Trabajo Práctico Final

Diseño de una investigación basado en una base de datos

Materia: Bioestadística

Alumnos:

- Axel Cesar Wood Niella
- Lucía Bernochi
- Priscilla Vanesa Tenas Vai

Elección de base de datos

Presentación del diseño de una investigación basado en una base de datos de interés que en este caso se trató de biomarcadores urinarios del cáncer de páncreas y posterior análisis estadístico de la misma.

TABLE OF CONTENTS

01.

Introducción

Introducción

Importancia

La tasa de supervivencia a cinco años es inferior al 10%.

Actualidad

No existen biomarcadores específicos con la capacidad de generar una detección temprana.

Objetivo

Encontrar una prueba de diagnóstica capaz de identificar posibilidades de desarrollar la enfermedad.

Objetivo

Creatinina

La elevación implica menor perfusión renal provocada en los casos graves de pancreatitis.

LYVE1

Receptor que tiene un papel activo en la linfangiogénesis y remodelación endotelial.

REGIB

Actúa como factor de crecimiento en la regeneración de los islotes pancreáticos.

TFF1

Factor involucrado en el desarrollo y la progresión de varios tipos de cáncer.

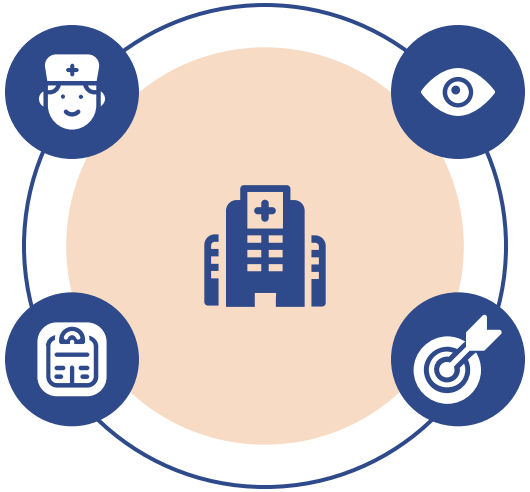
Pregunta PICO

Población (P)

Hombres y mujeres con un rango etario entre 26 y 89 años con historias clínicas diferentes.

Intervención (I)

Medición de biomarcadores.



Comparación (C)

Pacientes con adenocarcinoma ductal pancreatico vs. sin.

Outcomes (O)

Presencia o no de cáncer pancreático.

02.

Métodos

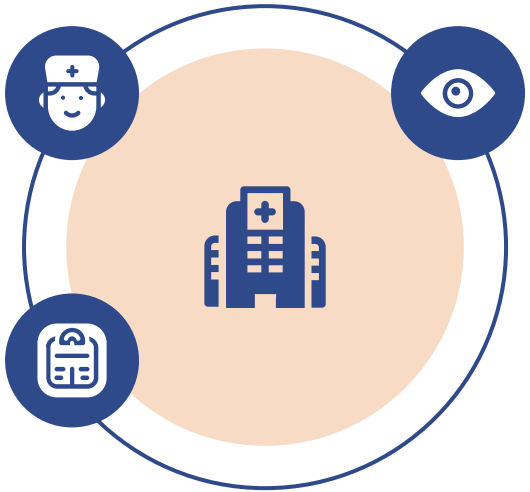
Diseño del estudio

Casos y controles

Baja prevalencia

Hospital Boston EEUU

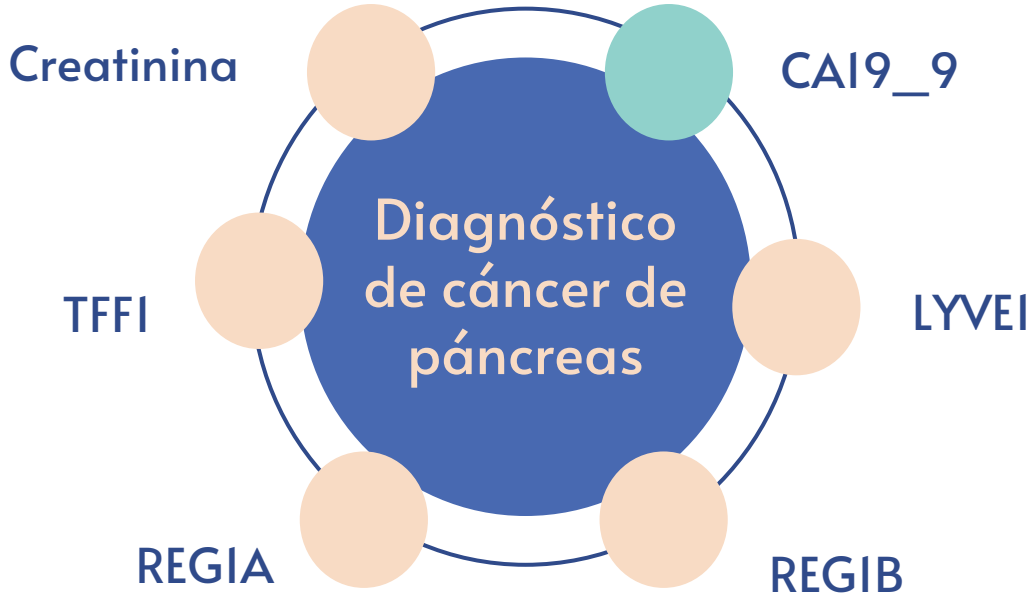
Previo a intervención quirúrgica o quimioterapéutico



Muestras

- Preservadas según procedimiento estándar.
- Sin afecciones pancreáticas
- Sin neoplasias malignas conocidas
- Adenocarcinoma ductal pancreatico
- Grupo benigno

Variables


Creatinina
CA19_9
TFF1
Diagnóstico de cáncer de páncreas
LYVE1
REGIA
REGIB

Data sources / measurement

Kit de Elisa

Grupos de 40

Muestras duplicadas para confirmación de resultados

Creatinina

ILab Aries del Laboratorio de Instrumentación

mmol/l

REGIB

TMB Substrate Set y Stop Solution de BioLegend

FLUOstar Omega Microplate Reader

Límites de detección:

CA19_9:0.3 U/ml

TFF1: 3.91 pg/ml

REG1B: 8 pg/ml

LYVE1: 56 pg/ml

Sesgos

Sesgo de Selección

Se usaron muestras de pacientes entre 26 y 89 años.

Las funciones renales de pacientes mayores de edad pueden no funcionar de la manera esperada

Tamaño del estudio

Determinado por la cantidad de personas que se sometieron una intervención quirúrgica o a un tratamiento quimioterapéutico para el tratamiento de cancer de pancreas durante 6 meses

Sexo

Se tomaron muestras tanto de hombres como mujeres

Edad

Se tomaron muestras de pacientes en el rango etario de interés

Muestras benignas

- Pancreatitis Cronica
- Casos de dolor Abdominal
- Enfermedades de vesícula biliar

Variables Cuantitativas

Variables Explicativas

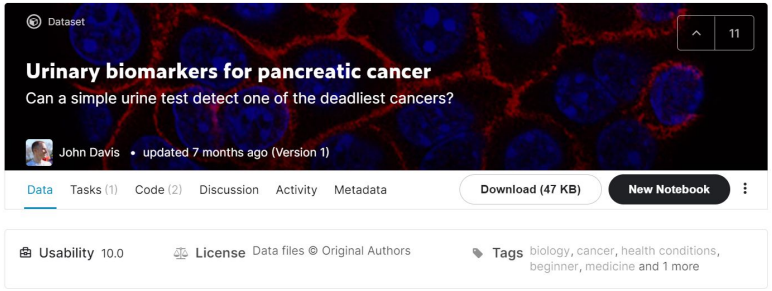
Fueron agrupadas dependiendo del diagnóstico del paciente. No se agrupo por edad.

Analisis Multivariado

Las variables fueron agrupadas en dos grupos utilizando un valor umbral

03.

Data Set



https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

1

Controles Saludables

2

Pacientes con afecciones pancreáticas no cancerosas, como pancreatitis crónica

3

Pacientes con adenocarcinoma ductal pancreático

Datos faltantes < 20%

En su lugar se reemplazó la media correspondiente a cada variable

04.

Resultados

Tabla 0: Identificación de variables, su forma de medirlas y su distribución.

Tabla 0: Identificación de variables, su forma de medirlas y su distribución.

Variable	Tipo
Sex	factor
Age	integer
Creatinine	numeric
LYVE1	numeric
REG1B	numeric
TFF1	numeric
Diagnosis	factor

Variable: Sex

	No	Si
Fem	68	12
Masc	29	16

Variable explicatoria cualitativa nominal categórica que en consecuencia posee una distribución binomial.

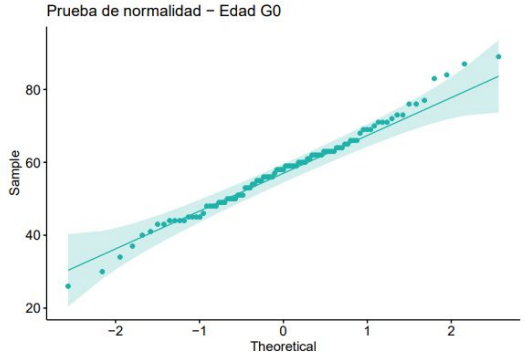
Comparación mediante TEST CHI CUADRADO.

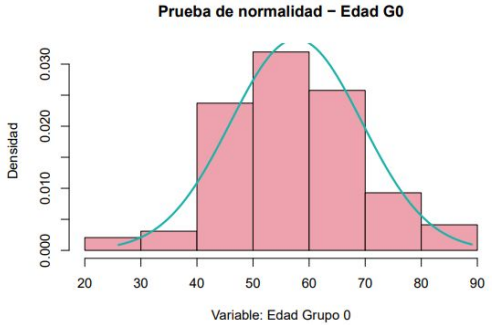
Variables: Age, Creatinine, LYVE1, REG1B, TFF1

En primer lugar: Análisis de normalidad.

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9901, p-value = 0.6925
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.058762, p-value = 0.5633
```

Grupo 0

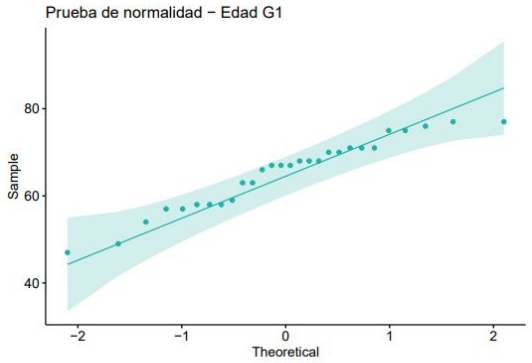




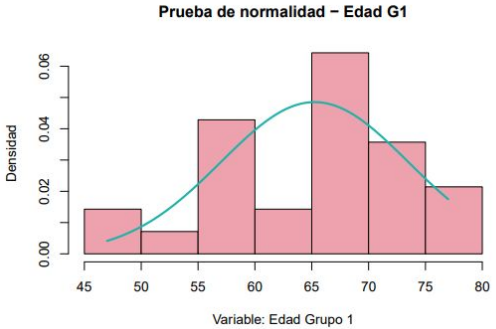
Variables: Age, Creatinine, LYVE1, REG1B, TFF1

En primer lugar: Análisis de normalidad.

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9459, p-value = 0.1561
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.1557, p-value = 0.08022
```



Grupo 1



Variables: Age, Creatinine, LYVE1, REG1B, TFF1

En segundo lugar: Comparación

Distribución Normal

Para ambos grupos la distribución fue del tipo Normal.

Test Z

Condiciones:

- ➔ Homocedasticidad
- ➔ Muestra mayor a 30

Tabla resumen

Variable	Respuesta o explicatoria	Tipo	Unidad	Dist. de Prob.	Test
Sex	Explicatoria	Cual. Dicotómica	-	Binomial	Chi Cuadrado
Age	Explicatoria	Cuant. cont.	Años	Normal	Test Z
Creatinine	Explicatoria	Cuant. cont.	mmol/l	No normal	Wilcoxon
LYVE1	Explicatoria	Cuant. cont.	pg/ml	No normal	Wilcoxon
REG1B	Explicatoria	Cuant. cont.	pg/ml	No normal	Wilcoxon
TFF1	Explicatoria	Cuant. cont.	pg/ml	No normal	Wilcoxon
Diagnosis	Respuesta	Cual. Dicotómica	-	Binomial	-


Tabla 1:
Análisis
Univariado

Tabla 1: Analisis Univariado

Chi Cuadrado

H0: Se trata de variables independientes.

H1: No se trata de variables independientes

Test Z

$H0 : \mu Grupo0 - \mu Grupo1 = 0$

$H1 : \mu Grupo0 - \mu Grupo1 \neq 0$

Wilcoxon

$H0 : \theta Grupo0 - \theta Grupo1 = 0$

$H1 : \theta Grupo0 - \theta Grupo1 \neq 0$

Tabla 1: Analisis Univariado

Chi Cuadrado

H0: Se trata de variables independientes.

H1: No se trata de variables independientes

Test Z

$H0 : \mu Grupo0 - \mu Grupo1 = 0$

$H1 : \mu Grupo0 - \mu Grupo1 \neq 0$

Wilcoxon

$H0 : \theta Grupo0 - \theta Grupo1 = 0$

$H1 : \theta Grupo0 - \theta Grupo1 \neq 0$

Variable: Sex

TEST CHI CUADRADO

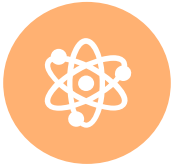
`>> chisq.test(tf, correct = TRUE)` donde tf es la tabla de frecuencias

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tf
## X-squared = 5.8681, df = 1, p-value = 0.01542
```

Menor al nivel de significancia (5%)

Sex y el cancer de pancreas no son variables independientes

Tabla 1: Analisis Univariado



Test Z

$H0 : \mu Grupo0 - \mu Grupo1 = 0$

$H1 : \mu Grupo0 - \mu Grupo1 \neq 0$

Chi Cuadrado

H0: Se trata de variables independientes.

H1: No se trata de variables independientes

Wilcoxon

$H0 : \theta Grupo0 - \theta Grupo1 = 0$

$H1 : \theta Grupo0 - \theta Grupo1 \neq 0$

Variable: Age

TEST DE LEVENE Y TEST Z

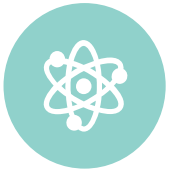
```
>> z.test(edad0,edad1, mu = 0, sigma.x=sigma.x, sigma.y =sigma.y,
conf.level = 0.95)

##  Two-sample z-Test
##
## data:  edad0 and edad1
## z = -3.8511, p-value = 0.0001176
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.375321  -3.701999
## sample estimates:
## mean of x mean of y
##  57.71134  65.25000
```

Menor al nivel de significancia (5%)

Grupo 1	65.25000
Grupo 0	57.71134

La diferencia entre las medias de edad del grupo 0 y del grupo 1 es distinta de cero

Tabla 1: Analisis Univariado



Wilcoxon

$H0 : \theta Grupo0 - \theta Grupo1 = 0$

$H1 : \theta Grupo0 - \theta Grupo1 \neq 0$

Chi Cuadrado

H0: Se trata de variables independientes.

H1: No se trata de variables independientes

Test Z

$H0 : \mu Grupo0 - \mu Grupo1 = 0$

$H1 : \mu Grupo0 - \mu Grupo1 \neq 0$

Variables: Creatinine, LYVE1, REG1B, TFF1

```
>> wilcox.test(creatinina0, creatinina1,
                    paired = F, conf.int =0.95)

##  Wilcoxon rank sum test with continuity correction
##
## data:  creatinina0 and creatinina1
## W = 1285.5, p-value = 0.6698
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.2375452  0.1470490
## sample estimates:
## difference in location
##             -0.03394606
```

TEST DE WILCOXON

Mayor al nivel de significancia (5%)

La mediana de la creatinina del grupo 0 no difiere de la del grupo 1.

Variables: Creatinine, LYVE1, REG1B, TFF1

```
>> wilcox.test(LYVE10, LYVE11, paired = F,
                conf.int =0.95)

##  Wilcoxon rank sum test with continuity correction
##
## data:  LYVE10 and LYVE11
## W = 285.5, p-value = 1.845e-10
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -31.79223 -31.61921
## sample estimates:
## difference in location
##              -31.78906
```

TEST DE WILCOXON

95 %

(-31.79223; -31.61921) pg/ml.

Mediana -31.78906 pg/ml

Menor al nivel de significancia (5%)

La mediana de la LYVE1 del grupo 0 difiere de la del grupo 1.

Tabla resumen: Analisis univariado

Variable	Test	Hipotesis	p-valor	Inferencia
Sex	Chi Cuadrado	H_0: Se trata de variables independientes	0.01542	Se rechaza H_0
Age	Test Z	$H_0 : \mu_{Grupo0} - \mu_{Grupo1} = 0$	0.0001176	Se rechaza H_0
Creatinina	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.6698	Se acepta H_0
LYVE1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	1.845e-10	Se rechaza H_0
REG1B	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.07389	Se acepta H_0
TFF1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	5.53e-05	Se rechaza H_0

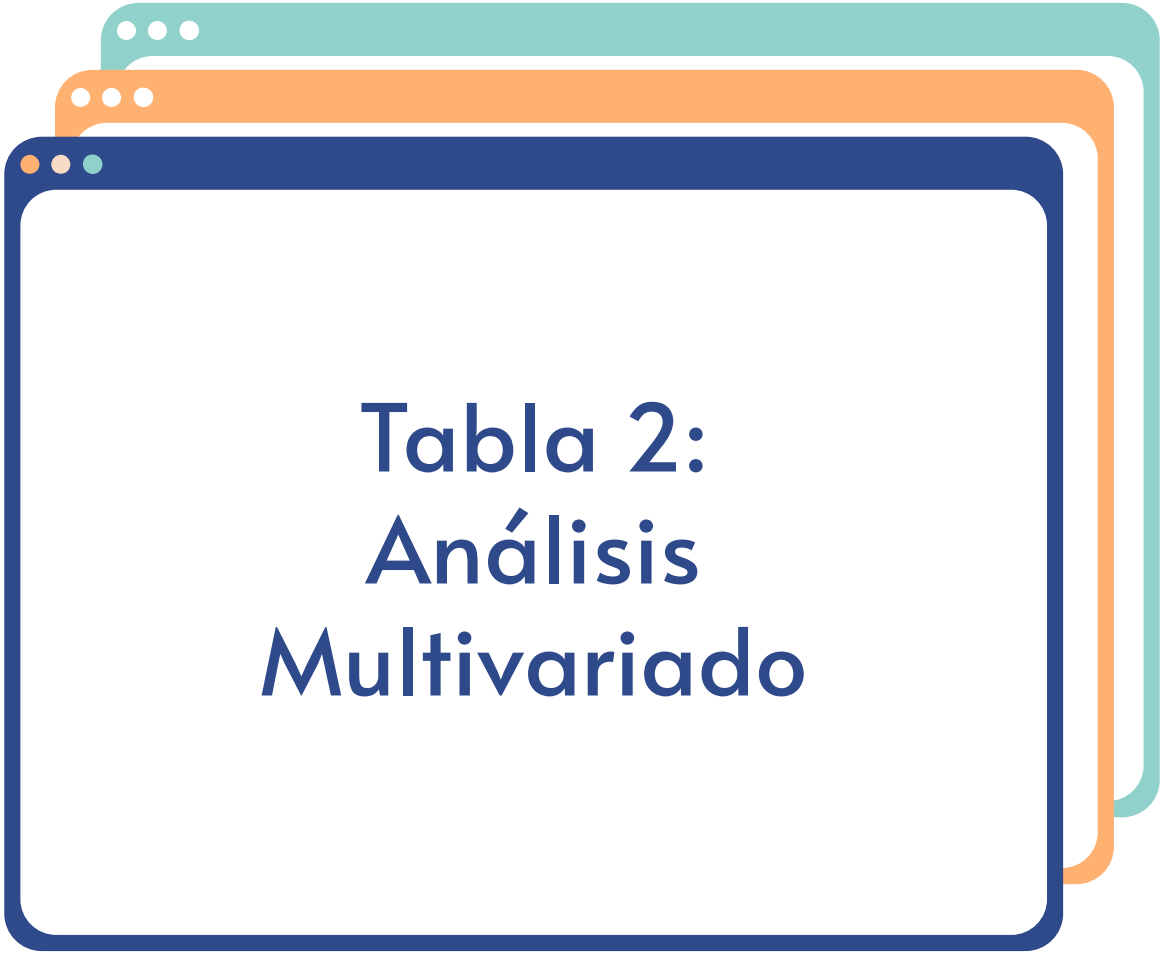


Tabla 2:
Análisis
Multivariado

Tabla resumen: Analisis univariado

Variable	Test	Hipotesis	p-valor	Inferencia
Sex	Chi Cuadrado	H_0: Se trata de variables independientes	0.01542	Se rechaza H_0
Age	Test Z	$H_0 : \mu_{Grupo0} - \mu_{Grupo1} = 0$	0.0001176	Se rechaza H_0
Creatinina	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.6698	Se acepta H_0
LYVE1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	1.845e-10	Se rechaza H_0
REG1B	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.07389	Se acepta H_0
TFF1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	5.53e-05	Se rechaza H_0

Variables dicotómicas

- Se tomaron las variables explicativas relevantes.
- Se las volvió variables dicotómicas utilizando un valor umbral (media)

LYVE1D	
0	93
1	28

TFF1D	
0	117
1	8

Test de Chi cuadrado de Pearson

- 80% de las celdas deben tener una frecuencia esperada mayor o igual a 5 y ninguna frecuencia menor a 1

	Cáncer de pancreas		
TFF1D	0	1	Row Total
0	79 72.168	18 24.832	97
1	14 20.832	14 7.168	28
Column Total	93	32	125

- P valor = 0.001854415

	Cáncer de pancreas		
LYVE1D	0	1	Row Total
0	95 90.792	2 6.208	97
1	22 26.208	6 1.792	28
Column Total	117	8	125

- P valor = 0.001153486

Odds Ratio

TFF1

```
Point estimates and 95% CIs:
-------------------------------------------------------------------------
Inc risk ratio                                1.25 (1.03, 1.52)
Odds ratio                                    12.95 (2.45, 68.55)
Attrib risk *                                 19.37 (3.91, 34.83)
Attrib risk in population *                   15.03 (-0.76, 30.82)
Attrib fraction in exposed (%)                19.77 (2.44, 34.03)
Attrib fraction in population (%)             16.06 (1.54, 28.43)
-------------------------------------------------------------------------
 Test that OR = 1: chi2(1) = 13.604 Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Un valor de TFF1 alto incrementa en 12.95 las chances de tener cancer de pancreas IC(2.45,68.55)

LYVE1

```
Point estimates and 95% CIs:
-------------------------------------------------------------------------
Inc risk ratio                                1.63 (1.11, 2.39)
Odds ratio                                    4.39 (1.78, 10.80)
Attrib risk *                                 31.44 (11.37, 51.51)
Attrib risk in population *                   24.40 (4.36, 44.44)
Attrib fraction in exposed (%)                38.61 (10.01, 58.12)
Attrib fraction in population (%)             32.80 (6.87, 51.51)
-------------------------------------------------------------------------
 Test that OR = 1: chi2(1) = 11.279 Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Un valor de LYVE1 alto incrementa en 4.39 las chances de tener cancer de pancreas IC(1.78,10.80)

```
Call:
glm(formula = diagnosis ~ TFF1D, family = "binomial", data = datafinal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0727  -0.5712  -0.5712  -0.5712   1.9460

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7304     0.2900  -5.967 2.41e-09 ***
TFF1D1        1.4791     0.4594   3.219  0.00128 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.98  on 124  degrees of freedom
Residual deviance: 122.66  on 123  degrees of freedom
AIC: 126.66

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = diagnosis ~ LYVE1D, family = "binomial", data = datafinal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -0.6454  -0.6454  -0.6454   1.8282

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4628     0.2366  -6.183  6.3e-10 ***
LYVE1D1       2.5614     0.8501   3.013  0.00259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.98  on 124  degrees of freedom
Residual deviance: 122.10  on 123  degrees of freedom
AIC: 126.1

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = diagnosis ~ TFF1D + LYVE1D, family = "binomial",
    data = datafinal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0290  -0.5258  -0.5258  -0.5258   2.0235

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9090     0.3117  -6.125 9.07e-10 ***
TFF1D1        1.3998     0.4842   2.891  0.00384 **
LYVE1D1       2.4311     0.8866   2.742  0.00611 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.98  on 124  degrees of freedom
Residual deviance: 113.86  on 122  degrees of freedom
AIC: 119.86

Number of Fisher Scoring iterations: 4
```

CURVAS ROC

- Morado - LYVE1
- Verde - TFF1
- Amarillo - LYVE1 + TFF1

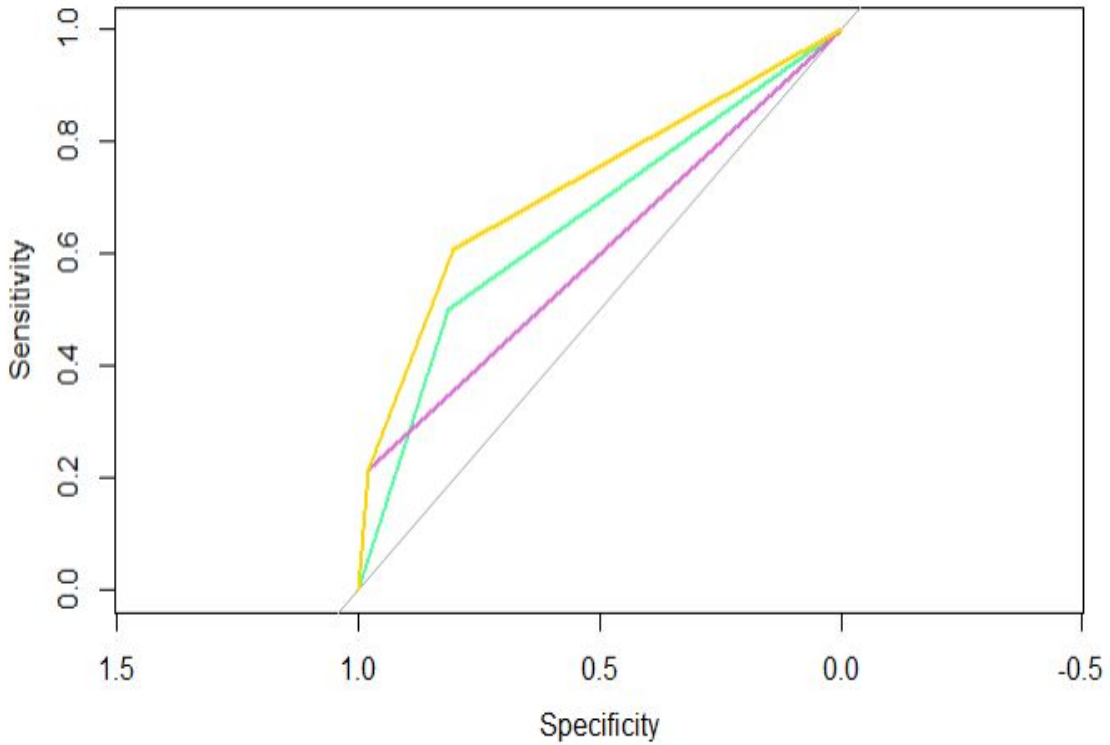


Tabla resumen: Analisis Multivariado

	Variables Utilizadas	AUC
Modelo 1	LYVE1	0.5968
Modelo 2	TFF1	0.6572
Modelo 3	LYVE1 + TFF1	0.7204

05.

Conclusiones

Conclusiones

Las variables explicativas LYVE1 y TFF1 pueden ser utilizadas como biomarcadores para la detección temprana del cáncer de páncreas.

El modelo de regresión logística que provee mejores resultados es el que usa las dos variables (LYVE1 y TFF1).

En futuros trabajos utilizar otros valores umbral de modo de analizar si se consiguen mejores resultados.


Muchas
Gracias!