PEC 1 - Análisis de datos ómicos

Malena Díaz Río

2024-11-03

Table of contents

1 Resumen ejecutivo **1**

2 Objetivos del estudio **2**

3 Materiales y métodos **2**
3.1 Preprocesado, control de calidad y normalización 3
3.2 Exploración de los datos 3
3.3 Análisis estadístico 4

4 Resultados **4**
4.1 Preprocesado, control de calidad y normalización 4
4.2 Exploración de los datos 7
4.3 Análisis estadístico 11

5 Conclusiones **14**

6 Discusión y limitaciones **15**

7 Bibliografía **15**

1 Resumen ejecutivo

En este estudio se realiza una exploración de datos sobre el set de datos público "human cachexia" con el fin de determinar qué metabolitos y qué valores de estos metabolitos podrían ser unos biomarcadores para poder diagnosticar de una forma preventiva la caquexia la cual padecen más del 80% de las víctimas con cáncer. Se ha realizado una limpieza y normalización de los datos, se han eliminado variables altamente correlacionadas y se han determiando las

variables que más impacto pueden tener para resolver este reto mediante un análisis de correlaciones, pruebas estadísticas de comparación de grupos así como a través del análisis de los coeficientes de una regresión logística. Así, se ha obtenido que las variables que más utilidad presentan para diferenciar pacientes con caquexia son la Glucosa, la N,N-Dimethylglycine, la 3-Hydroxybutyrate, el Succinato y el Acetato donde valores altos indican que el paciente padece de esta enfermedad. Otras variables que podrían resultar de interés para un clínico son Adipato, el Quinolinato, la Leucina y la Valina. La principal limitación de este estudio es la falta de registros de control y el desbalanceo de pacientes sanos frente al número de pacientes con caquexia. Así, no se puede afirmar que los resultados obtenidos sean lo suficientemente robustos como para integrarlos en la práctica clínica.

Todos el código utilizado para realizar este estudio está disponible en un repositorio Github que se encuentra en el siguiente enlace: https://github.com/mdiazrioUOC/Diaz-Rio-Malena-PEC1.git.

2 Objetivos del estudio

La caquexia se define como la pérdida de músculo esquelético y grasa como efecto secundario del cáncer y se traduce en síntomas como debilidad o anorexia. Hoy en día es la causa principal del 30% de las víctimas con cáncer [1]. El motivo inicial por el que surge la caquexia no está claramente determinado siendo este un resultado de múltiples factores [1]. Así, el diagnóstico temprano de esta enfermedad a partir de biomarcadores de metabolitos podría ser una gran aportación a la comunidad científica [2]. De ahí nace la finalidad de este estudio que consiste en la comparación en la distribución de metabolitos entre dos grupos de pacientes: un grupo que padece de caquexia y un grupo de control. El fin es identificar aquellas variables con más diferencias significativas entre ambos grupos así como los valores de estas variables que empujarían al clínico a diagnosticar una caquexia.

3 Materiales y métodos

El conjunto de datos utilizados para realizar este estudio es abierto y se puede obtener a través del siguiente enlace: https://rest.xialab.ca/api/download/metaboanalyst/human_cachexia.csv. Este dataset consiste en los datos de 77 pacientes de los cuales se han recogido 63 variables diferentes. De los 77 pacientes 30 pertenecen al grupo de control y 47 al grupo que padece de caquexia. Las 63 variables son de tipo numérica (doubles) e indican el valor de diferentes metabolitos que pueden ser aminoácidos, carbohidratos, lípidos...etc en un momento determinado. Este conjunto de datos no tiene ningún valor faltante.

Para llevar a cabo el análisis de los datos se ha utilizado el lenguaje de programación `R`. Además, se han explotado las librerías proporcionadas por el gestor de paquetes de Bioconductor que ayudan al usuario de datos ómicos a manipular los diferentes conjuntos de una manera eficaz.

Todo el código utilizado para la exploración de datos se encuentra dispuesto en un archivo de R markdown (.Rmd) con el fin de facilitar la visualización de los resultados.

El estudio de las variables diferenciadoras entre pacientes con caquexia y sin ella se ha llevado a cabo en tres pasos siguiendo el flujo natural del proceso de datos ómicos. Una vez planteada la pregunta biológica y obtenidos los datos crudos se procedió a realizar un control de calidad, preprocesado y normalización de los datos. Más tarde se realizó una exploración de los datos estudiando tanto sus distribuciones por separado como las interacciones entre las diferentes variables del conjunto de datos. Finalmente, se procedió a realizar un análisis estadístico que pretende realizar la comparación en las variables de ambos grupos.

3.1 Preprocesado, control de calidad y normalización

El primer paso al recibir los datos crudos fue la adaptación de estos a un formato fácilmente manejable. De esta forma, los datos se vuelcan en una estructura de tipo *Summarized Experiment* que permite condensar toda la información relativa al estudio en un único objeto. Así, se llevaron a cabo una serie de transformaciones en el set de datos original. Por una parte, se separó el dataset en dos conjuntos de datos, uno conteniendo la variable dependiente, es decir, el grupo de cada paciente en forma de factor y otro conjunto con todas las variables independientes. Este segundo conjunto de datos, para cumplir con el formato impuesto por el *Summarized Experiment* se traspuso obteniendo un dataset con una columna por paciente y una fila por variable. Además, se definió una serie de metadatos sobre el experimento incluyendo el autor y la fuente de los datos con el fin de garantizar la integridad de estos.

Una vez creada la estructura necesaria, se procedió a realizar el control de calidad. Este proceso trata de resolver tres preguntas claves: estudio de valores faltantes, estudio de valores anómalos e identificación de correlaciones altas. Para la identificación de valores anómalos se utilizó el criterio del rango IQR que consiste en identificar como valor anómalo todo aquel que quede fuera del rango $[\hat{x} - 1,5IQR, \hat{x} + 1,5IQR]$, siendo $\hat{x}$ el estimador de la media de cada variable en el conjunto de datos. Aquellos valores que salían del rango permitido se imputaron al mínimo o máximo valor que entrara dentro del rango.

Por último, se procedió a normalizar los datos. Este paso permite evitar que variables que tengan una escala mayor empujen al análisis estadístico a concluir que las diferencias entre ambos grupos de pacientes para esas variables es mayor frente a otras con una escala menor. Además, la mayoría de los análisis estadísticos requieren que las variables presenten una distribución normal. Por lo tanto, se calculó el logaritmo natural de todas las variables y este dataset se añadió al objeto *arrays* del objeto *SummarizedExperiment*.

3.2 Exploración de los datos

Para entender el comportamiento del conjunto de variables se realizaron comparaciones de media y desviación estándar de los dos grupos. Para ello y asumiendo condiciones de normal-

idad, se realizó una prueba T-Test. Junto con el p-valor de esta prueba, se estudió el valor del FDR que pretende estimar la corrección de este valor frente a un conjunto de datos con un alto número de variables independientes. Además, se representaron las distribuciones de las variables con menor p-valor, esto es, con más diferencias significativas entre ambas poblaciones. Por último, se estudió la correlación de cada variable con la variable objetivo mediante el coeficiente de correlación de Pearson. Se realizó también un estudio de las correlaciones entre variables del set de datos con el fin de eliminar variables que tuvieran correlaciones muy altas y que, por lo tanto, estuvieran aportando información duplicada.

3.3 Análisis estadístico

Una vez observadas las variables con más impacto en la variable objetivo así como sus distribuciones se llevó a cabo un análisis estadístico que permitiera extraer conclusiones robustas. Para ello se realizaron los siguientes pasos. Por una parte, se seleccionaron las 20 variables con menor p-valor y se ajustó una regresión logística para determinar si había diferencias significativas entre los pacientes con caquexia y los pacientes del grupo control. Además, se crearon una serie de modelos de regresión logística con diferentes conjuntos de variables para obtener aquella combinación que mejor rendimiento demostrase. De este modo, se utiliza el método de selección hacia delante. Mediante este análisis se obtuvieron las variables más significativas para el diagnóstico de caquexia. Por último, se realizó un estudio de los coeficientes de regresión mediante los cuales se observa qué valores de cada variable indican la presencia de esta enfermedad.

4 Resultados

4.1 Preprocesado, control de calidad y normalización

El dataset human_cachexia presentó un 0% de nulos por lo que no hizo falta afrontar valores faltantes. Para el cálculo de los valores anómalos, se estudió la distribución de las variables mediante la representación de histogramas. La distribución para todas las variables resultó no cumplir con las condiciones de normalidad y seguir más bien una distribución exponencial como se puede apreciar en la figura 1. Así, se normalizó la distribución calculando el logaritmo natural de cada una de estas variables. La distribución original y la distribución después de calcular el logaritmo se pueden visualizar en las siguientes figuras:

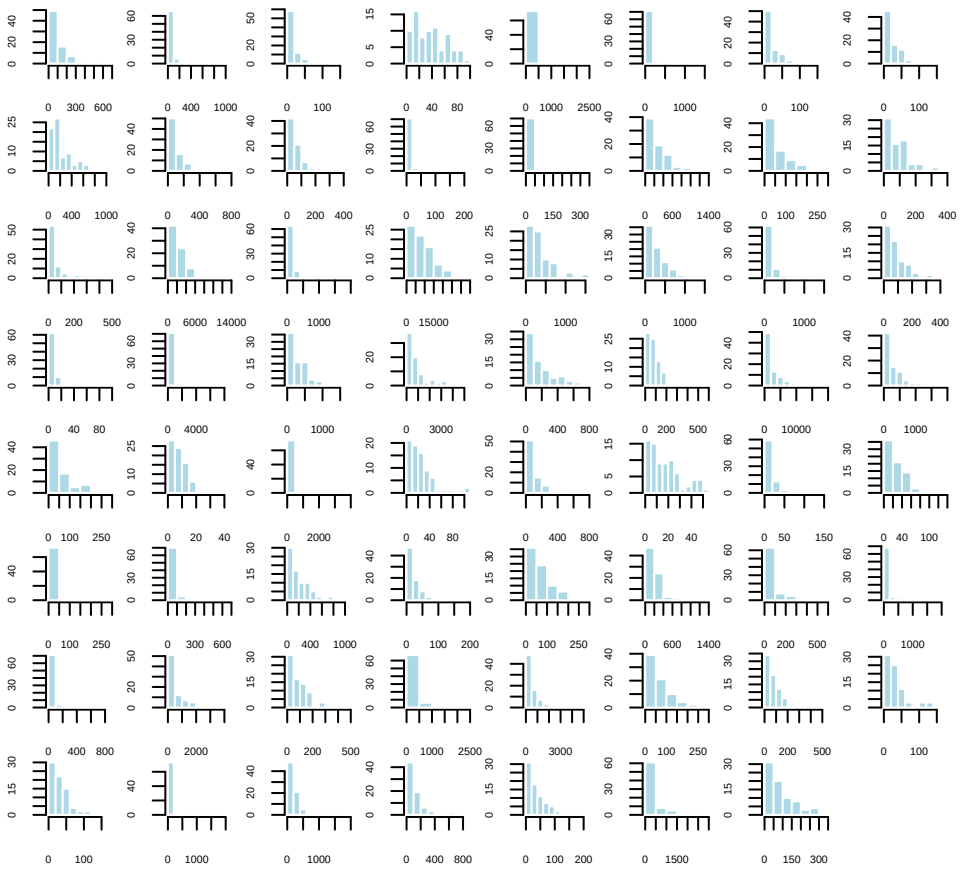


Figure 1: Distribución de las variables del set de datos human cachexia

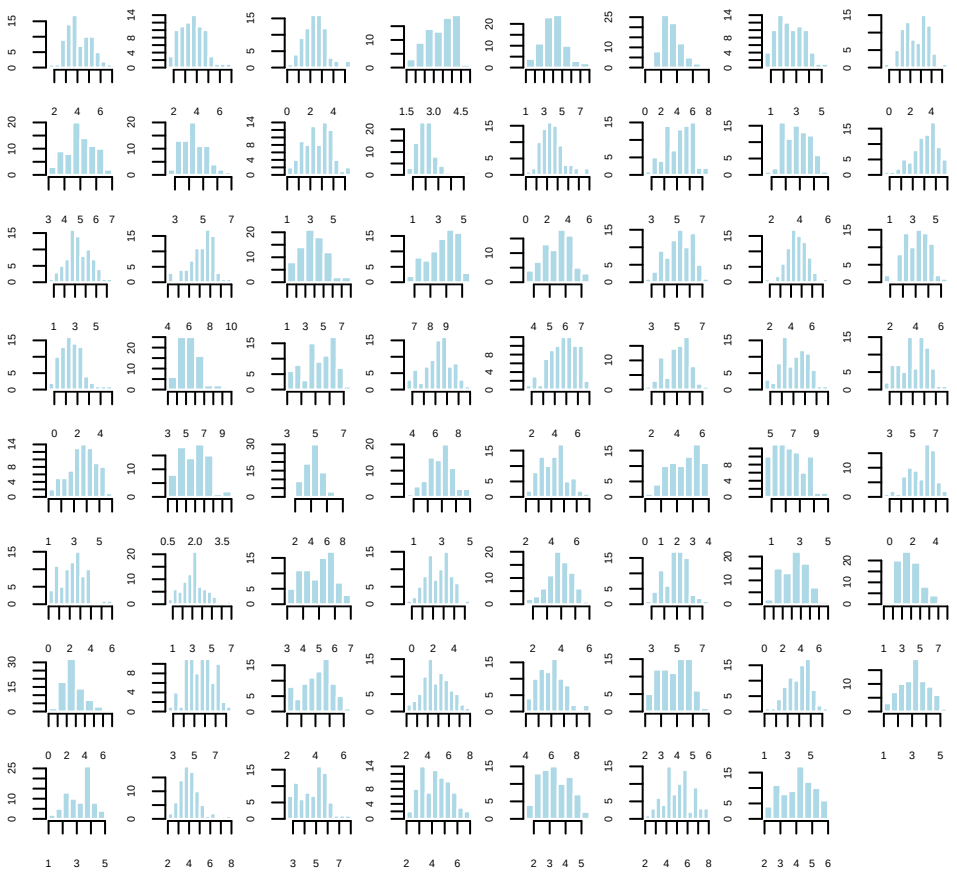


Figure 2: Distribución de los logaritmos de las variables del set de datos human cachexia

Una vez calculado el logaritmo y calculados los intervalos de valores válidos a partir del IQR, se obtuvieron 32 valores anómalos los cuales pertenecían a las variables 1-Methylnicotinamide, 2-Aminobutyrate, 2-Oxoglutarate, 3-Aminoisobutyrate, Acetone, Adipate, Betaine, Citrate, Formate, Fumarate, Glucose, Lactate, Pantothenate, Sucrose, Tartrate y Xylose. Las variables Adipate y 2-Oxoglutarate presentaron 4 valores anómalos, esto es, un 5%. Con el fin de confirmar la validez de los rangos calculados en la figura 3 se puede observar los límites de los rangos frente a la distribución original de aquellas variables que presentaron valores anómalos.

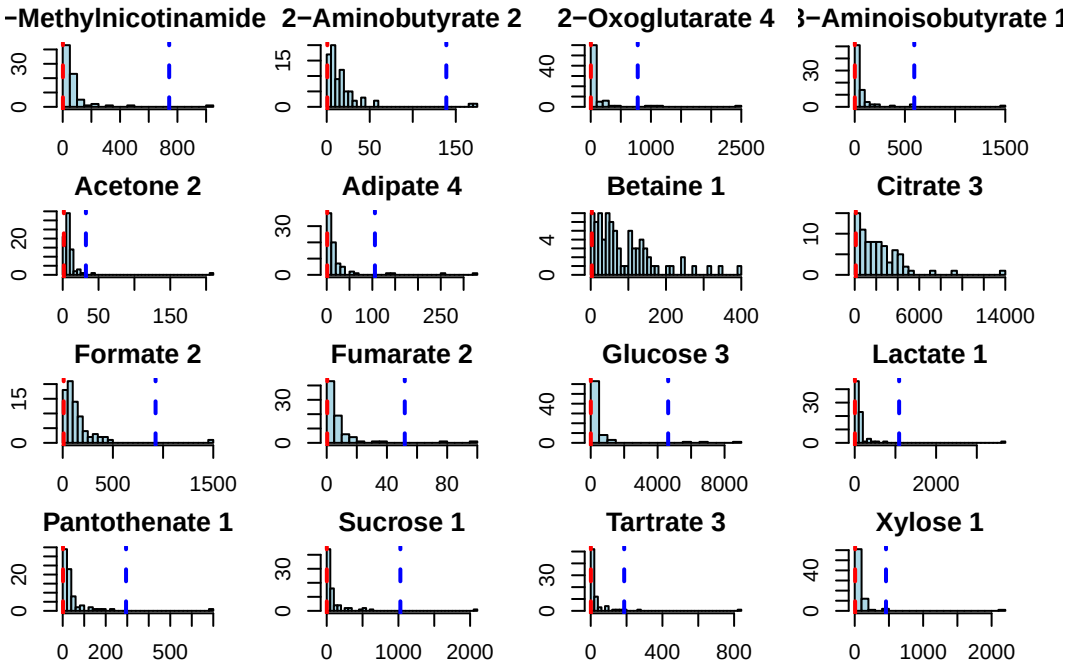


Figure 3: Distribución de las variables con valores anómalos y rangos

A través de estos histagramas se acepta la validez de los rangos ya que los valores que caen fuera de ellos se encuentran en posiciones extremas. Como se ha indicado en la sección de materiales y métodos estos valores anómalos fueron modificados de tal forma que cogieran el valor más cercano que estuviera dentro de los rangos IQR.

4.2 Exploración de los datos

En un primer lugar, se procedió a la comparación de cada variable individualmente, realizando un t-test frente a las variables normalizadas. Mediante el p-valor observamos las variables que estadísticamente muestran más diferencias entre ambas poblaciones. A continuación se presenta la tabla que muestra las 15 variables con menor p-valor. En esta tabla también se observa la media y desviación estándar de cada grupo lo que nos da idea de qué valores podrían indicar la presencia de caquexia.

Table 1: 15 variables más significativas ante un t-test para la comparación de dos grupos

	Grupo 1 media +- sd	Grupo 2 media +- sd	p-valor	FDR
Glucose	665.84 ± 1096.38	140.96 ± 99.2	0.00000	0.00000
Adipate	25.46 ± 28.2	8.99 ± 10.43	0.00001	0.00021
Quinolinate	83.75 ± 53.41	39.32 ± 33.76	0.00001	0.00021

	Grupo 1 media +- sd	Grupo 2 media +- sd	p-valor	FDR
Leucine	31.26 ± 24.23	13.56 ± 9.16	0.00002	0.00032
Valine	45.58 ± 32.49	20.13 ± 15.1	0.00003	0.00032
myo-Inositol	181.84 ± 197.77	62.64 ± 70.45	0.00003	0.00032
3-Hydroxyisovalerate	27.61 ± 27.44	12.31 ± 17	0.00004	0.00032
N,N-Dimethylglycine	34.49 ± 26.59	13.6 ± 13.44	0.00004	0.00032
3-Hydroxybutyrate	29.26 ± 30.73	9.9 ± 7.99	0.00006	0.00042
Succinate	79.63 ± 98.58	29.84 ± 44.99	0.00007	0.00044
Creatine	174.91 ± 335.36	51.5 ± 87.24	0.00009	0.00052
Glutamine	391.41 ± 311.57	174.43 ± 195.36	0.00011	0.00054
Acetate	85.63 ± 90.6	35.6 ± 42.93	0.00012	0.00054
cis-Aconitate	276.03 ± 331.29	91.72 ± 84.9	0.00012	0.00054
Pyroglutamate	270.29 ± 212.19	119.26 ± 98.63	0.00014	0.00059
Alanine	347.59 ± 281.73	157.58 ± 156.14	0.00017	0.00067
Dimethylamine	453.58 ± 347.08	208.68 ± 139.19	0.00021	0.00073
Tryptophan	81.82 ± 59.08	41.83 ± 42.08	0.00021	0.00073
Methylamine	21.22 ± 13.78	11.36 ± 12.03	0.00022	0.00073
Betaine	112.25 ± 83.2	95.69 ± 218.79	0.00025	0.00079

De las 63 variables disponibles en el set de datos 54 variables obtienen un p-valor de 0.05 en esta prueba, esto es un 85%. Este es un número bastante elevado de variables lo que puede indicar dos cosas. Por una parte, que el número de registros (de cada grupo 47 y 33) no sea suficientemente representativo y que, por lo tanto, casos extremos de uno de los grupos empujen a la prueba estadística a pensar que se tratan de dos poblaciones diferentes. Sin duda, los resultados de esta prueba serían más fiables si los datos perteneciesen a poblaciones más grandes. Por otra parte, este resultado puede ser verídico, es decir, que sí existan esas diferencias en la distribución en ambas variables. Para reforzar la hipótesis de que estas variables sí que presentan diferencias significativas procedemos al análisis de correlaciones.

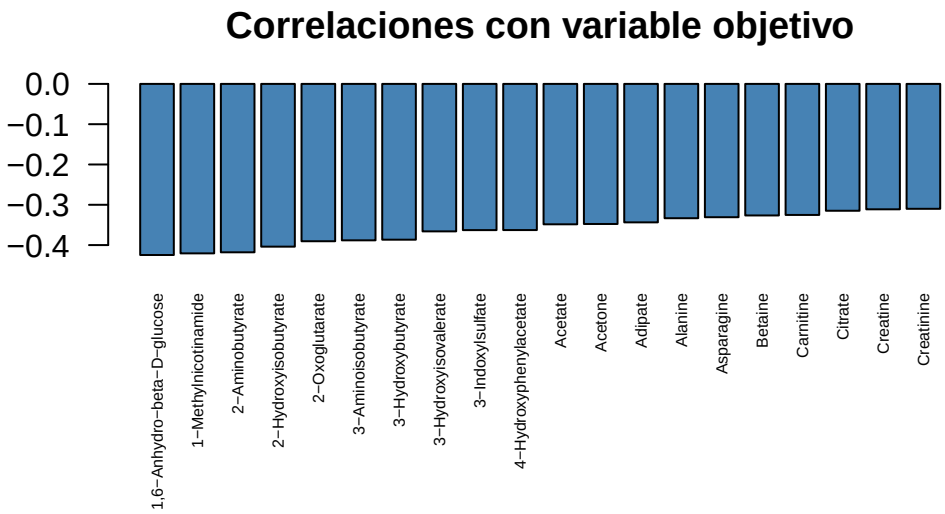


Figure 4: Índice de correlación de Pearson para las 20 variables con más correlación con la variable 'grupo'

Si comparamos estos resultados con los obtenidos anteriormente, observamos que la Glucosa, que en el análisis de comparación de medias parecía presentar diferencias significativas entre el grupo de control y el grupo que padece la enfermedad, no sale en la lista de variables con más correlación. Esto ocurre también para unas cuantas variables como la Acetona, o la Alanina. La variable 1,6-Anhydro-beta–D-glucose sería el caso contrario: alta correlación pero menor p-valor. Desde una perspectiva global, este análisis nos permite identificar variables que con alta probabilidad tienen una distribución diferente entre ambos grupos. Estas variables son aquellas que aparecen en la Tabla 1 y que además tienen una correlación elevada con la variable objetivo. Como se puede observar, un ejemplo de estas variables es la 3-Hydroxyisovalerate, el acetato o la creatina.

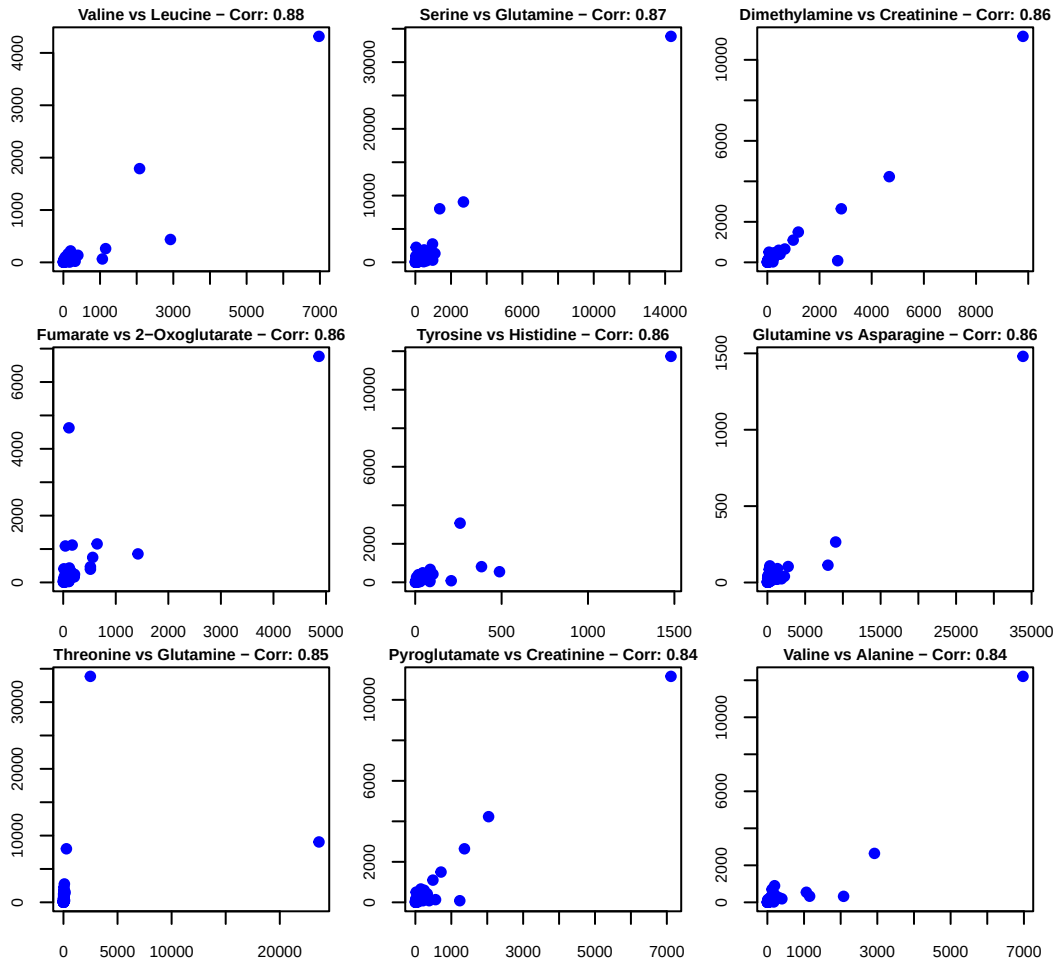
Antes de proceder con el análisis estadístico, realizamos un estudio de correlaciones entre todo el conjunto de variables independientes para evitar que dos variables altamente correlacionadas puedan afectar a los resultados del análisis estadístico. Después de calcular la correlación de Pearson para cada par de variables, mostramos en la Tabla 2 todas aquellas variables que presentan más de 0.8 de correlación y que por lo tanto habrá que filtrar.

```
kable(cor_long[cor_long$Correlation > 0.8, ],format = "html", caption = "Variables con una co
```

Table 2: Variables con una correlación de Pearson mayor a 0.8

	Variable1	Variable2	Correlation
928	Serine	Asparagine	0.8021854
1666	Glycine	Glutamine	0.8027463
1221	Fucose	Creatinine	0.8048195
424	Serine	3-Hydroxybutyrate	0.8051757
1356	Hypoxanthine	Ethanolamine	0.8088955
1303	Pyroglutamate	Dimethylamine	0.8167289
1219	Ethanolamine	Creatinine	0.8190004
933	Threonine	Asparagine	0.8315868
1752	Threonine	Glycine	0.8352743
876	Valine	Alanine	0.8366967
1240	Pyroglutamate	Creatinine	0.8443799
1689	Threonine	Glutamine	0.8477669
909	Glutamine	Asparagine	0.8556964
2008	Tyrosine	Histidine	0.8572686
277	Fumarate	2-Oxoglutarate	0.8602892
1218	Dimethylamine	Creatinine	0.8623471
1684	Serine	Glutamine	0.8674403
2262	Valine	Leucine	0.8804148

Tras comparar cada par de variables de la Tabla 2 y seleccionar aquella con menos p-valor en la Tabla 1 completa, las siguientes variables fueron eliminadas del experimento: Valine, Serine, Creatinine, 2-Oxoglutarate, Histidine, Asparagine, Threonine, Alanine, Glycine, Ethanolamine Dimethylamine, Hypoxanthine, Fucose. En la siguiente figura se observa un diagrama de puntos junto con el valor de correlación de los 9 pares de variables que mayor valor presentaban.



4.3 Análisis estadístico

Finalmente, para comprobar qué subconjunto de variables podrían ayudar más a un médico a decantarse por diagnosticar la enfermedad de caquexia, se ha realizado una regresión logística con las 20 variables que menos p-valor han presentado en la exploración de datos. El resumen del modelo lineal se presenta a continuación.

```
summary(linear_model)
```

```
Call:
glm(formula = as.numeric(se$group) - 1 ~ top_20_pvalue, family = binomial)

Coefficients:
```

```
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       3.277067   1.066774   3.072  0.00213 **
top_20_pvalueGlucose             -0.019432   0.009323  -2.084  0.03713 *
top_20_pvalueAdipate              0.009123   0.057112   0.160  0.87309
top_20_pvalueQuinolinate         -0.013501   0.017422  -0.775  0.43838
top_20_pvalueLeucine             -0.073837   0.076723  -0.962  0.33586
top_20_pvaluemyo-Inositol        -0.005248   0.004076  -1.288  0.19785
top_20_pvalue3-Hydroxyisovalerate -0.011047   0.045536  -0.243  0.80832
top_20_pvalueN,N-Dimethylglycine -0.120911   0.053328  -2.267  0.02337 *
top_20_pvalue3-Hydroxybutyrate   -0.163760   0.079959  -2.048  0.04056 *
top_20_pvalueSuccinate            0.061447   0.021771   2.822  0.00477 **
top_20_pvalueCreatine            -0.006741   0.003086  -2.184  0.02894 *
top_20_pvalueGlutamine            0.012290   0.006897   1.782  0.07475 .
top_20_pvalueAcetate             -0.033844   0.017658  -1.917  0.05529 .
top_20_pvaluecis-Aconitate        0.002121   0.005981   0.355  0.72289
top_20_pvaluePyroglutamate        0.014229   0.007589   1.875  0.06079 .
top_20_pvalueTryptophan          -0.039738   0.025632  -1.550  0.12107
top_20_pvalueMethylamine          0.133378   0.080968   1.647  0.09950 .
top_20_pvalueBetaine              0.023410   0.012720   1.840  0.06572 .
top_20_pvalueFormate             -0.008537   0.008780  -0.972  0.33091
top_20_pvalueSucrose              0.006523   0.006701   0.973  0.33033
top_20_pvalueLactate              0.012723   0.010794   1.179  0.23854
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102.960  on 76  degrees of freedom
Residual deviance:  50.282  on 56  degrees of freedom
AIC: 92.282

Number of Fisher Scoring iterations: 10
```

En el modelo, observamos que diferentes variables presentan un p-valor en sus coeficientes de regresión menor a 0.05. Esto es, hay baja probabilidad de que estas variables no tengan efecto alguno sobre la variable objetivo lo que se traduce en que sí que hay diferencias entre las dos poblaciones. Las variables que cumplen con esta característica son: la Glucosa, la N,N-Dimethylglycine, la 3-Hydroxybutyrate, el Succinato y el Acetato.

Como se puede observar a partir de los coeficientes de regresión, todas estas variables menos el Succinato tienen un coeficiente de regresión negativo por lo que a menor valor, será más probable que el sujeto pertenezca a la clase control. Observamos la diferencia de distribuciones en la siguiente figura:

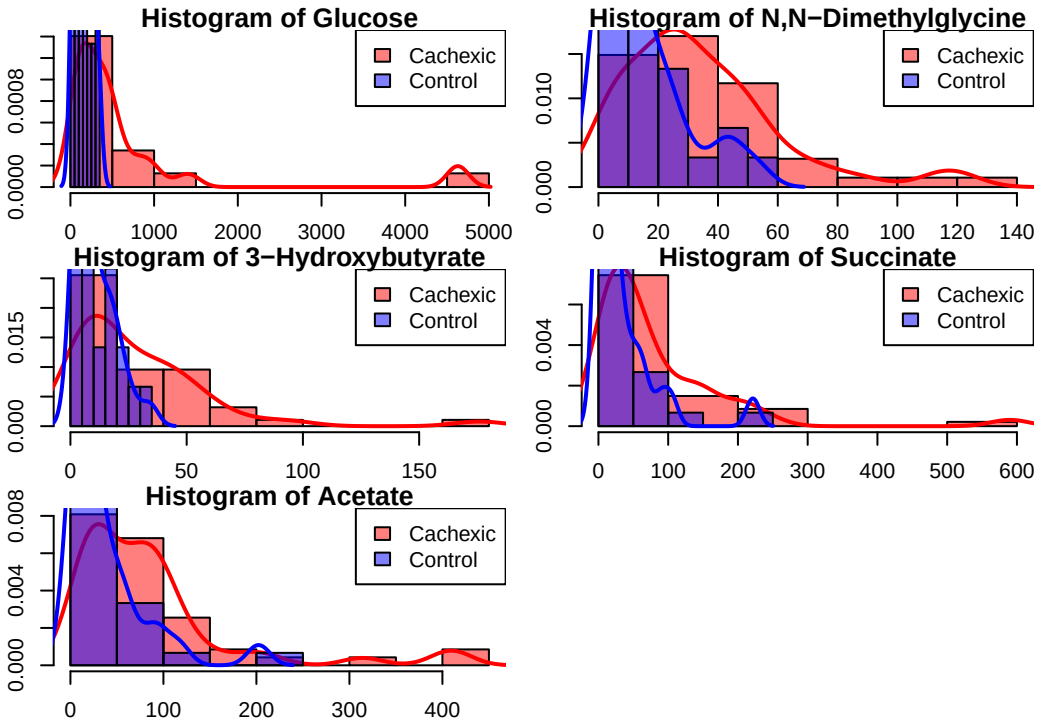


Figure 5: Comparación de distribuciones de las variables más significativas entre grupos

En estos gráficos se observa que los pacientes con caquexia presentan una distribución descentrada con una larga cola derecha. Esta es la principal razón por la cual el modelo de regresión logística asigna los valores negativos a los diferentes coeficientes de las variables más significativas. En todo caso, parece que el grupo de pacientes de caquexia tiene valores más extremos que aquellos del grupo control.

Se utilizó la técnica de la selección hacia delante para crear un subconjunto de variables que minimizara el rendimiento de la regresión logística y se obtuvo el siguiente modelo lineal:

```
# Resumen del modelo seleccionado
summary(step.model)
```

```
Call:
glm(formula = group_data ~ 1, family = binomial, data = as.data.frame(top_20_pvalue))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  -0.4490     0.2337  -1.921   0.0547 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 102.96  on 76  degrees of freedom
Residual deviance: 102.96  on 76  degrees of freedom
AIC: 104.96

Number of Fisher Scoring iterations: 4
```

El modelo seleccionado es precisamente aquel que no coge ninguna variable. Esto indica que la suma de variables en conjunto no es capaz de diferenciar entre pacientes del grupo con caquexia y el grupo de control mejor que si lo hiciéramos al azar (30% probabilidad grupo control y 70% grupo caquexia).

5 Conclusiones

A través de esta exploración de datos extraemos una serie de conclusiones. El objetivo principal de este estudio era determinar si existían metabolitos que pudiesen indicar la presencia de caquexia en pacientes con tal de poder realizar un diágnostico temprano de la enfermedad. Tras haber analizado los datos a través de pruebas estadísticas concluimos que las variables de Glucosa, la N,N-Dimethylglycine, la 3-Hydroxybutyrate, el Succinato y el Acetato son los biomarcadores que más importancia tendrían para tomar esta decisión. Valores más pequeños de estas variables en general indican que el paciente no padece la enfermedad mientras que valores extremos suelen indicar un estado de salud menos favorable. Otras variables que podrían ser interesantes son el Adipato, el Quinolinato, la Leucina y la Valina. Todas estas variables han demostrado tener distribuciones diferentes al enfrentarlas a una comparación de grupos mediante la prueba T. La Valina también puede ser un biomarcador interesante pero en este estudio se ha determinado que no es un metabolito tan relevante como los otros ya que presenta una correlación alta con la Leucina que presenta un menor p-valor en la comparación de distribuciones por grupo.

Por otra parte, el análisis del método de selección hacia delante en el estudio de la regresión logística también nos indica cómo se comportan las variables en conjunto. Observamos que el modelo no encuentra un subconjunto de variables que maximicen el rendimiento de una forma significativa. Esto puede suponer que muchas veces estos valores extremos que empujan a pensar que el paciente padece caquexia no se dan en todas las variables, si no que es realmente solo una variable la que empujaría al modelo a tomar la decisión.

6 Discusión y limitaciones

Como se ha observado en la última parte del análisis estadístico, observamos que las principales diferencias que el modelo encuentra entre ambos grupos es la presencia de valores extremos en los pacientes de caquexia. Por una parte, esto podría ajustarse a la realidad lo que implicaría que pacientes que padecen la enfermedad tienen en efecto valores más elevados. Sin embargo, también podría ser una consecuencia del alto número de pacientes con caquexia de los que disponemos frente al número de pacientes del grupo control. Cuanto menor es la muestra de la que se dispone, la distribución tenderá a tener menos valores extremos lo que podría ser la razón por la cual el grupo control presenta una distribución menos dispersa. Así, este desbalanceo perjudica enormemente la fiabilidad de los resultados obtenidos en este estudio. Para poder concluir con cierta robustex qué variables presentan diferencias entre ambos grupos, se necesitaría un mayor número de registros en general y en especial de pacientes del grupo control.

Por otra parte, en este estudio sólamente se ha recurrido a técnicas basadas en estadística tradicional para determinar las variables con más impacto en los pacientes con caquexia. Sin embargo, hoy en día las técnicas más punteras están basadas en Machine Learning y en la literatura presentan mejores resultados que las técnicas utilizadas en este estudio. Por ello modelos basados en estructura de árbol como Random Forest o XGBoost podrían proporcionar a este problema un análisis más preciso.

7 Bibliografía

[1] Instituto Nacional del Cáncer. (n.d.). Avances en el tratamiento de la caquexia Por Cáncer. Avances en el tratamiento de la caquexia por cáncer. https://www.cancer.gov/espanol/cancer/tratamiento/investigacion/caquexia

[2] Costa, G. (1977). Cachexia, the Metabolic Component of Neoplastic Diseases. *Cancer Research*, *37*, 2327–2335. https://doi.org/10.1159/000385967