$

El objetivo del ADL es maximizar el cociente entre la varianza entre grupos y la varianza intragrupos.

$$Entre = \sum_{j=1}^{q} n_j(\bar{x}_{.j} - \bar{x})\ (\bar{x}_{.j} - \bar{x})^T \quad (2)$$

$$Intra = \sum_{j=1}^{q} \sum_{i=1}^{n} \left(\bar{x}_{ij} - \bar{x}_{.j}\right)\left(\bar{x}_{ij} - \bar{x}_{.j}\right)^T \quad (3)$$

Destacando que el ADL proporcionará una función discriminante menos que los subgrupos que se tengan, es decir, si la variable categórica tiene dos subgrupos, obtendremos una función discriminante, si tiene tres subgrupos obtendremos dos y así sucesivamente. Este concepto se ve muy claro desde un punto de vista gráfico, ya que si existen tres subgrupos, el procedimiento lo que hará será establecer dos funciones que separen a un grupo de otro (ver figura 2). En la construcción de las funciones discriminantes, el procedimiento permite incluir a todas las variables o usar un procedimiento de selección paso a paso que incluye solamente algunas variables que son estadísticamente significativas para discriminar sobre los grupos. Así también, las funciones discriminantes creadas pueden utilizarse para clasificar nuevos casos dentro de los grupos (Rencher, 2007).

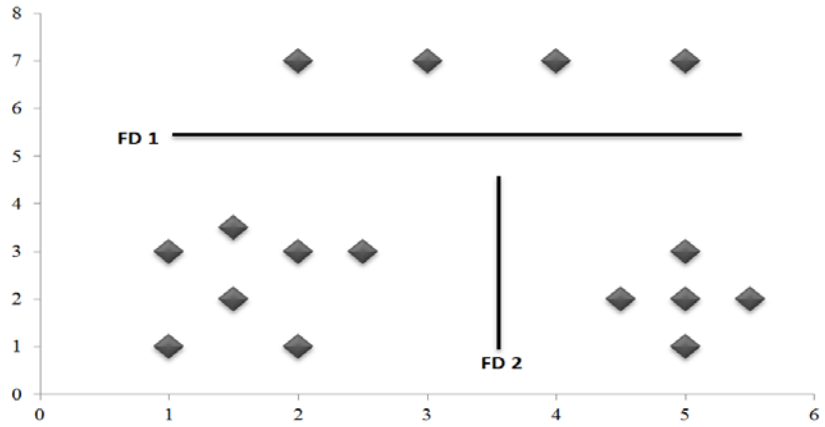


Figura 2. Ejemplo gráfico de 3 variables dependientes o subgrupos y dos funciones discriminantes para separarlos. FD = función discriminante.

Árbol de Regresión y Clasificación (CART, por sus siglas en inglés)

En 1984 Breiman desarrolló el algoritmo CART, cuyo resultado es en general un árbol de decisión, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición), formando así grupos homogéneos respecto a las observaciones que se desean discriminar.

CART es un método no paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado *raíz* y representa el total de observaciones, el nodo raíz es dividido en subgrupos (dos o más) determinados por la partición de una variable predictora elegida, generando nodos *hijos*. Los nodos hijos a su vez son divididos usando la partición de una nueva variable y el proceso recursivo se repite para los nuevos nodos hijos de manera sucesiva hasta cumplir una condición de parada. Algunos de los nodos resultantes son terminales, mientras que otros continúan dividiéndose hasta llegar a un nodo terminal.

Para dividir las observaciones se requiere un criterio de partición el cual determinará la medida de impureza de cada nodo; esta medida de impureza, denotada por $i(t)$, establecerá el grado de homogeneidad entre los grupos. Existen distintas medidas de impureza, la más utilizada para árboles de clasificación es el índice Gini:

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t) \quad (4)$$

En cada división el índice Gini tiende a separar la categoría más grande en un grupo aparte.

$$\Delta i = -\sum_{j=1}^{k} \left[p_j(t)\right]^2 \quad (5)$$

El análisis de CART generalmente consiste en tres pasos:

a) Construcción del árbol máximo. El árbol máximo es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es sobreajustado, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo.
b) Poda del árbol. La poda busca un balance entre el aumento del error de clasificación y la reducción del tamaño del árbol, además, ayuda a reducir el sobreajuste que tenga el árbol, ya que el árbol completo puede clasificar perfectamente al conjunto de entrenamiento pero

puede no ser tan efectivo con un conjunto de prueba (observaciones no utilizadas en la fase de construcción).

c) Selección del árbol óptimo mediante "validación cruzada". De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo, por tanto se requiere estimar con precisión el error de predicción utilizando validación cruzada. El objetivo es encontrar la proporción óptima entre la tasa de error de clasificación y la complejidad del árbol.

La idea básica de la "validación cruzada" es tomar del total de datos una muestra para entrenamiento, y el resto para la fase de prueba. Con la muestra de entrenamiento se calculan los estimadores, y el subconjunto de prueba es usado para verificar el desempeño de los estimadores obtenidos. El desempeño basado en el error de predicción es acumulado para obtener el error medio absoluto del conjunto de prueba.

Random Forest

Los métodos basados en árboles de decisión presentan dos principales limitaciones. En primera instancia, las reglas de asignación son extremadamente sensibles a pequeñas perturbaciones en los datos, es decir, a variaciones mínimas de nuevos datos a clasificar (inestabilidad). En segundo lugar, muchos estudios (de genómica, por ejemplo) obtienen datos de grandes dimensiones: miden gran número de variables "*p*" (número de genes) en una "*n*" problema pequeña (Zhang, 2003). Para superar estas dos debilidades, se ha sugerido el método de Random Forest (RF) el cual se basa en el desarrollo de muchos árboles de clasificación, logrando mayor estabilidad en el método y haciéndolo menos propenso al error de predicción como consecuencia de la perturbación de datos (Breiman, 2001). Por otro lado, el RF no requiere validación cruzada o separación de un subconjunto de datos para la fase de prueba, ya que la estimación de la tasa de error de clasificación la realiza utilizando una técnica de remuestreo de los datos originales, mejor conocida como "bootstraping". Además, el RF ofrece una estimación de las variables que son importantes en la clasificación mediante el índice Gini, el cual es una "medida de desorden", tomando el siguiente sentido: a mayor medida, mayor importancia en el modelo de clasificación creado, ya que valores próximos a 0 para el índice Gini implican un mayor desorden. Por el contrario, valores cercanos a 1 implican un menor desorden. Si computamos una medida de la disminución del índice Gini, cuanto mayor sea este valor, más variabilidad aporta a la variable dependiente (grupo), esta puntuación es también conocida como **Mean Decrease Gini**.

Curva ROC (por sus siglas en inglés, Receiver Operating Characteristic)

La curva ROC es usada para estimar la precisión de una prueba de discriminación, por ejemplo, enfermos versus no enfermos. La limitante de muchas pruebas es que la distribución de los resultados con frecuencia se traslapa, como se muestra en la figura 3. Dependiendo del punto de corte seleccionado para discriminar entre una muestra de enfermo, de una de no enfermo, la distribución de muestras correctamente clasificadas puede variar. La curva ROC es un gráfico de sensibilidad versus 1 - especificidad, de diferentes puntos de corte. La *sensibilidad* es la probabilidad de que la prueba sea positiva para muestras de enfermos, mientras que la *especificidad* es la probabilidad que la prueba sea positiva para muestras de no enfermos (verdaderos negativos a la enfermedad). Cuanto más se acerca la curva a la esquina superior izquierda (100% de sensibilidad y 100% de especificidad) mayor es la precisión de la prueba. Cada punto en la curva ROC representa un par de sensibilidad/especificidad correspondiente a cierto punto de corte elegido.

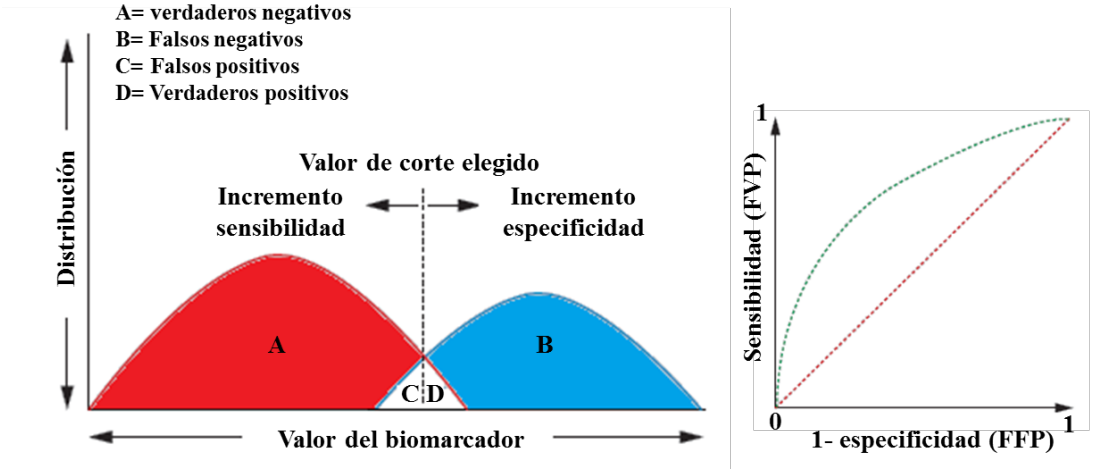


Figura 3. La curva ROC representa la relación entre sensibilidad y especificidad y es usada por ejemplo, para evaluar la eficiencia de un biomarcador, a diferentes puntos de corte. Un gráfico ideal es aquel que muestra el área bajo la curva (ABC) máxima. En el ejemplo dado, la curva roja representa una prueba inútil (ABC = 0.50), mientras que la curva verde representa una prueba útil (ABC = <1.00), más no perfecta (ABC = 1). FVP = fracción de verdaderos positivos; FFP = fracción de falsos positivos.

3. JUSTIFICACIÓN

En las últimas décadas, se han propuesto diversos biomarcadores para cáncer de pulmón con propósitos de diagnóstico y/o pronóstico, los cuales pueden medirse directamente en tejido o fluidos biológicos como suero/plasma, esputo y líquido pleural. No obstante, en su mayoría, los biomarcadores evaluados de manera individual no presentan suficiente sensibilidad/especificidad, lo que limita su utilidad clínica. Lo anterior ha llevado al estudio de biomarcadores en combinación, con el fin de mejorar su capacidad diagnóstica.

Diversos "métodos de clasificación supervisados" han sido ampliamente utilizados en la identificación de biomarcadores, sobre todo de tipo genómico, resultando ser herramientas eficientes en la tarea de analizar cientos de biomarcadores, de manera simultánea, con dos objetivos principales: 1) crear un modelo de predicción/clasificación y 2) identificar las variables (biomarcadores) con mayor peso en la discriminación entre individuos sanos y enfermos (Chen, 2011). Por lo tanto, el presente estudio se enfoca a la búsqueda y definición de un panel optimizado de proteínas, de utilidad diagnóstica, generado a partir de la aplicación de algunos métodos de clasificación supervisados, con el fin de proponer un modelo de clasificación, utilizando como variables de entrada 14 biomarcadores con diferente poder de discriminación entre individuos con cáncer de pulmón y sujetos control (sin la enfermedad).

4. HIPÓTESIS

El análisis combinado de un conjunto de proteínas, utilizando una serie de herramientas estadísticas multivariadas, permitirá identificar entre el grupo de biomarcadores seleccionados en el presente estudio, aquellos con mayor peso en la discriminación entre pacientes con cáncer de pulmón e individuos control. Este panel optimizado de proteínas presentará mayor poder diagnóstico que cada una de las proteínas evaluadas de manera individual, lo que permitirá proponer un modelo de clasificación de individuos, para la detección temprana de cáncer de pulmón.

5. OBJETIVOS

5.1 GENERAL

Determinar el panel óptimo de proteínas que en conjunto incremente el poder de discriminación entre pacientes con cáncer de pulmón e individuos control, utilizando métodos de análisis multivariado.

5.2 ESPECÍFICOS

- Determinar si existe diferencia estadísticamente significativa en la concentración de las proteínas séricas evaluadas, entre pacientes con cáncer de pulmón e individuos control.

- Determinar el conjunto de biomarcadores óptimo para discriminar entre pacientes con cáncer de pulmón e individuos control, mediante métodos estadísticos multivariados.

- Comparar la capacidad diagnóstica de biomarcadores individuales contra el panel obtenido mediante parámetros de sensibilidad, especificidad y precisión.

6. MATERIALES Y MÉTODOS

6.1 DESCRIPCIÓN DE LAS MUESTRAS

Para la presente investigación se incluyeron pacientes en protocolo de estudio por sospecha de cáncer de pulmón del Centro Médico Nacional de Occidente y de la Clínica 110 (ambas Instituciones localizadas en Guadalajara, Jalisco y pertenecientes al Instituto Mexicano Seguridad Social, IMSS), así como del Instituto de Seguridad Social del Estado de México y Municipios (ISSEMyM). Del total de muestras captadas, se seleccionaron 64, provenientes de pacientes con diagnóstico histopatológico confirmado de cáncer pulmonar, que a su vez cumplieron con los criterios de inclusión descritos en la tabla 4. Para efectos de comparación se conformaron dos grupos control. El primero incluyó 29 sujetos fumadores, a los cuales se les aplicó un cuestionario referente a su estado de salud, antecedentes patológicos, tiempo que llevan fumando, entre otros, con el fin de concentrar la información relevante de cada uno en su expediente respectivo. El segundo grupo estuvo integrado por 60 pacientes diagnosticados con enfermedades no neoplásicas como la enfermedad pulmonar obstructiva crónica (EPOC), asma y/o enfisema, procedentes de la Clínica 110 (IMSS) y del Hospital Civil de Guadalajara, sumando el estudio un total de 153 muestras. Los grupos control fueron seleccionados con base en características demográficas similares a los pacientes con cáncer de pulmón, como edad y sexo, y que además compartieran los principales factores de riesgo: exposición al humo de tabaco y/o leña e inflamación.

6.2 COLECTA Y PREPARACIÓN DE LA MUESTRA

Se obtuvieron 2 tubos de 10 ml de sangre periférica, extraída de la vena cubital, tanto de pacientes con cáncer de pulmón, como de los individuos control. Previo a la toma de muestra, el donador firmó la carta de consentimiento, aprobada por el Comité de Ética de la Institución participante, en la cual se da a conocer el destino de la muestra, la finalidad del estudio y los riesgos que implica la toma de sangre (anexo 3). Las muestras fueron procesadas en los laboratorios de biotecnología médica y farmacéutica del Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco (CIATEJ), siguiendo los lineamientos para muestras biológicas. Todas las muestras fueron tratadas de la misma manera, se centrifugaron a 3000 rpm por 10 min, y el suero resultante se alicuotó y almacenó a -70°C.

Tabla 4. Criterios a considerar para la toma de muestra

Criterios de inclusión	Criterios de exclusión
Paciente mexicano	Cáncer metastásico a pulmón
Cáncer de pulmón primario	Expediente clínico incompleto
Virgen de tratamiento	Diagnóstico dudoso
Tipo histológico, etapa, edad y sexo indistintos	Muestra de sangre insuficiente y/o hemolisada

6.3 CUANTIFICACIÓN DE BIOMARCADORES EN SUERO Y ANÁLISIS ESTADÍSTICO

Se realizó la cuantificación de 14 biomarcadores circulantes en los tres diferentes grupos: pacientes con cáncer, fumadores (Control 1) y pacientes con EPOC (Control 2), mediante la técnica de ELISA (Enzyme-linked Immunosorbent Assay), utilizando kits comerciales (véase tabla 5). Cada proteína fue cuantificada siguiendo el protocolo anexo en el kit. Se realizaron diluciones de las muestras sólo para las proteínas que así lo indicaba el kit. Las concentraciones de todas las proteínas se calcularon, utilizando el Software MasterPlex® 2010, y posteriormente se analizaron mediante los siguientes métodos estadísticos:

Evaluación individual de biomarcadores. Se utilizó estadística descriptiva (media, mediana, distribución, rangos, diagrama de caja y bigote) para obtener la distribución de los valores de concentración asociados a cada biomarcador, utilizando el software StatGraphics Centurion® XV. Así también, por medio de la prueba no paramétrica de suma de rangos Kruskall-Wallis, fue posible evaluar diferencias en la concentración de los biomarcadores entre los diferentes grupos de estudio, a un nivel de significancia del 95%. Las curvas ROC (Receiver Operating Characteristic) se calcularon utilizando el software Simca-P® 7.01, con el fin de determinar la precisión de cada biomarcador individual para discriminar entre pacientes con cáncer de pulmón y grupos control, por medio de parámetros como sensibilidad, especificidad, área bajo la curva (ABC) y valor P[2].

[2]El valor P determina si el área bajo la curva es significativamente diferente de 0.5, según la prueba de suma de rangos de Mann-Withney.

Tabla 5. Kits de ELISA comerciales utilizados en este estudio para evaluar las proteínas de interés.

Marcador tumoral	Kit comercial
Antígeno de Cáncer 125 (CA125)	ALPCO Immunoassays
Antígeno Carcinoembrionario (CEA)	ALPCO Immunoassays
Metaloproteinasa de matriz 9 (MMP-9)	R&D Systems
Haptoglobina	ASSAYPRO
Enolasa específica de neuronas (NSE)	ALPCO Immunoassays
Activador de plasminógeno tipo urocinasa (uPA)	ASSAYPRO
Transferrina	ASSAYPRO
Alfa-1-antitripsina	ASSAYPRO
Metaloproteinasa de matriz 1 (MMP-1)	R&D Systems
Apolipoproteína A-I (ApoA-1)	ASSAYPRO
Proteína C Reactiva (PCR)	ASSAYPRO
Proteína de unión a retinol (RBP)	ASSAYPRO
Fragmento de citoqueratina 19 (CYFRA 21.1)	DRG Diagnostics
YKL-40	QUIDEL CORPORATION

Análisis multivariado. Se realizaron tres diferentes métodos de clasificación supervisados, los cuales fueron previamente descritos. Primero, se obtuvo un modelo de discriminación con base en el análisis discriminante lineal (ADL) utilizando el método de Lambda de Wilks, el cual determina las variables más influyentes en la discriminación de los grupos (cáncer, control 1 y control 2), con el objeto de simplificar el modelo; el análisis se ejecutó en el software SPSS statistics® versión 19.0. Posteriormente, se utilizó el algoritmo CART para crear un árbol de clasificación capaz de asignar a los individuos de estudio a los grupos *a priori* (cáncer, control 1 y control 2) en función de las variables (proteínas), mediante el paquete RPART del software estadístico R® 2.13.0. Finalmente, se aplicó el algoritmo RF usando el paquete Random Forest en R® 2.13.0, para seleccionar la combinación de variables óptima, mediante la construcción de numerosos árboles de clasificación (500 en el presente estudio) con subconjuntos de proteínas, tomados de manera aleatoria. El bosque elige a qué grupo pertenece cada observación, tomando en cuenta el mayor número de votos sobre todos los árboles del bosque. Cada árbol se construyó por validación cruzada, donde un conjunto de entrenamiento (aproximadamente dos tercios de los datos) es elegido del total de datos, de manera aleatoria, y cada árbol crece lo más extenso posible

(sin poda) usando este conjunto. El árbol resultante es entonces utilizado para predecir el grupo al que pertenece cada observación del conjunto de prueba (el tercio de los datos restantes), que se denomina predicción out-of-bag (OOB). Este proceso se repite 500 veces, dando lugar a otro conjunto de entrenamiento seleccionado aleatoriamente y al crecimiento de un nuevo árbol que a su vez será usado para realizar otra predicción OOB. La precisión de clasificación del RF es medida por el promedio de error de predicción OOB de todo el bosque, denominado tasa de error OBB. Las variables (biomarcadores) con mayor peso en el modelo de clasificación, se tomaron con base en el valor de Mean Decrease Gini, previamente descrito. Por último, se construyó la curva ROC para el modelo de clasificación obtenido y, tomando en cuenta parámetros como área bajo la curva, sensibilidad, especificidad y porcentaje de casos correctamente clasificados, fue posible comparar el rendimiento de los biomarcadores individuales frente al panel elegido.

7. RESULTADOS

7.1 CARACTERÍSTICAS DEMOGRÁFICAS PRINCIPALES DE LOS GRUPOS DE ESTUDIO

Se colectaron 101 muestras de pacientes en protocolo de estudio por sospecha de cáncer de pulmón, de las cuales 69 resultaron ser carcinomas broncogénicos, confirmados por histopatología; sin embargo, 5 se descartaron por ser probables carcinomas metastásicos a pulmón. Por otro lado, se excluyeron 15 muestras de pacientes diagnosticados con neumopatías benignas, entre las que se encuentran hiperplasia epitelial, metaplasia escamosa y displasia. Las 17 muestras restantes no presentaron diagnóstico concluyente debido a limitaciones en la toma de la biopsia. Los diferentes métodos diagnósticos incluyeron, lavado, cepillado y biopsia bronquial, obtenidos por broncoscopía, biopsia por aspiración con aguja fina (BAAF), biopsia pleural y citología de líquido pleural, siendo la biopsia bronquial el método más utilizado (figura 4). Por otra parte, el tipo histológico predominante fue el cáncer de pulmón de células no pequeñas (63%), en particular el adenocarcinoma, seguido del cáncer de pulmón de células pequeñas (17.7%) y por último el cáncer de células grandes, con una frecuencia relativa del 1.6%, resultados que concuerdan con lo reportado en la literatura (Morales, 2000; Santos, 2004; Villalba, 2004; Gurrola, 2008). El 19.3% restante corresponde a carcinomas poco diferenciados. En la figura 5 se muestran las frecuencias de los tipos histológicos por sexo. Al realizar la comparación no se encontraron diferencias significativas entre sexo y tipo histológico. El 100% de los casos estudiados se presentó en estadios III y IV, donde las únicas opciones terapéuticas fueron quimioterapia, radioterapia o incluso sólo cuidados paliativos.

Los 60 pacientes con EPOC que conforman el grupo Control 2, fueron diagnosticados por espirometría, con niveles de FEV1/FVC <70%. En su mayoría, los tres grupos de estudio presentan al menos uno de los principales factores de riesgo, tabaquismo (activo/pasivo), humo de leña y/o factor ocupacional. La tabla 6 muestra las características clínicas y patológicas de los individuos bajo estudio.

Tabla 6. Perfiles demográficos y clínicos de los participantes en el estudio.

Demográficos	Control 1 (n=29)	Control 2 (n=60)	Cáncer (n=64)
Edad (años)	48.03±15.63	67.01±10.6	64.19±13.2
Rango	19-78	48-92	31-89
Sexo			
Femenino	10	19	28
Masculino	19	41	36
Fumador			
Activo	27	34	32
Índice tabáquico	18.7±16	31.2±28	36.34.2± 31.56
Pasivo		2	3
Exposición humo de leña		4	4
Tabaquismo + humo de leña	2	15	4
Ninguno			5
Sin dato		5	16
Etapa de la enfermedad			
III			7
IV			18
Histología			
Carcinoma de células no pequeñas			3
Adenocarcinoma			27
Carcinoma de células escamosas			4
Carcinoma epidermoide			5
Carcinoma de células pequeñas			12
Carcinoma de células grandes			1
Carcinoma poco diferenciado			12

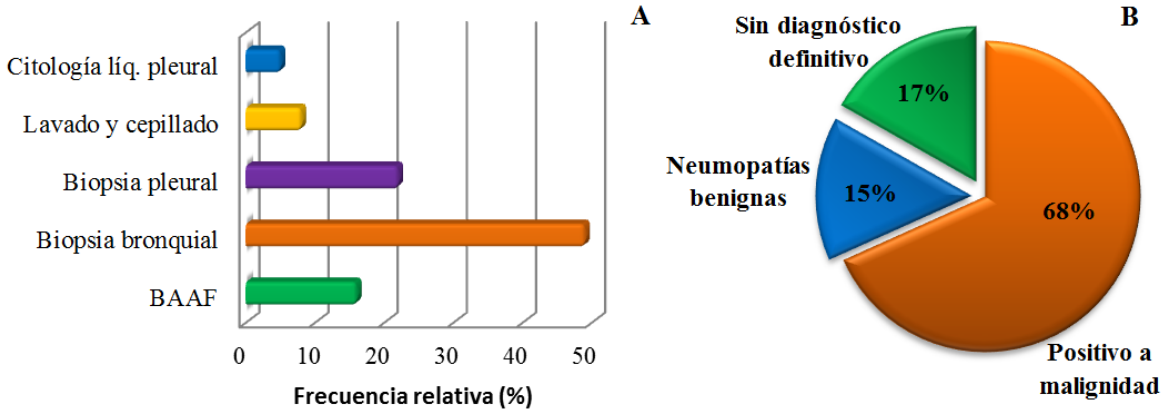


Figura 4. (A) Métodos diagnósticos empleados en los casos positivos a malignidad. En los pacientes en quienes se realizó más de un método diagnóstico, para el análisis se consideró únicamente el método que arrojó el resultado positivo a malignidad. (B) Distribución del total de muestras colectadas.

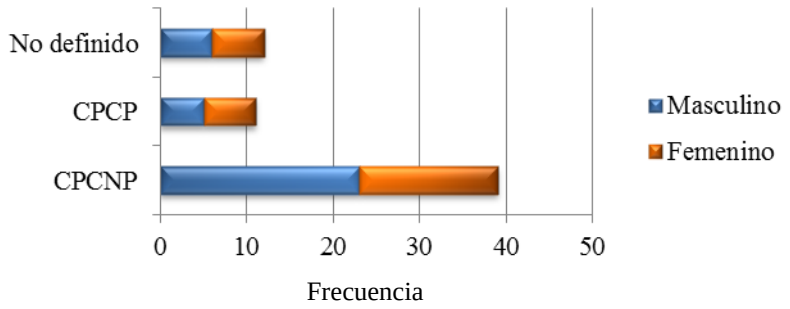


Figura 5. Frecuencia de los tipos histológicos en los pacientes con cáncer de pulmón. No se encontraron diferencias significativas entre sexos.

7.2 ANÁLISIS ESTADÍSTICO DE LOS BIOMARCADORES INDIVIDUALES

Se midió la concentración en suero de las 14 proteínas de interés, en las 64 muestras de pacientes con cáncer de pulmón, de 29 fumadores y de 60 pacientes con EPOC, y con estos valores de concentración se construyeron gráficos de caja y bigote para cada proteína. Se encontró que las concentraciones en suero de **CEA, HPT, MMP9, MMP1, CA125, α1ATTP, CYFRA 21.1, CRP y YKL-40**, fueron significativamente mayores en el grupo de cáncer, comparadas con los grupos control, a un nivel de significancia del 99% (figura 6). Por el contrario, los niveles de **TF** fueron significativamente menores en el grupo de cáncer, al mismo nivel de significancia. El resto de las proteínas (ApoA1, NSE, RBP y uPA) no presentaron diferencia significativa en concentración, entre el grupo de cáncer y los grupos control (anexo 4). Por otra parte, las 10 proteínas anteriormente mencionadas, presentaron un área bajo la curva mayor a 0.60 y valor $P<0.001$, es decir, podrían ser útiles en la diferenciación entre pacientes con cáncer de pulmón y sujetos control.

En la figura 7A se muestran las curvas ROC de los 9 biomarcadores que tienden a aumentar en los pacientes con cáncer de pulmón. La curva ROC de TF se graficó por separado debido a que el comportamiento del biomarcador va en sentido opuesto, es decir, tiende a disminuir en el grupo de cáncer con respecto a los controles, con un área bajo la curva de 0.69 y valor $P <0.001$ (figura 7B). La relevancia estadística de estas proteínas se presenta en la tabla 7, en donde puede observarse que CA125 aparece como el mejor biomarcador individual, debido a que presenta la mayor área bajo la curva (0.85). El nivel de corte óptimo se estableció en 13.67 U/ml, obteniéndose una sensibilidad del 83.15%, a una especificidad del 78.13%.

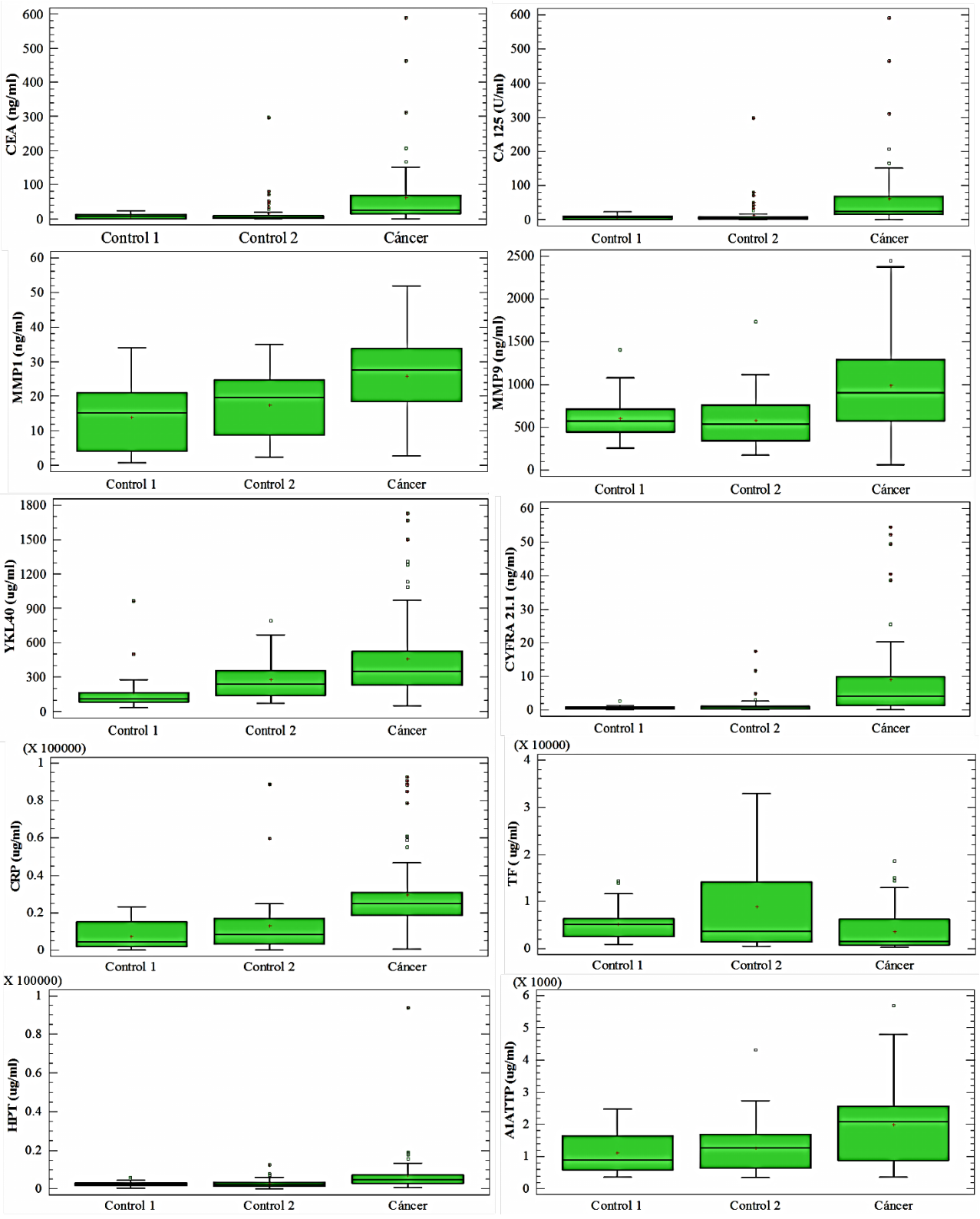


Figura 6. Gráficos de caja y bigote de los 10 biomarcadores que presentaron diferencia significativa en las concentraciones séricas entre el grupo de cáncer de pulmón y los grupos control. En el eje de las abscisas, Control 1 = Fumadores, Control 2 = Pacientes con EPOC, Cáncer = Pacientes con carcinoma broncogénico. Los asteriscos indican diferencias significativas (*P<0.01) en el valor de la mediana entre grupos.

Tabla 7. Biomarcadores con relevancia estadística entre los grupos de estudio.

Proteína	Control 1		Control 2		Cáncer		Curva ROC	
	Mediana	Rango	Mediana	Rango	Mediana	Rango	ABC	valor P
CA125*	6.219	0.0-23.56	4.0685	0.0-296.57	25.1839	4.288-310.19	0.85	<0.001
CYFRA21.1£	0.6048	0.0-2.6556	0.9198	0.0-17.5584	4.2468	0.0-40.4867	0.84	<0.001
CRP ¥	4499.83	0.0-23434.5	8487.64	0.0-59850	24702.4	186.57-26580.3	0.83	<0.001
HPT ¥	2667.01	555.47-6053.18	2290.3	225.24-12423.1	4637.09	3650.6-92781.6	0.78	<0.001
MMP-1£	15.2243	0.89-33.06	19.7	2.5-34.85	27.4595	2.93-37.59	0.74	<0.001
α1ATTP ¥	891.423	360.65-2480.08	1271.2	335.832-2725.1	2091.36	368.89-3261.12	0.74	<0.001
MMP-9 £	574.122	260.88-1409.67	536.02	70.817-1044.63	903.825	249.87-2441.23	0.72	<0.001
YKL-40 ¥	110.887	33.38-965.47	241.528	81.32-563.21	351.284	88.75-364.44	0.72	<0.001
TF ¥	5123.67	976.2-14394.8	3603.8	3603.8-32897.6	1482.6	259-18335.6	0.69	<0.001
CEA £	1.1278	0.0-5.28	1.5695	0.97-31.27	2.6538	0.14-126.07	0.67	<0.001

* = U/ml
¥ = ug/ml
£ = ng/ml

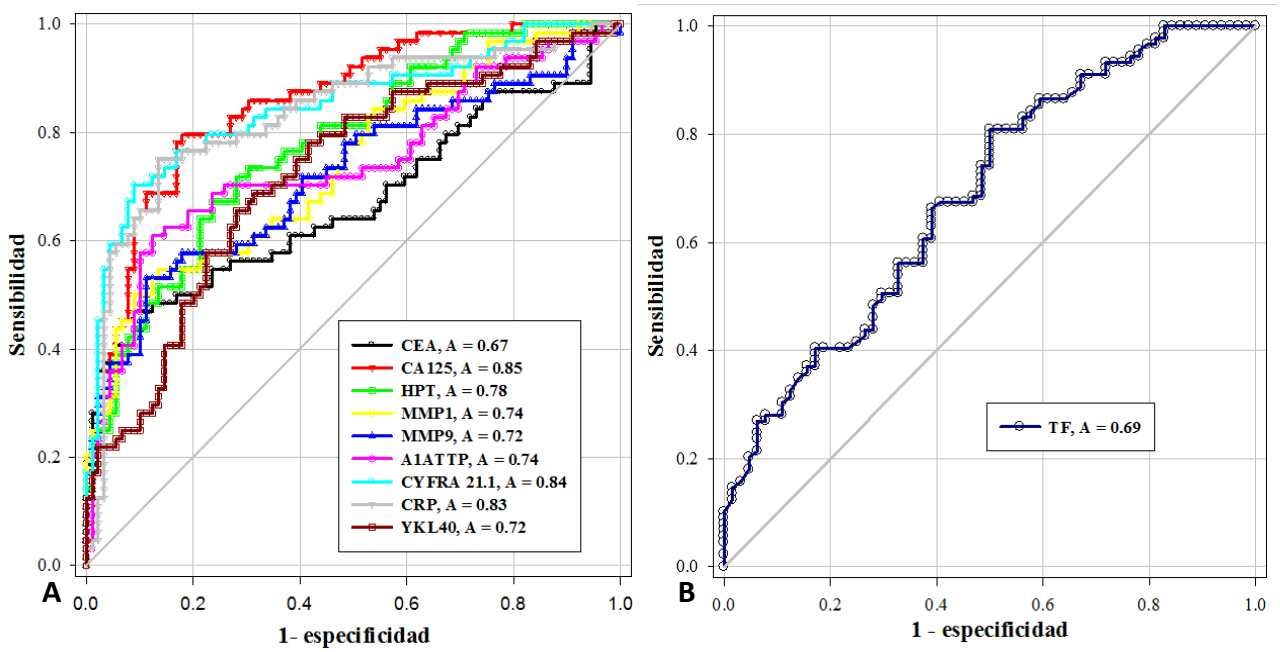


Figura 7. Curvas ROC de los 10 biomarcadores con área bajo la curva mayor a 0.60 y valor P < 0.001. En el análisis de curvas ROC, CA125 presenta la mayor área bajo la curva (0.85), seguida de CYFRA 21.1 con 0.84

7.3 ANÁLISIS MULTIVARIADO

7.3.1 ANÁLISIS DISCRIMINANTE LINEAL

El análisis discriminante lineal (ADL) tuvo por objetivo determinar cuáles variables (biomarcadores) cuantifican mejor las diferencias entre los grupos definidos *a priori*, así también, establecer un modelo para clasificar a un individuo con base en los valores del conjunto de biomarcadores. Se construyó un ADL tomando en cuenta a los 3 grupos de estudio (Cáncer, Control 1 y Control 2) y, mediante un algoritmo de selección de variables con el método Lambda de Wilks, se obtuvieron 2 funciones discriminantes compuestas por 7 variables (**HPT, MMP9, A1ATTP, TF, CYFRA 21.1, ApoA1 y YKL40**), las cuales resultaron significativas en la predicción de los grupos. Ambas funciones discriminantes presentan valor P menor a 0.05, es decir, existen diferencias entre las medias de cada grupo para las funciones discriminantes, indicando que ambas discriminan los grupos, a un nivel de significancia del 95%. La figura 8 muestra la puntuación que obtiene cada individuo de estudio con base en las funciones discriminantes creadas, pudiéndose observar que la función discriminante 1 logra separar a la mayoría de los pacientes con cáncer de pulmón, de los controles. A continuación se muestra el modelo conformado por 3 funciones, usadas para clasificar a los individuos de estudio, las cuales podrían predecir a qué grupo pertenecen nuevas observaciones[3].

$$\boldsymbol{Control\ 1} = 5.396 + 0.0000633067 * \text{HPT} + 0.00382081 * \text{MMP9} + 0.00189665 * \text{A1ATTP} + 0.000345599 * \text{TF} - 0.00919494 * \text{CYFRA 21.1} + 0.000348926 * \text{ApoA1} - 0.000322373 * \text{YKL40}$$

$$\boldsymbol{Control\ 2} = -8.10412 + 0.000067446 * \text{HPT} + 0.00371882 * \text{MMP9} + 0.0021079 * \text{A1ATTP} + 0.0004931 * \text{TF} + 0.0024242 * \text{CYFRA 21.1} + 0.0005089 * \text{ApoA1} + 0.0008038 * \text{YKL40}$$

$$\boldsymbol{Cáncer} = -10.2792 + 0.00016030 * \text{HPT} + 0.005709 * \text{MMP9} + 0.003466 * \text{A1ATTP} + 0.0003874 * \text{TF} + 0.0911 * \text{CYFRA 21.1} + 0.0001478 * \text{ApoA1} + 0.003011 * \text{YKL40}$$

La tabla 8 muestra el resultado de utilizar las funciones anteriores para clasificar cada una de las 153 observaciones empleadas para ajustar el modelo. Del total de observaciones el 69.28% fueron correctamente clasificadas. Cabe señalar que el bajo porcentaje de observaciones correctamente clasificadas se debe a que el modelo no es capaz de discriminar entre los grupos de

[3]Cada observación es asignada al grupo que corresponde el valor C_{ij}* $priori_j$ más grande, donde C_{ij} es la puntuación obtenida de cada función de clasificación y $priori_j$ es la probabilidad *a priori* de que un individuo provenga del grupo j, en este caso es de 0.333 asumiendo que existe la misma probabilidad de pertenecer a cualquiera de los 3 grupos.

fumadores y EPOC. Por lo tanto, se construyó un ADL tomando en cuenta el grupo de cáncer de pulmón y un sólo grupo control (fumadores más EPOC). El modelo obtenido es similar al anterior, con la diferencia de que el porcentaje de observaciones correctamente clasificadas aumentó a 86.93% utilizando las mismas 7 proteínas. Dicho porcentaje comprende 76.56% de verdaderos positivos y 94.38% de verdaderos negativos (véase tabla 9).

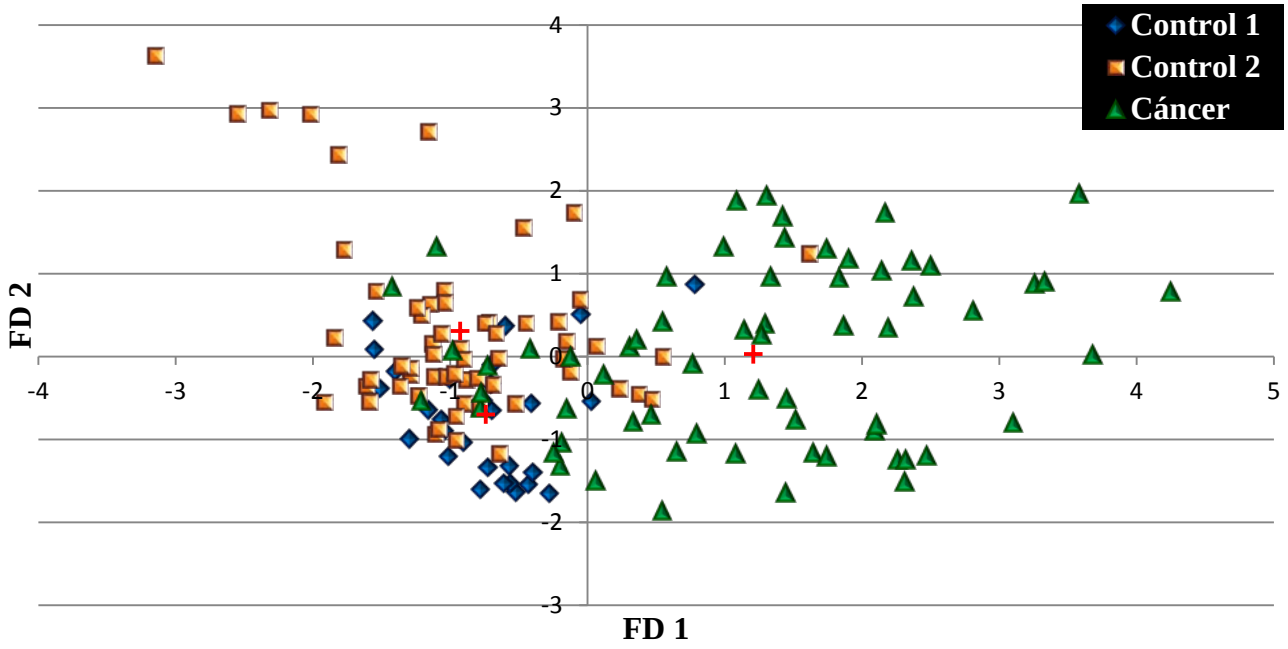


Figura 8. Gráfico de funciones discriminantes, donde cada punto representa una observación o individuo de estudio. Las funciones discriminantes creadas con 7 biomarcadores logran discriminar de manera similar al modelo con el total de variables, 72.55% de casos correctamente clasificados con 13 biomarcadores, frente a 69.28% utilizando sólo 7 variables. La FD1 logra separar en su mayoría al grupo de cáncer, de los controles, mientras que la FD2 no muestra una discriminación tan evidente entre fumadores y EPOC. FD1 = Función Discriminante 1, FD2 = Función Discriminante 2.

Tabla 8. Clasificación de las observaciones en tres grupos de estudio, usando el modelo conformado por 7 biomarcadores, obtenido a partir del ADL.

Grupo real	Tamaño de grupo	Grupo predicho		
		Cáncer	**Control 1**	**Control 2**
Cáncer	64	47 (73.44%)	11 (17.19%)	6 (9.38%)
Control 1	29	1 (3.45%)	22 (75.86%)	6 (20.69%)
Control 2	60	4 (6.67%)	19 (31.67%)	37 (61.67%)

Tabla 9. Clasificación de las observaciones en dos grupos de estudio, usando el modelo conformado por 7 biomarcadores, obtenido a partir del ADL.

Grupo real	Tamaño de grupo	Grupo predicho	
		Cáncer	**Control**
Cáncer	64	49 (76.56%)*	15 (23.44%)
Control	89	5 (5.62%)	84 (94.38%)*

* Los porcentajes de la diagonal corresponden a la sensibilidad y especificidad respectivamente.

Al analizar las variables que utiliza el modelo para discriminar, es de particular interés la proteína ApoA1, debido a que en el análisis univariado (prueba de Kruskall Wallis) no presentó diferencias significativas en la concentración entre el grupo de cáncer de pulmón y los controles; de igual manera, el área bajo la curva ROC no es significativamente diferente de 0.5 (prueba de Mann Whitney), lo cual lleva a pensar que el modelo podría carecer de precisión en la clasificación de nuevas observaciones. Este resultado controversial puede ser atribuible a que el ADL requiere una distribución normal multivariante de las variables (proteínas), que la variabilidad de éstas sea homogénea y que sus distribuciones no sean extremadamente asimétricas. Por lo tanto, fue necesario aplicar otro método de clasificación, debido a que los datos no presentan distribución normal multivariante.

Con el fin de definir un modelo más preciso en la predicción de observaciones, y por otro lado, identificar el panel de proteínas que sea capaz de diferenciar entre pacientes con cáncer pulmonar y controles, se realizó un árbol de clasificación con el algoritmo CART (figura 9), el cual utiliza **CYFRA 21.1, CA125, YKL40, CRP, NSE y HPT** para discriminar entre los 3 grupos de estudio. En la tabla 10 se presentan los indicadores sobre la precisión del modelo de clasificación, como la sensibilidad y especificidad.

Tabla 10. Matriz de confusión obtenida del árbol de clasificación tomando en cuenta 3 grupos de estudio.

Grupo real	Tamaño de grupo	Grupo predicho		
		Cáncer	**Control 1**	**Control 2**
Cáncer	64	54(84.4%)	7(10.9%)	3(4.7%)
Control 1	29	0	27(93.1%)	2(6.9%)
Control 2	60	11(18.4%)	11(18.4%)	38(63.2%)

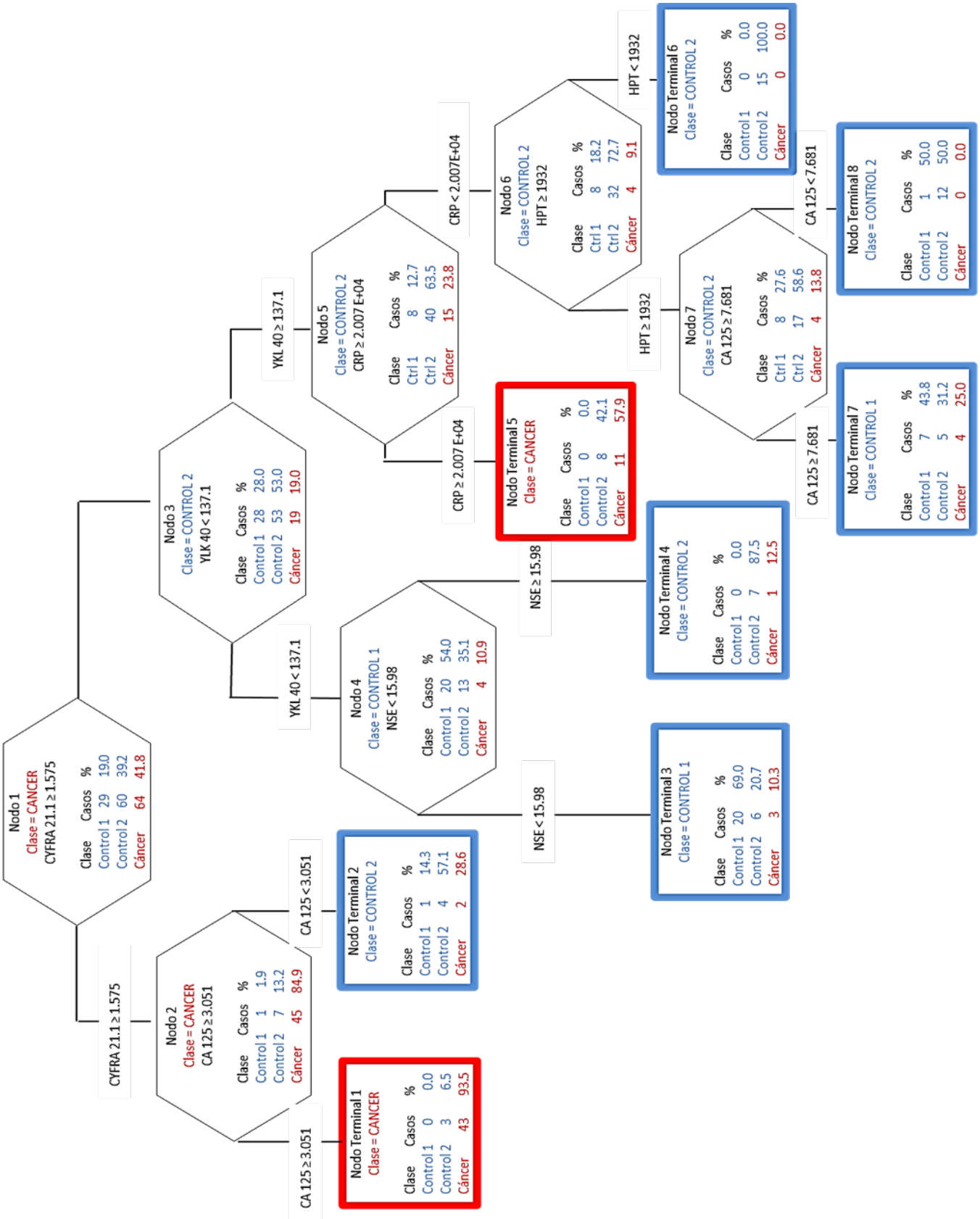


Figura 9. Árbol de clasificación para predecir si la observación pertenece a cáncer, control 1 o control 2. Cada nodo del árbol muestra la variable y la concentración umbral utilizada para dividir los individuos de estudio. Las observaciones continúan a lo largo de cada brazo dependiendo si el valor medido es menor, igual o mayor que el valor de corte de la proteína, hasta llegar a algún nodo terminal donde se decide a qué grupo pertenece dicha observación.

El valor de sensibilidad aumentó a 84.4% comparado con 76.56% del ADL, sin embargo, se observa un comportamiento similar al modelo anterior ya que el árbol de clasificación no es tan eficiente en la discriminación entre fumadores y EPOC. Por lo tanto, se construyó un árbol de clasificación con el grupo de cáncer y un solo grupo control (fumadores más EPOC); el resultado se muestra en la figura 10. Este último logró clasificar correctamente el 89.54% del total de observaciones, con un nivel de sensibilidad del 79.7% y 96.6% de especificidad, utilizando **CYFRA 21.1, CA125 y CRP** como variables predictoras y 5 nodos terminales. El 80% (51 de 64) del total de casos de cáncer de pulmón fueron agrupados en los nodos terminales 1 y 4, mientras que solo 3.3% (3 de 89) de los casos control fueron asignados en dichos nodos. Por consiguiente, si un paciente es asignado a uno de estos nodos, la probabilidad de tener cáncer de pulmón es de 94.4% (51 de 54). Es posible calcular la probabilidad de presentar cáncer de pulmón para cada uno de los 5 nodos terminales; los porcentajes se observan de color rojo dentro de cada nodo terminal (véase figura 10). La figura 12 muestra la curva ROC del árbol de clasificación, donde se observa un aumento del área bajo la curva (0.91) con respecto a CA125 (0.85).

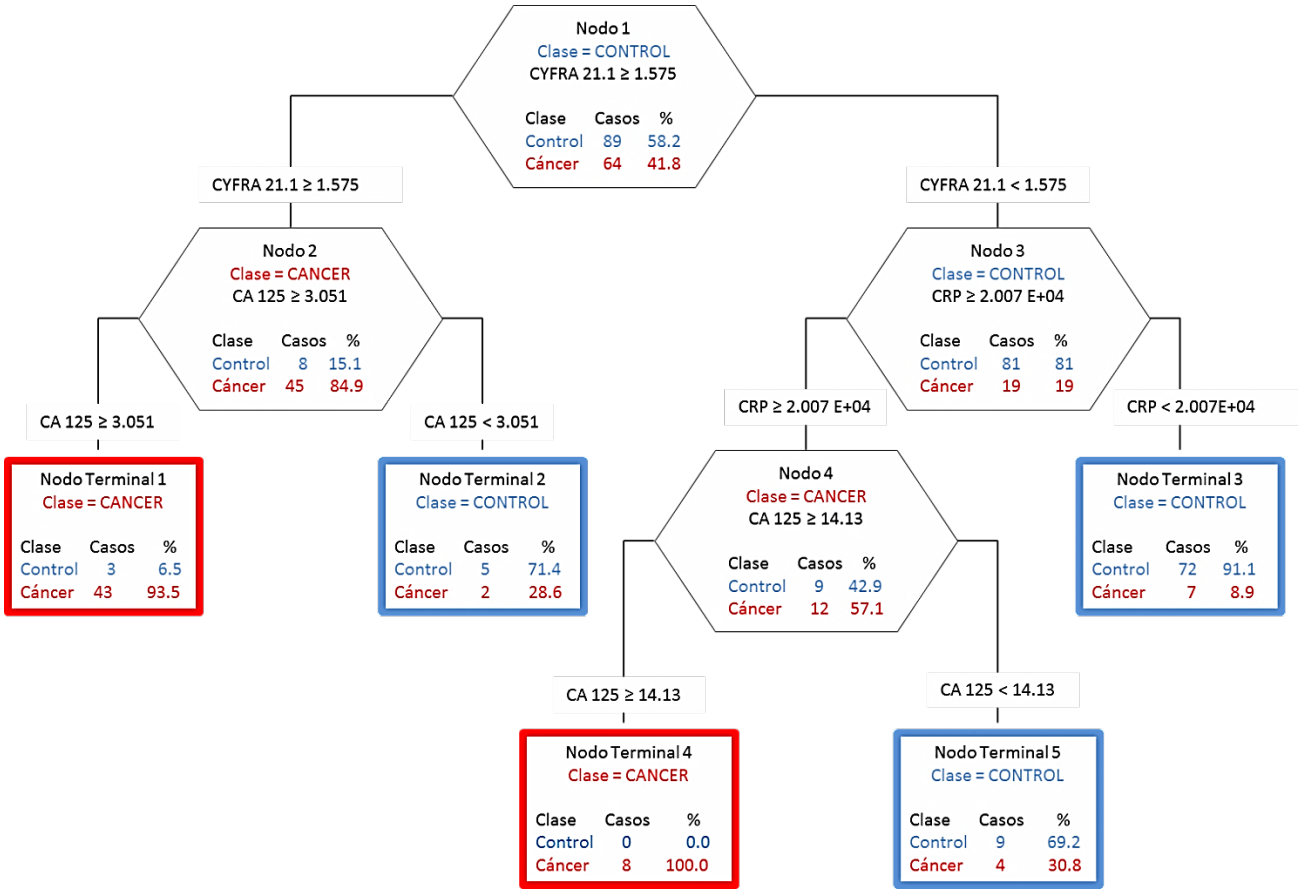


Figura 10. Árbol de clasificación para predecir si las observaciones pertenecen al grupo de cáncer o al grupo control. El algoritmo utiliza sólo 3 variables (CYFRA 21.1, CA125 y CRP) para clasificar correctamente la mayoría de observaciones, a diferencia del árbol de clasificación anterior donde fueron utilizadas 6 variables.

Tabla 11. Matriz de confusión obtenida del árbol de clasificación tomando en cuenta 2 grupos de estudio.

Grupo real	Tamaño de grupo	Grupo predicho	
		Cáncer	**Control**
Cáncer	64	51(79.7%)*	13(20.3%)
Control	89	3(3.4%)	86(96.6%)*

* Los porcentajes de la diagonal corresponden a la sensibilidad y especificidad respectivamente.

Ahora bien, una de las limitantes de este árbol de clasificación, es que es altamente sensible a variaciones mínimas en los datos provenientes de nuevas observaciones a clasificar, por esta razón, se utilizó el algoritmo random forest (RF) el cual permite identificar el conjunto de variables que arroja el mejor desempeño en la predicción de observaciones, y que se basa en la construcción de 500 árboles de clasificación. Por otra parte, fue posible estimar la precisión del modelo para predecir nuevas observaciones, ya que cada árbol fue construido por validación cruzada, calculando con ello la tasa de error de predicción OOB (para mayor detalle consultar apartado 7.3). Como se muestra en la figura 11, las proteínas con mayor peso en la clasificación fueron **CA125, CYFRA 21.1 y CRP**.

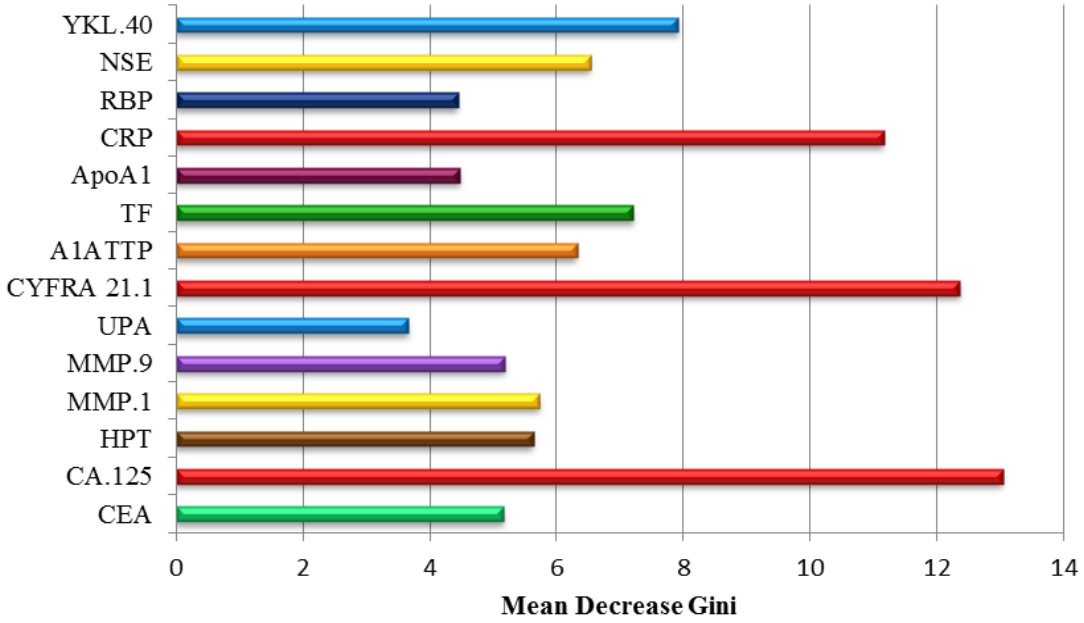


Figura 11. Las variables con mayor peso en la clasificación fueron determinadas con base en el valor Mean Decrease Gini; entre mayor sea dicho valor, mayor es la capacidad de diferenciar entre pacientes con cáncer de pulmón y controles.

Finalmente, se calculó la curva ROC para comparar la capacidad de discriminación del modelo creado con random forest frente al mejor biomarcador individual; los resultados se muestran en la figura 12. El área bajo la curva del modelo es de 0.93, frente a 0.85 del mejor biomarcador individual (CA125). Por otra parte, el punto de corte óptimo para el modelo se obtuvo asumiendo que existe la misma probabilidad (0.5) de pertenecer a cualquiera de los dos grupos (cáncer versus control), dando como resultado una sensibilidad de 86%, a una especificidad de 94.4%. A un nivel de especificidad similar al modelo (94.4%), CA125 mostró una sensibilidad de 43.75%, lo que indica que con el modelo fue posible incrementar 42.25% la sensibilidad (86%). Por último, el modelo fue capaz de clasificar correctamente el 91.5% de las observaciones, a diferencia de CA125 que al punto de corte óptimo logró clasificar 79.7% del total de observaciones. La tabla 12 concentra los parámetros de precisión obtenidos del modelo de clasificación, frente a los biomarcadores individuales que utiliza el mismo.

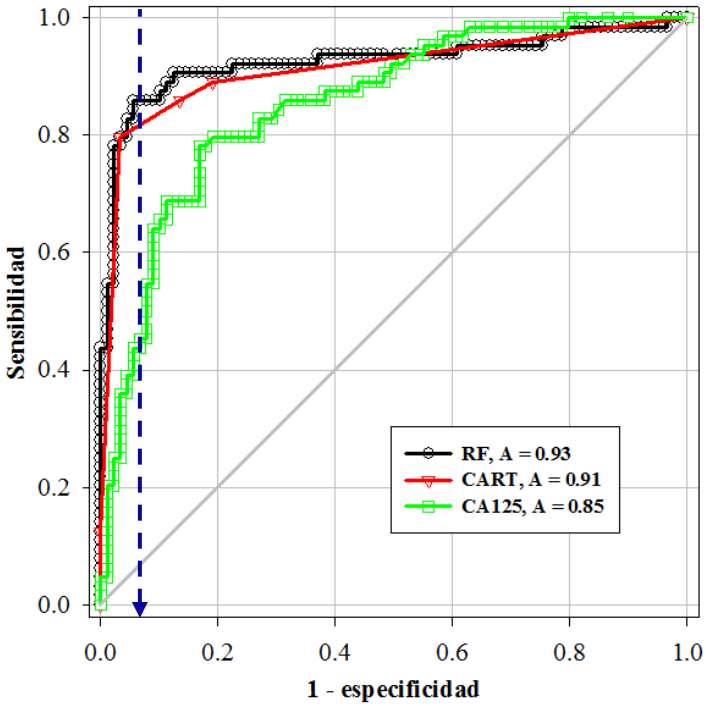


Figura 12. Comparación de la curva ROC del modelo obtenido con random forest frente al biomarcador individual con mayor área bajo la curva (CA125). Se estableció un nivel de especificidad de 0.944 para ambas curvas, como se indica con la flecha amarilla, observándose un incremento en la sensibilidad del modelo de clasificación de 42.25%, así como un aumento de 11.8% de casos correctamente clasificados.

Tabla 12. Rendimiento del modelo de clasificación Random Forest, frente a los mejores biomarcadores individuales.

	ABC	**Punto de corte óptimo**	**Sensibilidad***	**Especificidad***	**Precisión*ł**
Random Forest	0.93	0.5**	86	94.4	91.5
CA125	0.85	13.67 U/ml	83.15	78.13	79.7
CYFRA 21.1	0.84	1.575 (ng/ml)	70.31	91.01	78.4
CRP	0.83	19347 (ug/ml)	75	86.52	79.7

* Valores representados en porcentaje. ABC= área bajo la curva ROC

ł La precisión de la prueba está dada por el porcentaje de casos correctamente clasificados

**Cada observación presenta la misma probabilidad de pertenecer a cualquiera de los dos grupos

8. DISCUSIÓN

Actualmente no existen técnicas mínimamente invasivas que brinden un diagnóstico satisfactorio y oportuno de cáncer de pulmón. Hasta el momento, la técnica más promisoria en protocolos de screening de pacientes asintomáticos, es la tomografía computarizada; sin embargo, dada la baja especificidad de la prueba (64%) y la capacidad limitada para refinar la población objetivo, con base en criterios de selección como edad e índice tabáquico (Bach, 2003; McWilliams, 2003; MacRedmond, 2004), los biomarcadores séricos siguen siendo una alternativa prometedora. A pesar de la gran variedad de biomarcadores reportados para cáncer de pulmón, su limitada sensibilidad y especificidad impiden su uso en la clínica; no obstante, se ha demostrado que la aplicación de herramientas estadísticas (métodos basados en árboles de clasificación, particularmente), en la identificación de paneles de biomarcadores para diversos tipos de cáncer, puede aumentar los parámetros de sensibilidad/especificidad y, por consiguiente, incrementar la precisión de la prueba (Zhang, 2003; Yang, 2005; Díaz, 2006; Patz, 2007; Statnikov, 2007; Datta, 2008; Barrett, 2008; Farlow, 2010).

En el presente estudio se desarrolló una estrategia para extraer una combinación óptima de biomarcadores, partiendo de 14 proteínas tomadas de la literatura por su valor diagnóstico previamente reportado (véase figura 13); así también, se obtuvo un modelo capaz de discernir entre pacientes con cáncer de pulmón e individuos control, el cual presenta mayor robustez comparado con los biomarcadores de manera independiente. Dicha estrategia podría ser empleada en la evaluación de biomarcadores adicionales, para determinar variables no redundantes que brinden mayor precisión al modelo de predicción. El modelo obtenido utiliza 3 proteínas para clasificar correctamente el 86% (56/64) de los pacientes con cáncer pulmonar y 94.4% (83/89) de individuos control. Cabe resaltar que 2 de los 6 controles mal clasificados por el algoritmo, corresponden a pacientes con EPOC, los cuales, en meses posteriores a la toma de muestra, fueron diagnosticados con cáncer de pulmón; se desconoce si al momento que fueron colectadas las muestras, los pacientes ya contaban con alguna neoplasia maligna.

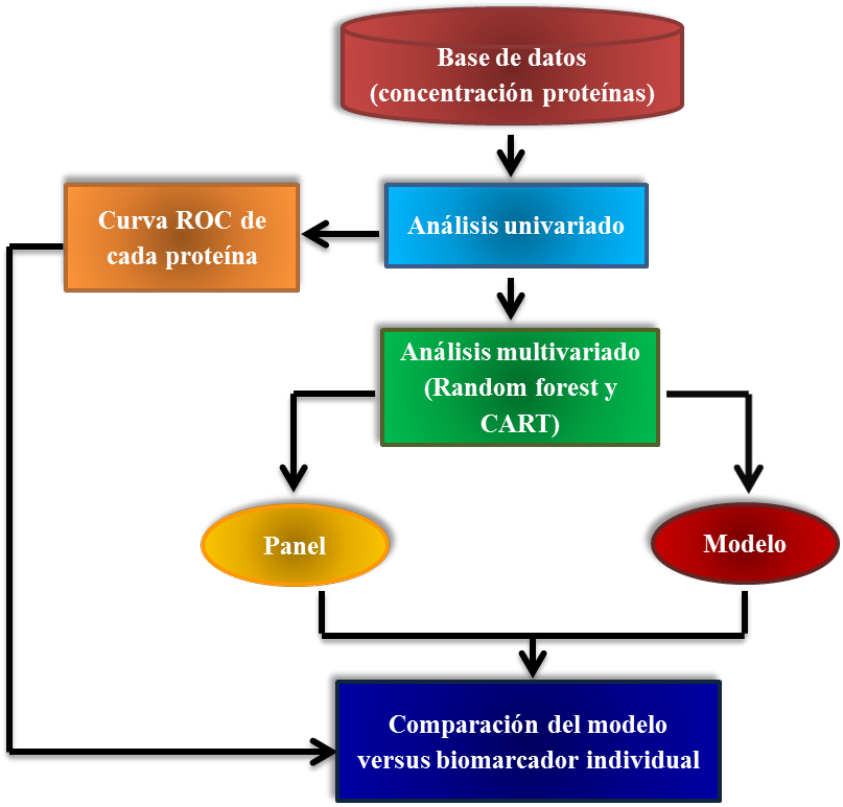


Figura 13. Diagrama de flujo de la estrategia metodológica empleada para identificar la combinación de variables (proteínas) en la discriminación de pacientes con cáncer de pulmón e individuos control.

Young RP y colaboradores (2009) analizaron la prevalencia de EPOC en pacientes diagnosticados con cáncer de pulmón y fumadores (los grupos fueron pareados en edad, sexo e índice tabáquico ≥ 10), y encontraron que aproximadamente 50% de los casos con cáncer de pulmón presentaron EPOC coexistente de moderada a severa, comparado con 8% en el grupo control. Además, uno de cada 4 casos contaba con diagnóstico de EPOC previo al cáncer de pulmón. Diversos estudios sugieren que la EPOC confiere un riesgo 6 veces mayor a desarrollar CP comparado con fumadores con función pulmonar normal (Skillrud, 1986; Mannino, 2003; Young, 2009). Dadas las estadísticas anteriores y los resultados obtenidos del modelo de clasificación, se ha dado seguimiento a los expedientes clínicos del grupo con EPOC, encontrando hasta el momento que siete de los 60 pacientes (11.7%) desarrollaron cáncer de pulmón y 20% (12/60) del total de pacientes presentan nódulos pulmonares. Por lo tanto, el escenario inmediato en el cual podría ser aplicado el panel de proteínas es justo en este tipo de pacientes, ya que la mayoría cuenta con los principales factores de riesgo: tabaquismo y/o exposición al humo de leña, más inflamación crónica. Aquellos pacientes con EPOC, con perfil

clínico de alto riesgo, y clasificados por el modelo como pacientes con cáncer de pulmón, podrían requerir intervención inmediata y con ello aumentar la probabilidad de detección temprana de carcinomas pulmonares.

Diversos autores han reportado combinaciones de biomarcadores para cáncer de pulmón, sin embargo, al comparar nuestros resultados con los obtenidos por otros grupos de trabajo, resulta evidente que se requiere la optimización y adecuación de los paneles de biomarcadores, antes de su aplicación a las diferentes poblaciones en el mundo. Desafortunadamente, hasta donde sabemos, en nuestro país no se han realizado estudios de esta naturaleza, lo que dificulta la atención a los pacientes con cáncer de pulmón de nuestra población. En 2007, Patz y colaboradores, mediante el algoritmo CART, obtuvieron un panel de 4 biomarcadores (antígeno carcinoembrionario, proteína de unión a retinol, α-1-antitripsina y antígeno de carcinoma de células escamosas); los niveles de sensibilidad y especificidad obtenidos con el conjunto de entrenamiento fueron de 89.3% y 84.7% respectivamente, mientras que para el conjunto de prueba la tasa de casos correctamente clasificados resultó disminuida (77.8% de sensibilidad y 75.4% de especificidad), a diferencia de nuestro trabajo, en donde logramos disminuir (2%) la tasa de error de clasificación en la etapa de validación, utilizando el algoritmo random forest, el cual construyó cada uno de los árboles del bosque por validación cruzada. Por otra parte, nuestro modelo muestra valores de sensibilidad y especificidad mayores (véase tabla 12), combinando el fragmento de citoqueratina 19 (CYFRA 21.1), el antígeno carbohidrato 125 (CA125) y la proteína C reactiva (CRP).

Por su lado, Farlow y colaboradores (2010), mediante random forest y CART identificaron un panel de 6 biomarcadores (factor de necrosis tumoral α, CYFRA 21.1, interleucina Ira, metaloproteinasa de matriz extracelular, proteína quimiotáctica monocítica I y selectina E) para discernir entre pacientes con CPCNP en etapa temprana e individuos con alto riesgo (pacientes diagnosticados con EPOC/asma o alguna lesión pulmonar no neoplásica resecada, granuloma por ejemplo). El panel fue capaz de clasificar correctamente 95% del total de casos, con 99% de sensibilidad y 95% de especificidad. Como puede apreciarse, la especificidad es similar a la obtenida en el presente estudio (94.4%), sin embargo, nuestra sensibilidad es menor (86%). Se desconoce la causa del limitado valor de sensibilidad, no obstante, cabe recordar que se trata de paneles de proteínas diferentes, aplicados a poblaciones diferentes, es decir, sería necesario evaluar los dos paneles en una misma población y determinar cuál presenta un desempeño

superior. Otra causa probable pudiera ser que 2 de las 3 proteínas de nuestro modelo (CYFRA 21.1 y CA125) han sido ampliamente reportadas por su valor diagnóstico en CPCNP (Rastel, 1994; Paone, 1995; Takada, 1995; Buccheri, 2002; O´Brien, 2007), razón por la cual se esperaría que las observaciones mal clasificadas correspondieran a pacientes con CPCP; no obstante, sólo 2 de los 8 casos pertenecen a este tipo histológico (el resto corresponde a adenocarcinomas). En este sentido, se construyó un modelo con RF en donde los grupos a clasificar fueron CPCNP y CPCP, sin embargo, el modelo mostró bajo rendimiento (resultados no mostrados). Los resultados no son significativos debido a que sólo el 19% del total de casos en nuestro estudio pertenece a CPCP.

En otro estudio, Gao y colaboradores (2005) evaluaron un microarreglo de 84 anticuerpos, en suero de pacientes con cáncer de pulmón, individuos sanos y pacientes con EPOC, y mediante análisis discriminante lineal encontraron un perfil de 4 proteínas: proteína C reactiva, amiloide de suero A, α-1-antitripsina y MUC1 (mucina unida a membrana), con valores de sensibilidad y especificidad de 70.83% y 92.85%, respectivamente. Los estudios anteriores (Gao, 2005; Patz, 2007) contrastan con los resultados del presente estudio, debido a que la α-1-antitripsina no mostró relevancia en nuestro modelo de clasificación propuesto (véase figura 11), lo que refuerza la idea de optimización de los paneles de proteínas en cada población en particular.

8.1 IMPLICACIONES BIOLÓGICAS DE LAS PROTEÍNAS INCLUIDAS EN EL MODELO DE CLASIFICACIÓN

CYFRA 21.1

El fragmento de citoqueratina 19[4], CYFRA 21.1, es quizá el biomarcador más ampliamente caracterizado con valor diagnóstico en CPCNP (Weiskopf, 1995). Estudios histopatológicos han demostrado que la citoqueratina 19 es abundante en carcinomas pulmonares. Niveles séricos anormales (>3.3 ng/ml) de dicho marcador tumoral han sido también encontrados en diversas enfermedades benignas, incluyendo fibrosis pulmonar, neumonía intersticial aguda, patologías del hígado y fallo renal (Molina, 2006; Nabi, 2009). De igual manera, CYFRA 21.1 se

[4]Las citoqueratinas son proteínas estructurales formadoras de subunidades de filamentos intermedios en el citoesqueleto celular de epitelios simples y pseudoestratificados (como el epitelio bronquial). Las citoqueratinas intactas son poco solubles, sin embargo, los fragmentos pueden ser detectados en suero (Karl, 2010). El fragmento de la citoqueratina 19 es reconocido específicamente por 2 anticuerpos monoclonales KS 19-1 y BM 19-21.

incrementa en neoplasias malignas diferentes a cáncer de pulmón, incluyendo tumores gastrointestinales, ginecológicos, mesoteliomas y de tipo urológico. Sin embargo, las mayores concentraciones se han encontrado en carcinomas de células no pequeñas, en particular de tipo escamoso. El rango de sensibilidad de CYFRA 21.1 va desde 30% hasta 75% para CPCNP, mientras que para CPCP es 20-60%. Lai y colaboradores (1995) entre otros autores, reportan niveles de corte entre 1.99 y 3.6 ng/ml; estos valores se encuentran por encima del punto de corte obtenido en el presente estudio (1.575 ng/ml) (Wieskopf, 1995; Schneider; 2002). Así también, ha sido evaluada la utilidad de CYFRA 21.1 para discriminar entre derrame pleural maligno y benigno, con niveles de sensibilidad que varían desde 38% hasta 82% (Toumbis, 1996; Satoh, 1995; Romero, 1996; Lai, 1999).

Estudios recientes en líneas celulares humanas, sugieren que las citoqueratinas 8, 18 y 19 presentan un papel en la progresión del tumor, ya que la expresión elevada de las mismas resulta en un incremento en las propiedades invasivas y metastásicas de las células o de su actividad migratoria (Nisman, 1998). Por otra parte, experimentos *in vitro* han mostrado que la liberación celular de fragmentos de citoqueratina 18 y 19 al espacio extracelular, ocurre como consecuencia de la digestión de caspasas (en particular, caspasa 3) durante el estadio intermedio de apoptosis, lo cual sugiere que la proteólisis de citoqueratinas mediada por caspasas probablemente facilite la formación de cuerpos apoptóticos y amplifique la señal apoptótica (Dohmoto, 2001).

CA125

El antígeno asociado a tumor CA125 es otro biomarcador de interés en neoplasias pulmonares. La mayoría de los estudios bioquímicos concluyen que CA125 es una glucoproteína de membrana, identificada en la línea celular OVCA 433 de cáncer ovárico seroso, la cual es reconocida por un anticuerpo monoclonal OC125 (Bast, 1981). Estudios recientes indican que CA125 es una molécula tipo mucina con un alto contenido de carbohidratos, así como un dominio citoplasmático carboxi-terminal con residuos de serina-treonina, véase anexo 5 (Yin, 2001; Kaneko; 2009). O´Brien y colaboradores (2009) indican que CA125 es fosforilada en el dominio C-terminal y proponen que podría estar ligada a la vía de transducción de señal del factor de crecimiento epidermal (EGF). Por otra parte, Kaneko y col. (2009) muestran que la

interacción entre CA125 y mesotelina[5] es significativa en la adhesión de células de cáncer de ovario a células mesoteliales presentes en la pared interna del peritoneo y en la superficie de otros órganos abdominales. Dicha adhesión célula-célula favorece la metástasis en cáncer ovárico.

El uso de CA125 para diagnóstico y seguimiento de cáncer ovárico ha sido bien definido, sin embargo, su expresión no se limita a este tipo de tumores; se ha encontrado en: carcinomas de pulmón, colon, páncreas, hígado, e incluso en tejido normal de origen urogenital (Diez, 1991; O´Brien, 2009). Para CPCNP, los niveles séricos de CA125 se han relacionado con el estadio del tumor, el tipo histológico y la tasa de sobrevivencia (Buccheri, 2002). De igual manera, pacientes con derrame pleural muestran un incremento en las concentraciones séricas de CA125, lo anterior es debido a que dicho antígeno normalmente está presente en células ectodérmicas del peritoneo y la pleura. Los niveles de corte reportados para CA125 en pacientes con CPCNP varían desde 15 hasta 35 U/ml (Diez, 1991; Casas, 1999; Polanski, 2006). Al igual que con CYFRA 21.1, el punto de corte óptimo obtenido en el presente estudio para CA125 es menor a lo reportado por la literatura (13.67 U/ml). Lo anterior reitera la necesidad de establecer puntos de corte específicos para nuestra población, de manera que dichos biomarcadores presenten mayor precisión en el diagnóstico de nuestros pacientes.

CRP

La proteína C reactiva es un reactante de fase aguda producido por hepatocitos estimulados por interleucina 6 (IL-6) como parte de la respuesta inflamatoria. Algunas referencias reportan concentraciones séricas en individuos sanos por debajo de 5 mg/l. Los niveles séricos de CRP reflejan la actividad de IL-6 y pueden ser útiles como un buen sustituto en la medición de dicha citocina. CRP forma homopentámeros, promueve la fagocitosis y complementa la fijación a fosforilcolina a través de uniones dependientes de calcio (anexo 6). Incrementos en los niveles de CRP son parte de la respuesta de fase aguda a la mayoría de formas de inflamación, infección, daño a tejido y neoplasias malignas. CRP también interactúa con el ADN e histonas para limpiar el material nuclear de células circulantes dañadas. Mientras que CRP es comúnmente poco eficiente como prueba simple, puede tener utilidad clínica como parte de un panel de

[5]La mesotelina es una glucoproteína anclada a glucosilfosfatidilinositol (GPI) presente en la membrana celular de tejido adulto normal como mesotelio. En contraparte, se sobre expresa en una gran variedad de cánceres como mesotelioma, cáncer ovárico, pancreático, e incluso en cáncer de pulmón de células pequeñas, especialmente en adenocarcinomas.

biomarcadores diagnóstico, especialmente en la evaluación de resultados de CT. CRP ha sido propuesto como un indicador pronóstico en carcinoma de esófago. Estudios han mostrado que CRP es un determinante independiente de sobrevivencia en CPCNP (Gao, 2005).

Heikkila y colaboradores en el 2007 analizaron 90 publicaciones, de las cuales sólo 5 estudios prospectivos proveen evidencia que CRP podría estar relacionado con cáncer de pulmón y colorectal. Sin embargo, ninguno proporciona evidencias del posible papel causal de CRP en cáncer. No obstante, Chaturvedi y colaboradores (2010) estimaron el riesgo absoluto de cáncer de pulmón a través de los niveles séricos de CRP en participantes del estudio de screening PLCO (Prostate, lung, colorectal and ovarian cancer screening trial), los individuos fueron fumadores actuales, fumadores que abandonaron el hábito al menos 15 años atrás e individuos sanos. Los resultados obtenidos muestran una relación entre los niveles elevados de CRP en fumadores actuales y la cantidad de cigarrillos fumada. Así también, es posible que a pesar de una exposición a tabaco similar, los niveles de CRP pueden identificar a los individuos con una respuesta inflamatoria mayor, y por consiguiente, mayor riesgo a padecer cáncer de pulmón. Además, se ha demostrado que la elevación de los niveles séricos de CRP se asocia con la incidencia de EPOC, así como con la progresión de displasias pulmonares, lo que indica que los niveles sistémicos de CRP pueden reflejar aspectos de la inflamación pulmonar, relevantes en la carcinogénesis broncogénica.

En conclusión, hasta el momento se desconoce si el incremento en la concentración sérica de CRP tiene alguna implicación directa en las células tumorales o, por otra parte, es consecuencia del microambiente creado por el tumor, por ejemplo la inflamación del tejido adyacente al tumor.

9. CONCLUSIONES Y PERSPECTIVAS

Los biomarcadores séricos en cáncer de pulmón pueden ser útiles como herramienta complementaria en pacientes con sospecha de dicha patología, sin embargo, los valores de sensibilidad y especificidad limitados impiden su aplicación como método diagnóstico. En el presente estudio fue posible identificar un panel óptimo que consta de tres proteínas, CYFRA 21.1, CA125 y CRP, el cual mostró mayor capacidad para discernir entre pacientes con cáncer de pulmón e individuos control, comparado con el rendimiento de cada proteína por separado. Por otra parte, nuestro análisis preliminar sugiere que el algoritmo RF presenta mayor utilidad para elucidar un conjunto de biomarcadores que permita incrementar la capacidad diagnóstica, comparado con marcadores individuales.

El modelo de clasificación obtenido por el algoritmo RF presenta una tasa de error de clasificación menor al 10%. A pesar de estimar la precisión del modelo por validación cruzada, es necesario evaluar el panel de proteínas en un nuevo conjunto de observaciones, con la finalidad de determinar qué tan reproducibles son los resultados obtenidos. Cabe señalar que existe un sesgo en el estudio, debido a la falta de pacientes con cáncer de pulmón en etapa temprana. Lo anterior es atribuible a la problemática actual en la detección temprana de dicha patología y, como consecuencia, la posibilidad de contar con un grupo de pacientes en etapas I y II es muy remota. No obstante, una alternativa viable es evaluar el panel en un estudio prospectivo de pacientes con EPOC, ya que dada la prevalencia de 40-70% de pacientes con EPOC diagnosticados con cáncer de pulmón (Congleton, 1995; Loganathan, 2006) sería posible determinar la capacidad del modelo para predecir aquellos pacientes con cáncer pulmonar, así como estimar el riesgo a desarrollar la patología basados en los niveles de las proteínas. Mientras que el panel identificado en el presente estudio requiere validación en una cohorte clínico más amplio, y podría no ser adecuado como una prueba independiente en el diagnóstico, su combinación con la TC podría proporcionar suficiente sensibilidad en el screening de cáncer de pulmón.

REFERENCIAS

1. Ashton RW, Jett JR. Screening for non-small cell lung cancer. Seminars in Oncology. 2005. 32: 253-258.
2. Bach PB, Kelley MJ, Tate RC, McCrory DC. Screening for lung cancer: a review of the current literature. Chest 2003.123:72–82.
3. Bain C, Feskanich D, Speizer FE, Thun M, Hertzmark E, Rosner BA, Colditz GA. Lung cancer rates in men and women with comparable histories of smoking. Journal of National Cancer Institute. 2004. 96(11):826
4. Barrett JH, Cairns DA. Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. Statistical Applications in Genetics and Molecular Biology 2008. 7(4)
5. Bast RC Jr, Feeney M, Lazarus H, Nadler LM, Colvin RB, Knapp RC. Reactivity of a monoclonal antibody with human ovarian carcinoma. The Journal of Clinical Investigation. 1981. 68: 1331–7
6. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. Journal of Computational Biology 2000, 7:559–583.
7. Breyman L. Random forest. Machine learning. 2001. 45(1):5-32
8. Black S, Kushner I, Samols D. C reactive protein. The Journal of Biological Chemistry. 2004. 279(47): 48487-48490.
9. Buccheri G and Ferrigno D. Lung tumor markers in oncology practice: a study of TPA and CA125. British Journal of Cancer. 2002. 87: 1112-1118.
10. Casas T, Tovar I, Bermejo J, Latour J, Parrilla P, Martínez P. Tumor markers in lung cancer: does the method of obtaining the cutoff point and reference population influence diagnostic yield?. Clinical Biochemistry. 1999. 32(6): 467-472.
11. Chaturvedi AK, Caporaso NE, Katki HA, Wong HL, Chatterjee N, Chanock SJ, Goedert J, Engels EA. C reactive protein and risk of lung cancer. Journal of clinical oncology. 2010. 28(16): 2719-2726.
12. Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. WIREs data mining and knowledge discovery.2011; 1:55-63.
13. Cho JY, Sung HJ. Proteomic approaches in lung cancer biomarker development. Expert Reviews Proteomics. 2009. 6:27-42.
14. Congleton J, Muers MF. The incidence of airflow obstruction in bronchial carcinoma, its relation to breathlessness, and response to bronchodilator therapy. Respiratory Medicine 1995; 89: 291–296.
15. Conrad DH, Goyotte J, Thomas PS. Proteomics as a method for early detection of cancer: A review of proteomics, exhaled breath condensate and lung cancer screening. Journal of General Internal Medicine, 2007. 23. 78-84.
16. Dalton WS, Friend SH. Cancer biomarkers- An invitation to the table. Science 2006. 312. 1165-1168.
17. Datta S. Classification of Breast Cancer versus Normal Samples from Mass Spectrometry Profiles Using Linear Discriminant Analysis of Important Features Selected by Random Forest. *Statistical Applications in Genetics and Molecular Biology.2008;* 7(7).
18. Diazgao R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3

19. Diez M, Cerdán FJ, Ortega MD, Torres A, Picardo A, Balibrea JL. Evaluation of serum CA125 as a tumor marker in non-small cell lung cancer. Cancer. 1991. 67:150-154.
20. Diez M, Torres A, Pollán M, Gómez A, Ortega D, Maestro ML, Granell J, Balibrea JL. Prognostic significance of serum CA-125 antigen assay in patients with non-small cell lung cancer. Cancer. 1994.73:1368-1376.
21. Dohmoto K, Hojo S, Fujita J, Yang Y, Ueda Y, Bandoh S, Yamaji Y, Ohtsuki Y, Dobashi N, Ishida T, Takahara J. The role of caspase 3 in producing cytokeratin 19 fragment (CYFRA 21-1) in human lung cancer cell lines. International Journal of Cancer. 2001. 91(4):468-73.
22. Dossat N, Mangé A, Solassol J, Jacot W, Lhermitte L, Maudelonde T, Daurès JP, Molinari N. Comparison of supervised classification methods for protein profiling in cancer diagnosis. Cancer Informatics. 2007. 3: 295-305.
23. Ezzati M, Henley SJ, Lopez AD, Thun MJ. Role of smoking in global and regional cancer epidemiology: current patterns and data needs. International Journal of Cancer. 2005. 116:963-971.
24. Farlow EC, Patel K, Basu S, Lee BS, Kim AW, Coon JS, Faber LP, Bonomi P, Liptay MJ, Borgia JA. Development of a multiplexed tumor associated autoantibody based blood test for the detection of non-small cell lung cancer. Clinical Cancer Research. 2010. 16(13): 3452-3462.
25. Farlow EC, Vercilla Ms, Coon JS, Basu S, Kim AW, Faber LP, Warren WH, Bonomi P, Liptay MJ, Borgia JA. A multi analyte serum test for detection of non small cell lung cancer. British Journal of Cancer. 2010. 103(8): 1221-1228.
26. Gao WM, Kuick R, Orchekowski R, Misek D, Qiu J, Greenberg A, Rom W, Brenner D, Omenn G, Haab B, Hanash S. Distinctive serum protein profiles involving abundant proteins in lung cáncer patients based upon antibody microarray analysis. BMC Cancer. 2005. 5(110).
27. Gaspar MJ, Diez M, Rodríguez A, Ratia T, Martin Duce A, Galván, Granell J, Coca C. Clinical value of CEA and CA-125 regarding relapsed and metastasis in resectable non-small cell lung cancer. Anticancer Research. 2003. 23:3427-3433.
28. Greenberg AK, Lee MS. Biomarkers for lung cancer: Clinical uses. Current Opinion Pulmonary Medicine. 2007. 13(4): 249-255.
29. Gurrola CM, González AE, Troyo R, Mendoza LA. Tipos histológicos y métodos diagnósticos en cáncer pulmonar en un centro hospitalario de tercer nivel. Gaceta Médica de México. 2009. 146: 97-101.
30. Hatzakis KD, Froudarakis ME, Bouros D, Tzanakis N, Karkavitsas N, Siafakas NM. Prognostic value of serum tumor markers in patients with lung cancer. Respiration 2002. 69:25-29.
31. Heikkila K, Ebrahim S, Lawlor DA. A systematic review of the association between circulating concentrations of C reactive protein and cancer. Journal of Epidemiology Community Health. 2007. 61:824-832.
32. Herbst RS, Heymach JV, Lippman SM. Molecular origins of cancer. The New England Journal of Medicine.2008. 359:13:1367-1380.
33. Jemal A, Bray F, Ferlay J, Ward E, Forman D. Global Cancer Statistics. CA A Cancer Journal for Clinicians.2011.
34. Kaneko O, Gong L, Zhang J, Hansen JK, Hassan R, Lee B, Ho M. A binding domain on mesothelin for CA125/MUC16. Journal of Biological Chemistry. 2009. 284(6): 3739-3749.

35. Lai RS, Chen CC, Lee PC, Lu JY. Evaluations of cytokeratin 19 fragment as a tumor marker in malignant pleural effusion. Japanese Journal of Clinical Oncology. 1999. 29 (9):421-424.
36. Lai RS, Hsu HK, Lu JY, Ger LP, Lai NS. CYFRA 21-1 enzyme-linked immunosorbent assay. Evaluation as a tumor marker in non-small cell lung cancer. Chest. 1996. 109(4):995-1000.
37. Lin RK, Hsieh YS, Lin P, Hsu HS, Chen CY, Tang YA, Lee CF, Wang YC. The tobacco specific carcinogen NNK induces DNA methyltransferase 1 accumulation and tumor suppressor gene hypermethylation in mice and lung cancer patients. The Journal of Clinical Investigation. 2010. 120 (2): 521-532.
38. Loganathan RS, Stover DE, Shi W, et al. Prevalence of COPD in women compared to men around the time of diagnosis of primary lung cancer. Chest 2006; 129: 1305–1312.
39. Malkinson AM. Primary lung tumors in mice: An experimentally manipulable model of human adenocarcinoma. Cancer Research 1992:52:2670s.
40. 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. Archives of Internal Medicine 2003; 163: 1475–1480.
41. McRedmond R, Logan PM, Lee M, Kenny D, Foley C, Costello RW. Screening for lung cancer using low-dose CT scanning. Thorax 2004. 59: 237–241.
42. McWilliams A, Mayo J, MacDonald S, leRiche JC, Palcic B, Szabo E, Lam S. Lung cancer screening: a different paradigm. American Journal of Respiratory and Critical Care Medicine 2003. 168:1143–1144.
43. Medina F, Salazar M. Frecuencia y patrón cambiante del cáncer pulmonar en México. Salud Pública de México. 2000. 42:333-336.
44. Molina R, Filella X, Augé JM. Tumor markers in lung cancer. European Oncological Disease. 2006.
45. Nabi E, Gomma N, Zeid A, Kantoush A, Ahmed M, Bushra S. Evaluation of Cyfra 21-1 as a diagnostic tool in lung cancer. Journal of Applied Science Research. 2009. 5(9): 1195-1201.
46. Nisman B, Lafair J, Heching N, Lyass O, Baras M, Peretz T, Barak V. Evaluation of tissue polypeptide specific antigen, CYFRA 21-1, the combined use of cytokeratin markers give any additional information?. Cancer. 1998. 82(10):1850-9.
47. O´Brien T, Beard J, Underwood L. CA125 gene and its use for diagnostic and therapeutic interventions. United States patent application publication. US 2009/0035819A1.
48. Paone G, Angelis de G, Munno R, Pallotta G, Bigioni D, Saltini C, Bisetti A, Ameglio F. Discriminant analysis on small cell lung cancer and non-small cell lung cancer by means of NSE and CYFRA 21.1. European Respiratory Journal.1995. 8:1136-1140.
49. Patz EF, Goodman PC, Bepler G. Screening for lung cancer. The New England Journal of Medicine. 2000. 343(22): 1627-1633.
50. Patz EF Jr, Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE 2nd. Panel of serum biomarkers for the diagnosis of lung cancer. Journal of Clinical Oncology. 2007. 25(35): 5578- 5583
51. Polanski M, Anderson NL. A list of candidate cancer biomarkers for targeted proteomics. Biomarker Insights. 2007. 1: 1-48.
52. Rastel D, Ramaioli A, Cornillie F, Thirion B. CYFRA 21-1, a sensitive and specific new tumor marker for squamous cell lung cancer. Report of the first European multicenter evaluation. European Journal of Cancer. 1994. 30(5): 601-606.

53. Rencher, A. Methods of multivariate analysis. Second edition. Wiley-Interscience, 2007. ISBN: 9780471271352.
54. Rivera MP, Detterbeck F, Mehta AC. Diagnosis of lung cancer: The guidelines. CHEST. 2003; 123: 129-136.
55. Romero S, Fernandez C, Arriero JM, Espasa A, Candela A, Martin C. CEA, CA 15-3, CYFRA 21-1 in serum and pleural fluid of patients with pleural effusions. European Respiratory Journal. 1996. 9: 17-23.
56. Rossi A, Maione P, Colantuoni G, Gaizo FD, Guerriero C, Nicolella D, Ferrara C, Gridelli C. Screening for lung cancer: New horizons?. Critical Reviews in Oncology Hematology. 2005. 56. 311-320.
57. Ruiz L, Rizo P, Sanchez F, Osornio A, García C, Meneses A. Mortality due to lung cancer in Mexico.2007. Lung Cancer. 58: 184-190.
58. Ruiz M, Rodríguez I, Rubio C, Revert C, Hardisson A. Efectos tóxicos del tabaco. Revista de toxicología. 2004. 21: 64-71.
59. Santos MJ, Curull V, Blanco ML, Macia F, Mojal S, Vila J y Broquetas JM. Características del cáncer de pulmón en un hospital universitario. Cambios epidemiológicos e histológicos en relación con una serie histórica. Archivos de Bronconeumología. 2005. 41(6):307-12.
60. Satoh H, Sumi M, Yagyu H, Ishikawa H, Sayama T, Naitoh T. Clinical evaluation of CYRA 21-1 in malignant pleural effusion. Oncology 1995. 52:211-14.
61. Schneider J, Bitterlich N, Velcovsky HG, Morr H, Katz N, Eigenbrodt E. Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer. International Journal of Clinical Oncology. 2002. 7(3):145-51.
62. Schreiber G, McCrory DC. Performance characteristics of different modalities for diagnosis of suspected lung cancer: Summary of published evidence. CHEST. 2003. 123: 115-128.
63. Skillrud DM, Offord KP, Miller RD. Higher risk of lung cancer in chronic obstructive pulmonary disease: a prospective matched controlled study. Annuals of International Medicine. 1986. 105:503–507.
64. Sone S, Yang ZG, Honda T, Maruyama Y, Takashima S, Hasegawa M, Kawakami S, Kubo K, Haniuda M, Yamanda T, Li F. Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. British Journal of Cancer 84:25-32, 2001
65. Spiro SG, Porter JC. Lung Cancer—Where Are We Today? Current Advances in Staging and Nonsurgical Treatment. American Journal and Respiratory Critical Care Medicine. 2002. 166: 1166-1196.
66. Statnikov A, Aliferis CF. Are Random Forests Better than Support Vector Machines for Microarray-Based Cancer Classification?. AMIA Annual Symposium Proceedings. 2007. 2007: 686–690.
67. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers: a different disease. Nature. 2007. 7:778-790.
68. Sung HJ, Cho JY. Biomarkers for the lung cancer diagnosis and their advances in proteomics. Biochemistry and Molecular Biology Reports.2008
69. Takada M, Masuda N, Matsuura E, Kusunoki Y, Matui K, Nakagawa K, Yana T, Tuyuguchi I, Oohata I and Fukuoka M. Measurement of cytokeratin 19 fragments as a marker of lung cancer by CYFRA 21-1 enzyme immunoassay. British Journal of Cancer. 1995. 71: 160-165.

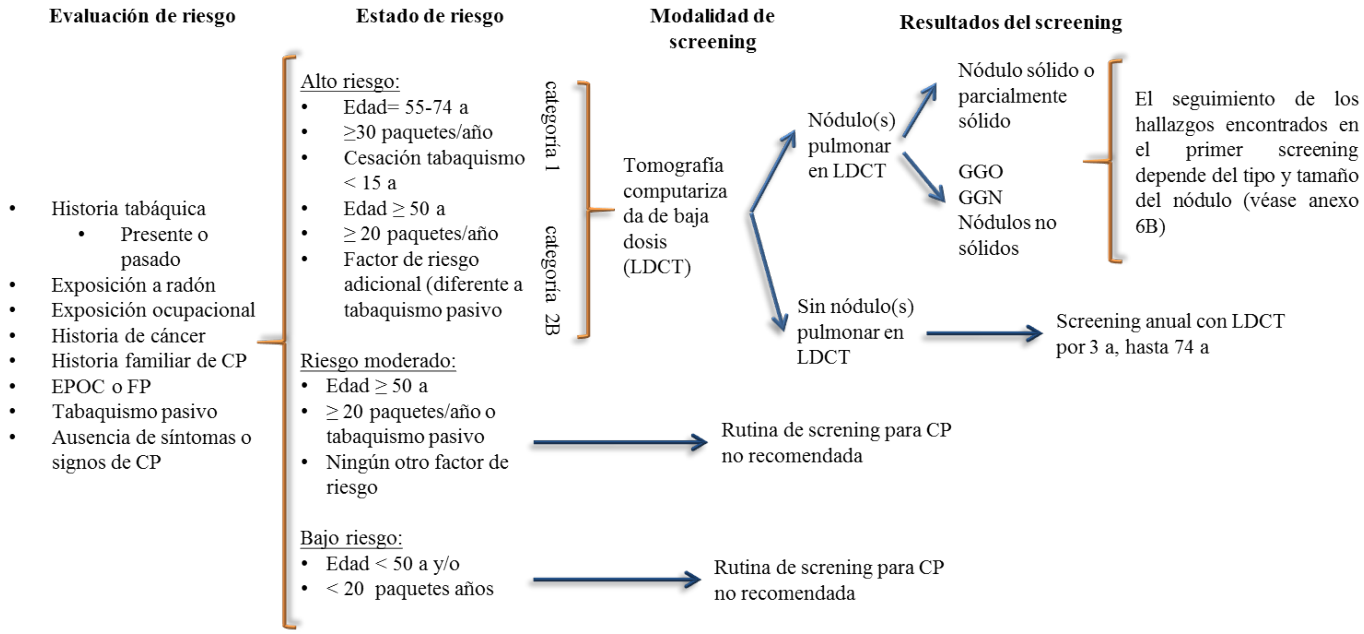
70. Toumbis M, Rasidakis A, Passalidou E, Kalomenidis J, Alchanatis M, Orphanidou D. Evaluation of CYFRA 21-1 in malignant and benign pleural effusion. Anticancer Research. 1996. 16:2101-4.
71. Travis WD, Brambilla E, Muller HK, Harris CC. Pathology and genetics of tumor of the lung, pleura, thymus and heart. International Agency for Research on Cancer. 2004. 2:10-124
72. Villalba J, Martínez R. Frecuencia del carcinoma broncopulmonar en pacientes fumadores y no fumadores diagnosticados en el Instituto Nacional de enfermedades Respiratorias en el año 2001. Revista del Instituto Nacional de Enfermedades Respiratorias. 2004. 17(1): 27-34.
73. Vlahou A, Schorge JO, Gregory BW, Coleman RL. Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. Journal of Biomedicine and Biotechnology. 2003. 2003(5):308–314.
74. Wieskopf B, Demangeat C, Purohit A, Stenger R, Gries P, Kreisman H, Quoix E. Cyfra 21-1 as a biologic marker of non- small cell lung cancer. Evaluation of sensitivity, specificity and prognostic role. Chest. 1995. 108(1):163-9.
75. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics. 2003. 19(13): 1636-1643.
76. Yang SY, Xiao XY, Zhang WG, Zhang LJ, Zhang W, Zhou B, Chen G, He DC. Application of serum SELDI proteomic patterns in diagnosis of lung cancer. BMC Cancer. 2005. 5(83).
77. Yin B, Lloyd K. Molecular cloning of the CA125 ovarian cancer antigen. Identification as a new mucin, MUC16. The Journal of Biological Chemistry. 2001. 276(29): 27371-27375.
78. Young RP, Hopkins RJ, Christmas T, Black PN, Metcalf P, Gamble GD. COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. European Respiratory Journal. 2009. 34: 380-386.
79. Zhang H, Yu CY, Singer B. Cell and tumor classification using gene expression data: Construction of forests. Proceedings of the National Academy of Sciences. 2003. 100(7):4168-4172.

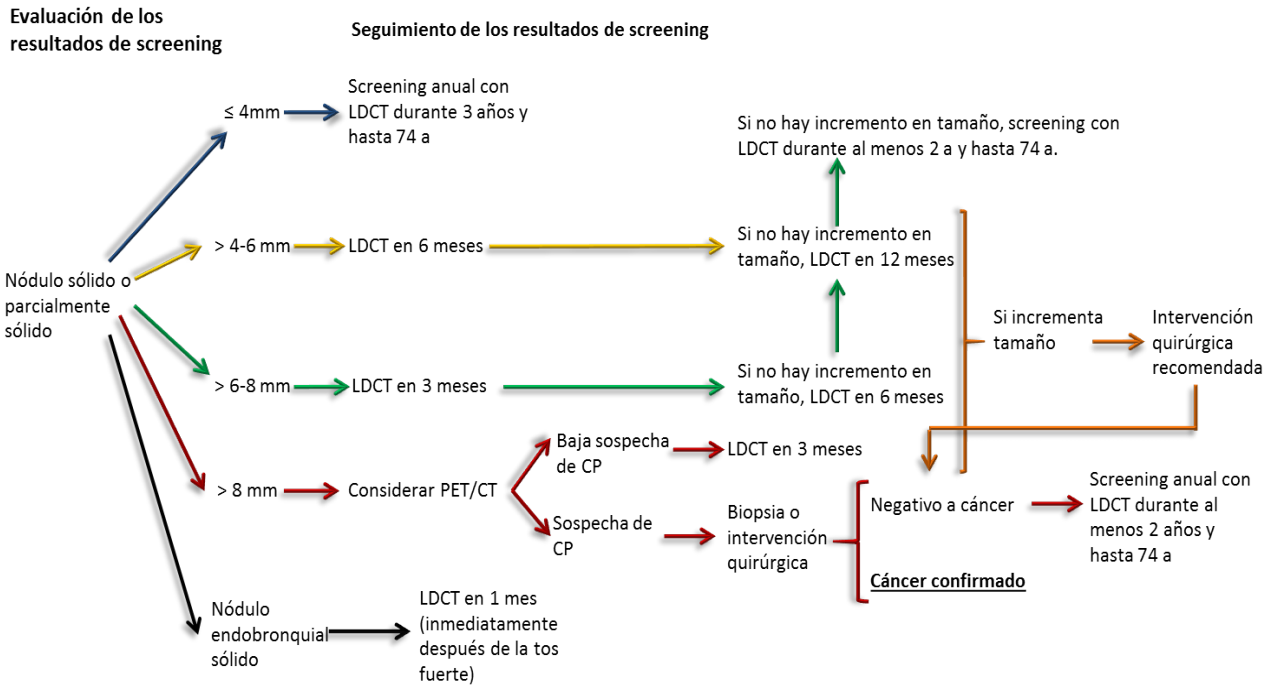
ANEXOS

Anexo 1. Sistema internacional de estadificación TNM para cáncer de pulmón.

Tumor primario		
TX	Células malignas +; lesión no visible	
T1	<3 cm diámetro	
T2	>3 cm diámetro, atelectasia distal	
T3	Extensión pleural parietal, diafragma o pericardio	
T4	Invasión a órganos mediastinales, derrame pleural maligno	
Nódulos linfáticos regionales		
N0	Sin nódulos	
N1	Nódulos hiliares o broncopulmonares ipsilaterales	
N2	Nódulos mediastinales, subcarinales o ipsilaterales	
N3	Nódulos supraclaviculares o hiliarmediastinales contralaterales	
Metástasis		
M0	ausencia	
M1	presencia	
Estadio	**TNM**	**Tasa sobrevivencia 5 años (%)**
IA	T1,N0,M0	>70
IB	T2,N0,M0	60
IIA	T1,N1,M0	50
IIB	T2,N1,M0	30
IIIA	T1-T3,N2,M0	10 a 30
IIIB	T4,N3,M0	<10
IV	M1	< 5

Anexo 2. (A) El siguiente esquema muestra las caracteristicas de selección de la población objetivo para screening de cáncer pulmonar, mediante tomografía computarizada de baja dosis (LDCT). **(B)** El seguimiento de los resultados obtenidos del primer screening depende del tamaño y tipo de nódulo, vease para nódulos sólidos o parcialmente sólidos. Para mayor detalle consulte NCCN Clinical Practice Guidelines in Lung Cancer Screening Version 1.2012. LDCT = Tomografía computarizada de baja dósis; a = años; GGO = ground glass opacity; GGN = ground glass nodule; CP = cáncer de pulmón; EPOC = enfermedad pulmonar obstructivo crónica.


Evaluación de riesgo
Estado de riesgo
Modalidad de screening
Resultados del screening
Historia tabáquica
Presente o pasado
Exposición a radón
Exposición ocupacional
Historia de cáncer
Historia familiar de CP
EPOC o FP
Tabaquismo pasivo
Ausencia de síntomas o signos de CP
Alto riesgo:
Edad= 55-74 a
≥30 paquetes/año
Cesación tabaquismo < 15 a
Edad ≥ 50 a
≥ 20 paquetes/año
Factor de riesgo adicional (diferente a tabaquismo pasivo
categoría 1
categoría 2B
Riesgo moderado:
Edad ≥ 50 a
≥ 20 paquetes/año o tabaquismo pasivo
Ningún otro factor de riesgo
Rutina de screning para CP no recomendada
Bajo riesgo:
Edad < 50 a y/o
< 20 paquetes años
Rutina de screning para CP no recomendada
Tomografía computarizada de baja dosis (LDCT)
Nódulo(s) pulmonar en LDCT
Nódulo sólido o parcialmente sólido
GGO
GGN
Nódulos no sólidos
El seguimiento de los hallazgos encontrados en el primer screening depende del tipo y tamaño del nódulo (véase anexo 6B)
Sin nódulo(s) pulmonar en LDCT
Screening anual con LDCT por 3 a, hasta 74 a


Evaluación de los resultados de screening
Seguimiento de los resultados de screening
Nódulo sólido o parcialmente sólido
≤ 4mm
Screening anual con LDCT durante 3 años y hasta 74 a
> 4-6 mm
LDCT en 6 meses
Si no hay incremento en tamaño, LDCT en 12 meses
Si no hay incremento en tamaño, screening con LDCT durante al menos 2 a y hasta 74 a.
> 6-8 mm
LDCT en 3 meses
Si no hay incremento en tamaño, LDCT en 6 meses
Si incrementa tamaño
Intervención quirúrgica recomendada
> 8 mm
Considerar PET/CT
Baja sospecha de CP
LDCT en 3 meses
Sospecha de CP
Biopsia o intervención quirúrgica
Negativo a cáncer
Screening anual con LDCT durante al menos 2 años y hasta 74 a
Cáncer confirmado
Nódulo endobronquial sólido
LDCT en 1 mes (inmediatamente después de la tos fuerte)

Anexo 3. Carta de consentimiento, aprobada por el Comité de Ética de la Institución participante, en la cual se da a conocer al donador el destino de la muestra, la finalidad del estudio y los riesgos que implica la toma de sangre.

CARTA DE CONSENTIMIENTO INFORMADO PARA PARTICIPACIÓN EN PROTOCOLOS DE INVESTIGACIÓN CLINICA

Lugar y Fecha _____

Por medio de la presente acepto participar en el protocolo de investigación titulado:

Análisis de la expresión de biomarcadores seleccionados específicos de cáncer de pulmón, y desarrollo de una prueba de ELISA prototipo para el diagnóstico de la enfermedad.

Registrado ante el Comité Local de Investigación o la CNIC con el número: _____

El objetivo del estudio es:

Analizar la expresión de un grupo de proteínas en sangre de personas con y sin cáncer de pulmón, para conocer qué conjunto de proteínas se asocian con el desarrollo de la enfermedad. El análisis de estas proteínas podría ayudarnos en un futuro a diagnosticar tempranamente a los pacientes con cáncer de pulmón.

Se me ha explicado que mi participación consistirá en:

1.- Se le tomará una muestra de sangre del antebrazo.

2.- Se le realizará una entrevista en donde se le solicitará información acerca de sus antecedentes médicos, estado de salud y antecedentes familiares.

Declaro que se me ha informado ampliamente sobre los posibles riesgos, inconvenientes, molestias y beneficios derivados de mi participación en el estudio, que son los siguientes:

Riesgos e inconvenientes: Los únicos riesgos posibles debidos a la toma de muestra de sangre incluyen solamente dolor moderado, sangrado o la formación de un moretón en la zona de punción que es el antebrazo.

Beneficios: este estudio no está diseñado para beneficiarle directamente en el corto plazo. Sin embargo, gracias a su participación, podremos estudiar la expresión de proteínas de cáncer de pulmón en sangre, lo que podría permitir en un futuro diagnosticar tempranamente a personas con esta enfermedad.

El Investigador Responsable se ha comprometido a darme información oportuna sobre cualquier procedimiento alternativo adecuado que pudiera ser ventajoso para mi tratamiento, así como a responder cualquier pregunta y aclarar cualquier duda que le plantee acerca de los procedimientos que se llevarán a cabo, los riesgos, beneficios o cualquier otro asunto relacionado con la investigación o con mi tratamiento.

Entiendo que conservo el derecho de retirarme del estudio en cualquier momento en que lo considere conveniente, sin que ello afecte la atención médica que recibo en el Instituto.

El Investigador Responsable me ha dado seguridades de que no se me identificará en las presentaciones o publicaciones que deriven de este estudio y de que los datos relacionados con mi privacidad serán manejados en forma confidencial. También se ha comprometido a proporcionarme la información actualizada que se obtenga durante el estudio, aunque esta pudiera cambiar de parecer respecto a mi permanencia en el mismo.

_____

Nombre y firma del paciente

_____

Nombre, firma y matrícula del Investigador Responsable.

Dr. Francisco Sánchez Llamas

Matrícula 99140833

Números telefónicos a los cuales puede comunicarse en caso de emergencia, dudas o preguntas relacionadas con el estudio:

Dr. Francisco Sánchez Llamas: Teléfono 3668 3000 Exts. 31326 y 31414.

Testigos

Nombre completo:

Parentesco con el paciente:

Dirección:

Teléfono:

Firma: Fecha:

Nombre completo:

Parentesco con el paciente:

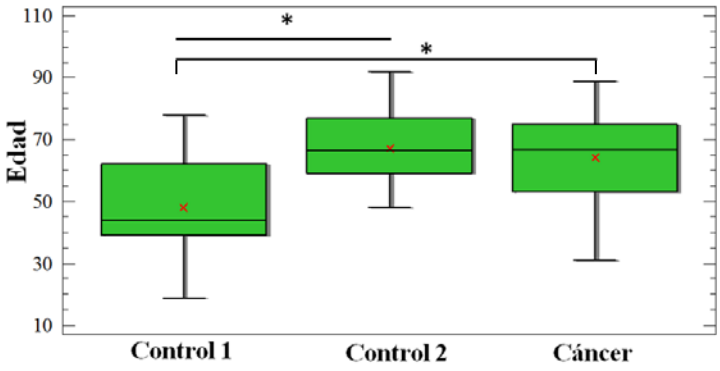
Dirección:

Teléfono:

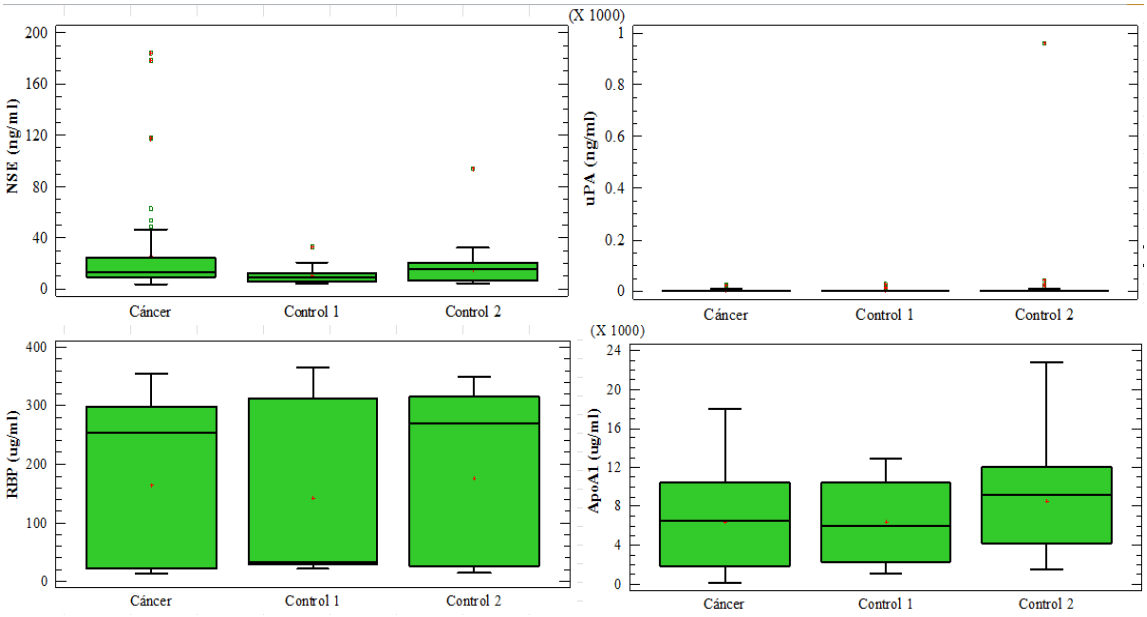
Firma: Fecha:

Clave: 2810 – 009 – 013

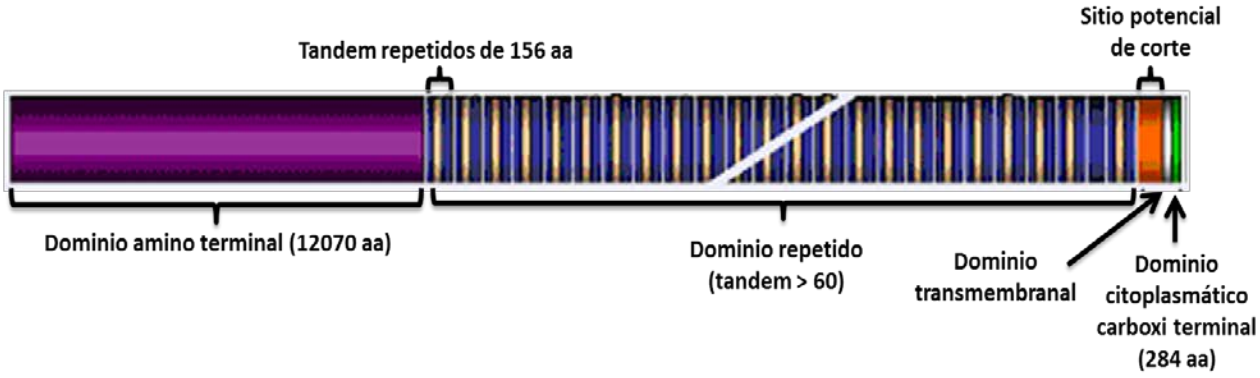
Anexo 4. Gráfico de caja y bigote de las edades en los tres grupos de estudio. Los datos provienen de una distribución normal, por lo tanto se aplicó una prueba *t*. Las diferencias significativas entre las medias de los grupos se muestran con barras sobre las cajas con * = valor P< 0.01.



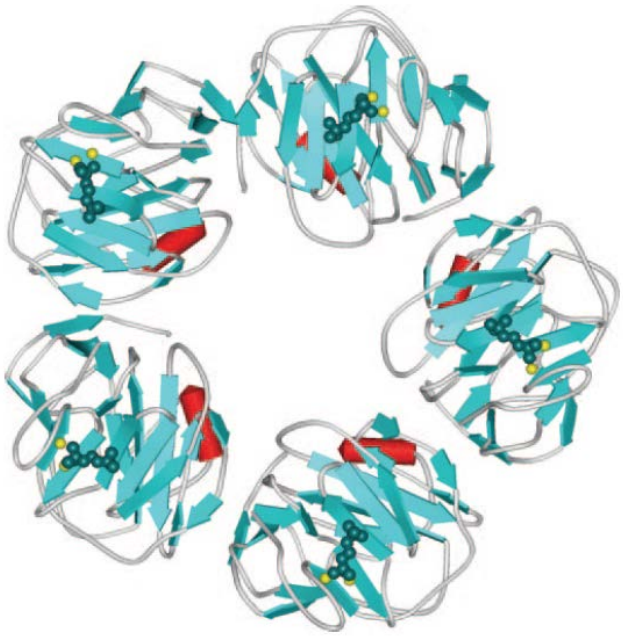
Anexo 5. Gráficas de caja y bigote de las proteínas que no presentaron diferencias significativas en las concentraciones entre los grupos de estudio.



Anexo 6. La figura muestra la organización estructural de la proteína CA125. El epitope para el anticuerpo anti-CA125 (OC125) se cree está presente en una pequeña región de anillo de cisteína en la región de repetidos en tándem. Los dominios extracelulares contienen numerosos sitios de O-glicosilación y N-glicosilación.



Anexo 7. Estructura cristalina del complejo proteína C reactiva y fosfocolina. Los iones de calcio se muestran en amarillo y la fosfocolina en azul fuerte. Diagrama de rayos-x reportado por Black y colaboradores (2004).



ANÁLISIS MULTIVARIADO DE UN CONJUNTO DE PROTEÍNAS SÉRICAS PARA MEJORAR LA CAPACIDAD DIAGNÓSTICA DE MARCADORES INDIVIDUALES EN CÁNCER DE PULMÓN

G. Leal[a], M.G. González[a], J.M. Flores[a], F. Sánchez[b], A. Rojas[c], P.A. Cabrera[c], M. Martínez[a]

[a]Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco/ Unidad de Biotecnología Médica y Farmacéutica, Guadalajara, Jalisco, gislealp@hotmail.com
[b]OPD Antiguo Hospital Civil, Fray Antonio Alcalde/ Servicio de Fisiología Pulmonar e Inhaloterapia, Guadalajara, Jalisco, drfcosanchez@yahoo.com
[c]Centro Oncológico Estatal ISSEMyM/ Coordinación de Cirugía Oncológica, Toluca, México.

RESUMEN
Antecedentes: Actualmente, los marcadores utilizados de manera individual, carecen de suficiente sensibilidad y especificidad para el diagnóstico, estratificación, pronóstico o respuesta a la terapia en pacientes con cáncer de pulmón. En el presente estudio se analizó un conjunto de marcadores tumorales previamente reportados con valor diagnóstico y/o pronóstico para cáncer de pulmón, con el propósito de mejorar la capacidad diagnóstica de los marcadores individuales.
Métodos: Se evaluó la concentración sérica de 14 proteínas en 3 grupos de estudio, compuestos por 20 pacientes con cáncer de pulmón, 30 con Enfermedad Pulmonar Obstructiva Crónica (EPOC) y 17 fumadores, utilizando kits de ELISA comerciales. Las concentraciones obtenidas fueron analizadas por estadística descriptiva y métodos estadísticos multivariados como Análisis de Componentes Principales (ACP), Análisis Discriminante (AD) y Curvas ROC.
Resultados: Del total de proteínas evaluadas, 7 presentaron diferencias estadísticamente significativas entre grupos, con un nivel de significancia del 95%. Al comparar los resultados de los diferentes métodos estadísticos multivariados, fue posible identificar un panel compuesto por: CYFRA 21-1, NSE, CEA, CA-125, MMP-9, CRP y YKL-40. CYFRA 21-1 mostró ser el mejor marcador individual, con una sensibilidad del 75% (a una especificidad de 80%), mientras que con el panel de proteínas fue posible incrementar la sensibilidad 11.4% (86.4%), al mismo nivel de especificidad.
Conclusiones: Es posible incrementar la capacidad diagnóstica de los marcadores tumorales individuales, a través de su análisis en combinación, utilizando las herramientas estadísticas adecuadas.

1. INTRODUCCIÓN

A nivel mundial, el cáncer de pulmón es el cáncer más común en términos de nuevos casos diagnosticados y el causante del mayor número de muertes asociadas a neoplasias malignas en hombres. En mujeres, fue el cuarto cáncer mayormente diagnosticado y la segunda causa de muerte por cáncer según lo reportado en el 2008[1]. En México, durante el 2008, la tercera causa de muerte se le atribuyó a tumores malignos con 67 048 defunciones, de las cuales, los tumores de tráquea, bronquios y pulmón constituyeron la primera causa de muerte por neoplasias malignas con 6 843 defunciones[2].
La asociación entre tabaco y cáncer de pulmón ha sido bien establecida; el riesgo relativo se incrementa de 10 a 20 veces en fumadores, comparado con no fumadores; alrededor del 85% de casos de cáncer de pulmón son atribuibles al tabaco. Estudios epidemiológicos han identificado otros factores de riesgo como la contaminación ambiental, combustión de leña y carbón, exposición a fibras de asbesto, sílica, arsénico, cadmio, níquel, cromo e incluso masa de partículas radiactivas, así también, factores hormonales, genéticos y virales como el Virus del Papiloma Humano (VPH).[3]

La distribución de los tipos histológicos constituye un parámetro importante en el tratamiento, evolución y pronóstico del cáncer pulmonar. En México, se ha reportado al adenocarcinoma como el tipo histológico más común, seguido del carcinoma de células escamosas y carcinoma de células pequeñas, siendo el carcinoma de células grandes el menos frecuente. Asimismo, el tratamiento y la tasa de sobrevivencia están en función del estadio de la enfermedad. Desafortunadamente, más del 75% de los pacientes son diagnosticados en etapas tardías, lo que conlleva a una tasa de sobrevivencia a 5 años de aproximadamente 15%, en el mejor de los casos[4]. Hasta el momento, la única opción curativa es la resección quirúrgica, por lo tanto, una detección temprana puede disminuir potencialmente la mortalidad del cáncer de pulmón.

El procedimiento diagnóstico en pacientes con sospecha de cáncer de pulmón involucra técnicas de imagenología como la Tomografía Axial Computarizada (TAC), la cual presenta una sensibilidad del 60% y una especificidad del 80%, así como métodos invasivos, fibrobroncoscopía (FBC), biopsia y cepillado bronquial. Recientemente se han propuesto diversos biomarcadores para diagnóstico y/o pronóstico, los cuales pueden ser medidos directamente en el tejido o en fluidos biológicos como suero, líquido pleural y lavado broncoalveolar. Sin embargo, la mayoría de ellos no presentan suficiente sensibilidad y especificidad para su uso en la clínica. Por lo tanto, en el presente estudio se evaluaron 14 biomarcadores tomados de la literatura por su utilidad en la discriminación entre individuos con cáncer de pulmón y población control, los cuales fueron cuantificados en un total de 67 individuos (cáncer de pulmón n=20, EPOC n=30, fumadores= 17) con el fin de identificar un panel de biomarcadores que cuente con valores de sensibilidad y especificidad superiores, comparado con el mejor biomarcador individual.

2. MATERIALES Y MÉTODOS

Se reclutaron 20 pacientes con diagnóstico histopatológico confirmado de cáncer pulmonar, 30 pacientes con Enfermedad Pulmonar Obstructiva Crónica (EPOC) como grupo control, y 17 fumadores sanos como segundo grupo control. Los grupos control se seleccionaron con base en características demográficas similares (edad y sexo) a los pacientes con cáncer de pulmón, y a los 2 principales factores de riesgo: tabaquismo e inflamación. Se obtuvieron 2 tubos de 10 ml de sangre periférica, extraída de la vena cubital tanto de pacientes con cáncer de pulmón, como de los individuos control. Previo a la toma de muestra, el donador firmó la carta de consentimiento, aprobada por el Comité de Ética de la Institución de Salud participante, en la cual se da a conocer el destino de la muestra, la finalidad del estudio y los riesgos que implica la toma de sangre. Las muestras se centrifugaron a 3000 rpm por 10 min y el suero se alicuotó y almacenó a -70°C.

La cuantificación de los 14 biomarcadores circulantes se realizó mediante la técnica de ELISA (Enzyme-linked Immunosorbent Assay), utilizando kits comerciales. Las concentraciones de todas las proteínas se calcularon utilizando el Software MasterPlex 2010. Los datos de concentración se analizaron mediante los siguientes métodos estadísticos:

- **a)** Estadística descriptiva (mediana, rangos, diagramas de caja y bigote) fue útil para la evaluación individual de los biomarcadores. Por medio de la Prueba de Kruskall- Wallis fue posible evaluar las diferencias en las concentraciones de los marcadores entre los diferentes grupos de estudio, a un intervalo de confianza del 95%.
- **b)** Curvas ROC (Receiver Operating Characteristic) para determinar parámetros como sensibilidad, especificidad y área bajo la curva (AUC).
- **c)** Se realizaron dos tipos de análisis multivariado, Análisis Discriminante y Análisis de Componentes Principales, con el fin de escoger el panel óptimo de biomarcadores. El criterio de inclusión para los marcadores individuales del panel fue la suma de rangos de Kruskall-Wallis con P-value menor a 0.05, es decir, diferencia estadística entre las distribuciones de los grupos de estudio, así como un área bajo la curva ROC mayor a 0.60.

3. RESULTADOS

La concentración de cada uno de los marcadores fue determinada en el suero de 20 pacientes con cáncer de pulmón, de 30 pacientes con EPOC y de 17 donadores fumadores. Las concentraciones en suero de MMP-9, CYFRA 21.1, CRP, CEA, CA 125, YKL-40 y NSE, fueron significativamente mayores en el grupo de cáncer a un nivel del significancia del 95% (véase figura 1). Mediante el análisis de las curvas ROC se encontraron 7 proteínas con área bajo la curva (AUC) > 0.60 (P-value<0.001), CA 125,

CEA, MMP-9, CYFRA 21.1, CRP, YKL-40 y NSE. Cabe resaltar que el biomarcador con mayor AUC es CYFRA 21.1, resultado que concuerda con lo reportado en la literatura. La figura 2 muestra las curvas ROC de los biomarcadores individuales con relevancia estadística.

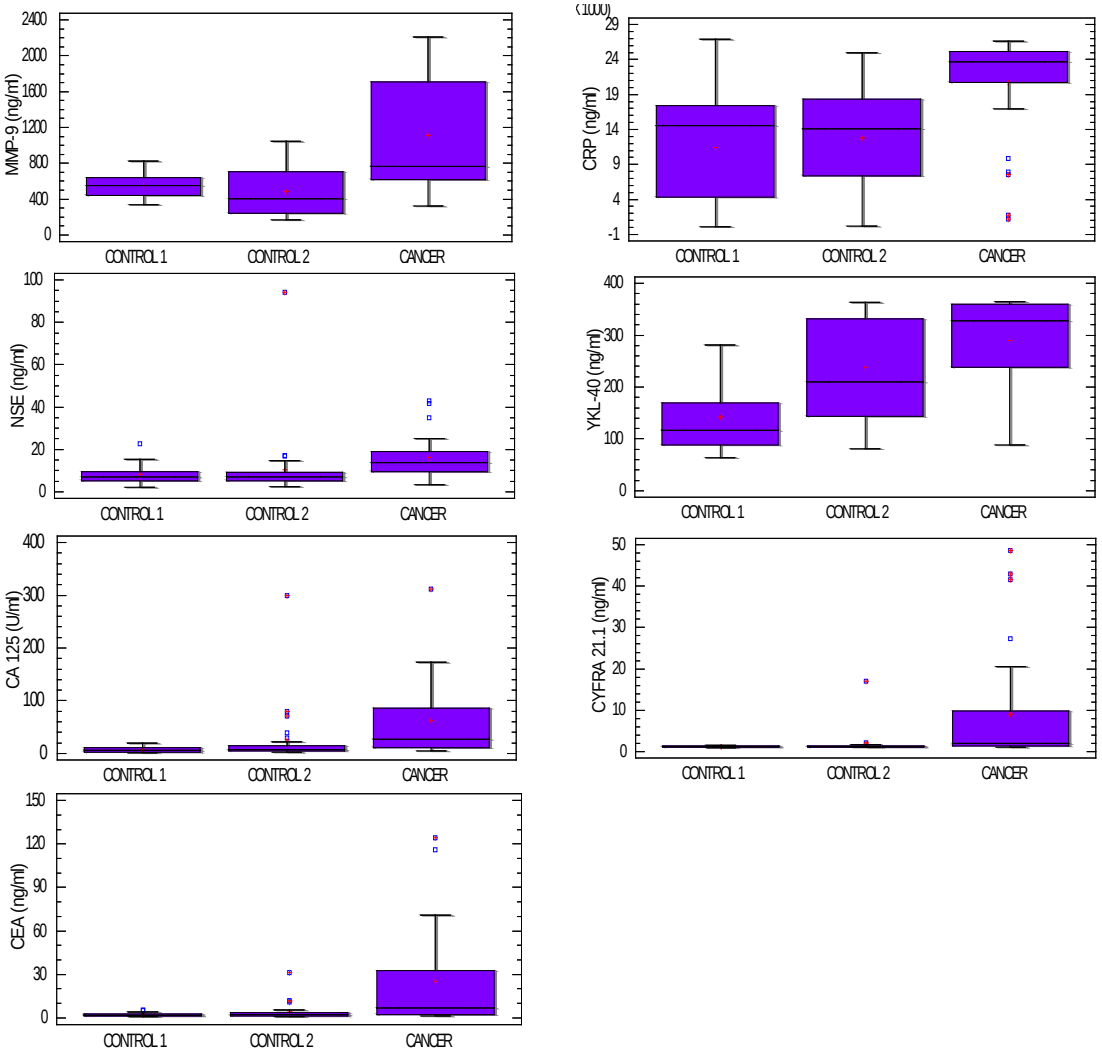


Figura 1. Comparación de medianas por el método de Kruskall-Wallis, a un nivel de significancia del 95%

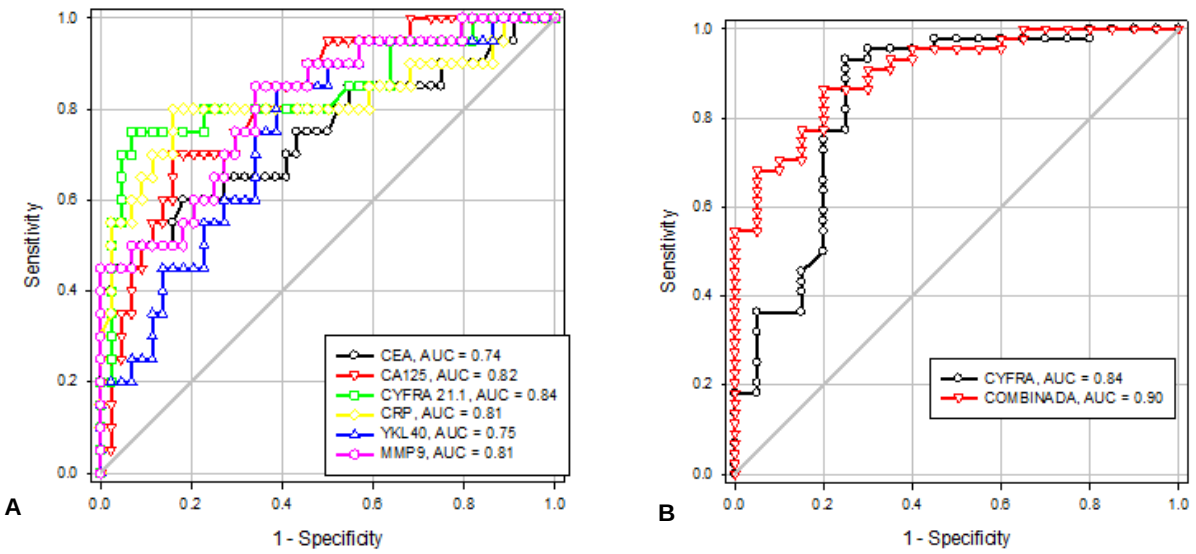


Figura 2. A) Curvas ROC de los biomarcadores con mayor área bajo la curva, siendo CYFRA 21.1 el mejor biomarcador individual (AUC= 0.84). **B)** Comparación de curvas ROC entre el panel de marcadores y CYFRA 21.1.

Con el fin de diferenciar los individuos de cada grupo con base en los niveles de los biomarcadores, se realizó un análisis de regresión discriminante linear con el software StatGraphics Plus. El análisis discriminante permite encontrar las combinaciones de proteínas, conocidas como *vectores discriminantes*, los cuales pueden discriminar entre los diferentes grupos de estudio (Cáncer, Control 1 y Control 2), véase figura 3.

Una vez obtenidos los vectores discriminantes, se analizaron los valores absolutos de los coeficientes que adquiere cada variable (proteína) y que representan la influencia de las variables correspondientes en la separación de los individuos de estudio. Con base en los coeficientes se designan las proteínas candidatas a mejores biomarcadores. Del total de proteínas se eligieron a CYFRA 21.1, CEA, CA 125, CRP, MMP-9, NSE y YKL-40 como el panel de biomarcadores para diferenciar entre el grupo de cáncer y los grupos control. Los resultados del Análisis Discriminante fueron comparados con un Análisis de Componentes Principales (PCA) con el fin de escoger las variables con mayor influencia en la clasificación de los grupos (véase figura 3). Una vez seleccionado el panel, se construyó la curva ROC combinando las 7 proteínas, obteniendo un área bajo la curva mayor (AUC= 0.90) comparado con el mejor biomarcador individual, CYFRA 21.1 (AUC= 0.84). A una especificidad del 80% CYFRA 21.1 mostró una sensibilidad del 75%, mientras que con el panel de biomarcadores fue posible incrementar 11.4% la sensibilidad (86.4%), al mismo nivel de especificidad.

4. CONCLUSIONES

La fibrobroncoscopía sigue siendo el principal método diagnóstico en el cáncer pulmonar, especialmente en lesiones que se encuentran visibles en la mucosa. Detterbeck y Rivera[6], considerando cifras de ocho publicaciones, encontraron que la sensibilidad promedio de la FBC en tumores centrales fue, para biopsia 83%, cepillado 64% y lavado 48%; en tumores periféricos, la biopsia (transbronquial) 60%, cepillado 48% y lavado 37%. Así mismo, se ha propuesto un gran número de biomarcadores como proteínas, microRNAs, autoanticuerpos, entre otros, sin embargo, carecen de suficiente sensibilidad y especificidad de manera individual.

En el presente estudio fue posible incrementar la capacidad diagnóstica de los biomarcadores individuales, a través de su análisis en combinación, utilizando las herramientas estadísticas adecuadas.

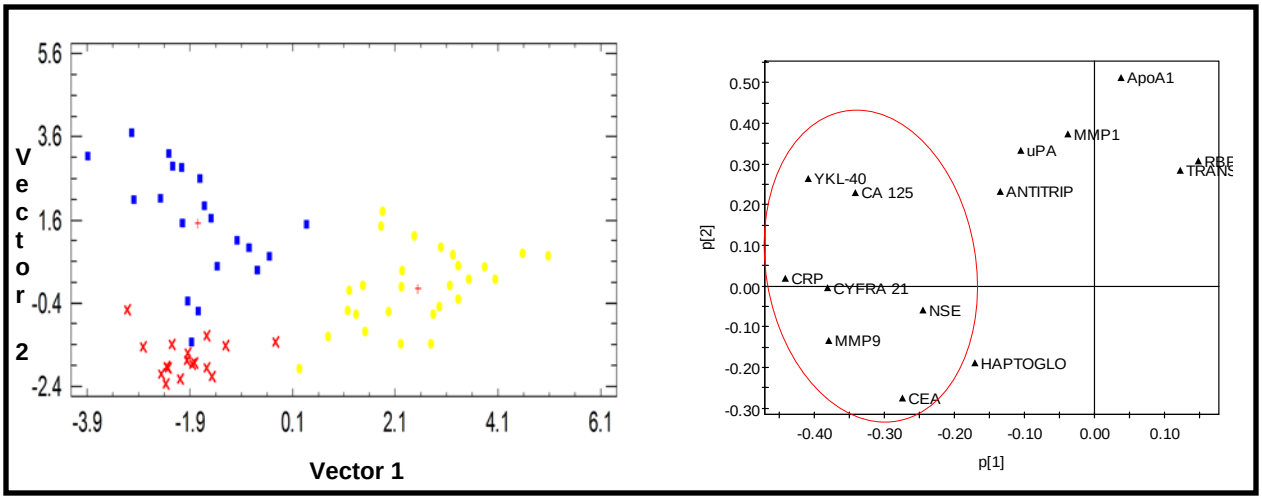


Figura 3. De lado izquierdo se muestra la representación gráfica del Análisis Discriminante, donde los ejes corresponden a los 2 vectores discriminantes que logran agrupar a los individuos de estudio, usando el total de biomarcadores. El gráfico derecho muestra los coeficientes adquiridos por cada biomarcador en el correspondiente vector, de los cuales, los encerrados en rojo lograron describir mejor al grupo de cáncer.

BIBLIOGRAFÍA

A. Jemal, F. Bray, J. Ferlay, E. Ward, D. Forman, "Global Cancer Statistics". *CA Cancer J. Clin.*2011.
Instituto Nacional de Estadística, Geografía e Informática "Boletín de estadísticas vitales", 2008.

S. Sun, J.H Schiller, A.F. Gazdar, "Lung cancer in never smokers, a different disease", Nature Vol. 7, 2007, pp. 778-790.

A. Rossi, P. Maione, G. Colantuoni, F. Gaizo, D. Nicolella, C. Ferrara, C. Gridelli, "Screening for lung cancer: New horizons?" Critical Reviews in Oncology Hematology, Vol. 56, 2005, pp. 311-320.

EC Farlow, M.S Vercillo, J.S Coon, S. Basu, A.W. Kim, L.P Faber, W.H Warren, P. Bonomi, M.J Liptay, J.A Borgia, "A multy-analyte serum test for the detection of non-small cell lung cancer" British Journal of Cancer Vol. 103, 2010, pp. 1221-1228.

F.C Detterbeck, P. Rivera "Clinical presentation and diagnosis", En: F.C Detterbeck, M.A Socinski, M.P Rivera, J.G Rosenman, editors. Diagnosis and treatment of lung cancer, Philadelphia: Saunders, 2000, pp. 54-56.