Minería de datos sobre los factores de riesgo del cáncer de mama

Guillermo Carretero Palacios
Máster universitario en Bioinformática y bioestadística UOC-UB
M0.190-TFM-Estadística y bioinformática 14

Romina Rebrij
David Merino Arranz

01/2018

A) Creative Commons:

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Minería de datos sobre los factores de riesgo del cáncer de mama
Nombre del autor:	*Guillermo Carretero Palacios*
Nombre del consultor/a:	Romina Rebrij
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	01/2018
Titulación::	Máster universitario en Bioinformática y bioestadística UOC-UB
Área del Trabajo Final:	*M0.190-TFM-Estadística y bioinformática 14*
Idioma del trabajo:	*Castellano*
Palabras clave	*Breast cancer; text mining, risk factor*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En este estudio se han enumerado y clasificado los factores de riesgo que pueden inducir una aparición del cáncer de mama. Estos factores se dividen en dos grandes grupos; factores de riesgo genéticos y factores de riesgo ambientales. Se ha pretendido observar como interaccionan entre e intragrupo. Para llevar a cabo el estudio se utilizó la técnica de minería de datos, concretamente la minería de texto. Esta herramienta nos ha permitido analizar grandes volúmenes de datos de una manera rápida y eficaz. Nuestra base de datos fue PubMed, que nos proporcionó los resúmenes de los artículos médicos necesarios para el estudio. Para analizar estos datos se utilizó PubMed.mineR, un paquete de R dedicado principalmente al *text mining* de PubMed. Después de aplicar esta metodología se obtuvieron diversos resultados. En primer lugar obtuvimos una lista de 4 genes principales, cuyas mutaciones tienen elevadas probabilidades de producir cáncer de mama, los dos principales responsables son BRCA1 y BRCA2. En segundo lugar obtuvimos diferentes factores de riesgo ambientales. Los resultados fueron más dispersos al existir una gran cantidad de posibilidades, pero aparentemente los factores más influyentes fueron la edad y una sobreexposición o producción de estrógeno. Otro componente con gran peso fue el factor hereditario. En conclusión, existen diversos factores con un gran peso en la posible aparición del cáncer de mama, guardando una estrecha relación entre factores genéticos y ambientales.

Abstract (in English, 250 words or less):

In this study we have listed and classified the risk factors that can induce an occurrence of breast cancer. These factors are divided into two large groups, genetic risk factors and environmental risk factors. The aim of this study is to observe the interaction between groups and intragroups. To carry out the study, *data mining technique* was used, specifically *text mining*. This tool allows us to analyze large volumes of data in a fast and efficient way. Our database was PubMed, which provided us the results of medical articles necessary for the study. To analyze this data, PubMed.mineR and R package were used. These programs are specifically for the analysis of PubMed texts. After applying this methodology, several results were obtained. First, we obtained a list of 4 genes, which are mutations that have high probabilities of producing breast cancer, two main genes are *BRCA1* and BRCA2. The results were more dispersed as there were a lot of possibilities, but the most influential factors were age and overexposure or estrogens production. Another component with great weight was the hereditary factor. In conclusion, there are factors with a great weight in the possible appearance of breast cancer, keeping a close relationship between genetic and environmental factors.

Índice

Introducción

1.1 Contexto y justificación del trabajo

Existen muchos factores de riesgo de cáncer de mama. Este tipo de cáncer es el más extendido entre las mujeres. Como la mayoría, este cáncer presenta un componente ambiental así como una predisposición genética a padecerlo. A causa de los problemas ambientales globales, el tipo de vida y el aumento de la población mundial, es posible que en los próximos años aumente el número de personas afectadas. Con este estudio, aplicando técnicas de minería de datos, pretendo identificar y clasificar el mayor número de factores de riesgo para poder evitarlos o minimizar su impacto.

1.2 Objetivos del trabajo

1.2.1. **Objetivos generales:**

- Identificar y clasificar los factores de riesgo del cáncer de mama.

1.2.2. **Objetivos específicos:**

- Determinar los factores genéticos del cáncer de mama.

- Determinar los factores ambientales del cáncer de mama.

- Hacer un ranking de factores de riesgo.

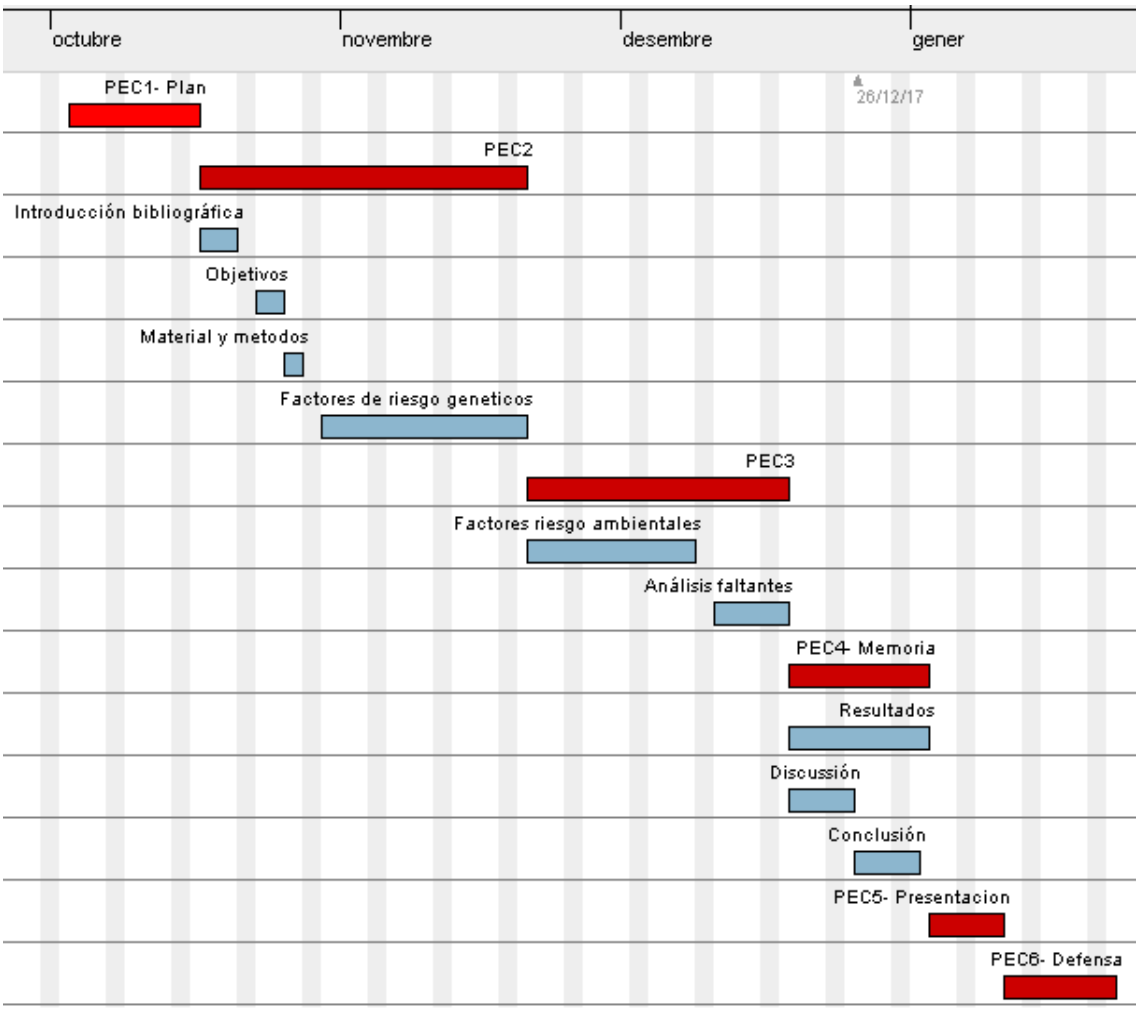
- Estudiar la interacción entre factores genéticos y ambientales.

1.3 Enfoque y método seguido

Para conseguir resultados en este trabajo se enfocara el análisis de los datos utilizando minería de datos combinado con *machine learning*. A diferencia de la lenta y tediosa, puede que incompleta, investigación manual la minería de datos permitirá un análisis rápido y directo después de reunir y combinar diferentes fuentes de información.

1.4 Planificación del trabajo

- Documentación
    - PECs
        - Redacción de la memoria
        - Preparar presentación y defensa
- Introducción bibliográfica
- Redacción de los objetivos
- Redacción de material y métodos
- Identificación de factores de riesgo genéticos
- Identificación de factores de riesgo ambientales
- Análisis de los datos
- Obtención de resultados
- Elaboración de la discusión
- Elaboración de la conclusión



1.5 Breve sumario de productos obtenidos

Lo que se obtiene al realizar este trabajo son los resultados obtenidos al realizar los análisis, que se pueden resumir en un ranking de factores de riesgo genéticos y ambientales, así como la relación entre ellos.

1.6 Breve descripción de los otros capítulos de la memoria

Tal y como se muestra en el índice, este TFM está estructurado de la misma forma que un articulo científico. En primer lugar encontramos la introducción bibliográfica, donde se contextualiza el trabajo y se hace una breve explicación de los conocimientos básicos necesarios. En segundo lugar encontramos el apartado de material y métodos, donde se muestra el procedimiento seguido en el trabajo, así como las instrucciones necesarias para replicar los resultados. En tercer lugar encontramos la exposición cruda de los resultados obtenidos. Seguido de una discusión donde se intenta justificar, entender e interpretar los resultados. Por último encontramos la conclusión, donde se resume la interpretación correcta de los resultados.

2.1 - Introducción

2.1.1 Cáncer

El cáncer es una carga en nuestro genoma, es biológicamente inevitable. Al envejecer las propias células se generan errores que provocan tumores (Weinberg, 2014). El cáncer se caracteriza por una proliferación anormal de células con el potencial de invadir otras partes del cuerpo, este proceso se conoce como metástasis. La inestabilidad genómica es una característica clave en los tumores, es un estado celular anormal asociado a la acumulación de errores en el genoma. Esta inestabilidad puede ser cromosómica o de microsatélites. A pesar de los errores genéticos que puedan desencadenar un tumor, el riesgo de desarrollar un cáncer está influenciado por el estilo de vida, es decir, por causas ambientales (Hoboken & Wiley-Blackwell; 2013). Los carcinógenos son compuestos que pueden inducir cáncer ya que actúan como mutágenos, pueden ser físicos o químicos. Los tumores se originan en células especializadas, ya sean epiteliales (carcinomas) o no epiteliales.

2.1.2 Cáncer de mama

El cáncer de mama es el más común entre las mujeres en todo el mundo, pues representa el 16% de todos los cánceres femeninos[(1).]

La mama está conformada por diferentes tejidos, que van de tejido muy adiposo a tejido muy denso. El cáncer se origina cuando las células saludables de la mama empiezan a cambiar y proliferar sin control, y forman una masa o un conglomerado. Generalmente, el cáncer de mama se origina en las células de los lobulillos, que son las glándulas productoras de leche, o en los conductos, que son las vías que transportan la leche desde los lobulillos hasta el pezón (Margolese *et al*.; 2003)

Hay dos tipos principales de cáncer de mama:

- Carcinoma ductal: Estos cánceres se originan en las células que recubren internamente los conductos de la leche y conforman la mayoría de los cánceres de mama.

- Carcinoma lobular: Este cáncer se origina en los lóbulos

Uno de los tipos de cáncer de mama más agresivos es el denominado triple negativo. Es aquel en que las células cancerígenas muestran resultados negativos en los receptores de estrógeno, progesterona y HER. Alrededor del 15% de los canceres de mama presentan este comportamiento (Sun *et al.;* 2017).

Las pruebas que se utilizan para la detección, diagnóstico y control, entre ellas: mamografías, ecografías, IRM, tomografía axial computarizada, exploración con TEP y más.

2.1.3 Síntomas del cáncer de mama

Al principio, es posible que el cáncer de mama no cause ningún síntoma. Puede que el bulto sea demasiado pequeño para ser palpable o para provocar cambios inusuales que puedas detectar por tu cuenta. Con frecuencia, aparece una zona anómala en una mamografía de detección (radiografía de la mama), lo que lleva a más análisis.

No obstante, en algunos casos el primer indicio de cáncer de mama es un bulto o masa reciente en la mama. Un bulto indoloro, duro y con bordes irregulares tiene más probabilidades de ser cáncer. Estos son los síntomas más usuales[(2)].

- Inflamación de la mama o parte de ella.
- Irritación cutánea o formación de hoyos.
- Dolor de mama
- Dolor en el pezón o inversión del pezón
- Enrojecimiento, descamación o engrosamiento del pezón o la piel de la mama
- Una secreción del pezón que no sea leche
- Aparición de nódulos en las axilas

2.1.4 Estadios del cáncer de mama:

Los estadios sirven para determinar y describir dónde se encuentra el cáncer, si ha crecido, si se ha diseminado y hacia dónde [(3)]. La herramienta más frecuentemente utilizada es el sistema de determinación de estadios TNM.

- Tumor (T): Tamaño y ubicación.
- Ganglio (N): Si ha habido diseminación a los ganglios. Ubicación y cantidad.
- Metástasis (M): Si ha habido metástasis. Ubicación y gravedad.

Estos resultados se combinan para determinar el estadio de cada paciente. Los estadios están subdivididos para poder concretar mucho más las características del tumor.

Estadio 0 y 1: Los estadios más bajos representan una detección temprana del cáncer. En estas etapas las células cancerosas están confinadas en un área limitada. En el estadio 1 ya representa un tumor invasivo que se ha podido diseminar a los ganglios linfáticos (estadio 1B) o no (estadio 1A).

Estadio 2 (A y B): El cáncer de mama en etapa 2 aún se encuentra en las etapas iniciales, pero hay evidencia de que el cáncer ha comenzado a crecer o diseminarse.

Todavía está contenido en el área de la mama y generalmente se trata de manera muy efectiva.

Estadio 3 (A,B, C): El cáncer de mama en etapa 3 se considera cáncer avanzado con evidencia de cáncer que invade los tejidos circundantes a la mama. El tamaño del tumor empieza a ser considerable y ha afectado a diversos ganglios.

Estadio 4: Metástasis. El cáncer de mama en estadio 4 indica que el tumor puede tener cualquier tamaño y se ha diseminado a otros órganos.

2.1.5 Tratamiento del cáncer de mama

Para determinar el tipo de tratamiento necesario en cada paciente, un equipo multidisciplinar evalúa las características del tumor, el estadio en el que se encuentra y las particularidades del paciente, para poder determinar el tratamiento más adecuado.
El primer paso, en general, recomienda cirugía con el fin de extirpar el tumor, esto puede combinarse con quimioterapia o terapia hormonal con el objetivo de reducir el tamaño del tumor antes de la operación.
El tratamiento post quirúrgico (terapia adyuvante) puede consistir en radioterapia, quimioterapia, terapia dirigida o terapia hormonal para minimizar el riesgo de la recurrencia.

2.1.6 Minería de datos aplicada a la identificación de factores de riesgo del cáncer de mama

Los datos son un recurso valioso, y el trabajo de un científico es dar sentido a esos datos (Witten et al.; 2011). Actualmente existe una cantidad abrumadora de datos. A nivel personal cada elección que hacemos y decisión que tomamos es registrada y almacenada. En el mundo de la economía, el comercio, la sanidad, la industria estos datos son prácticamente infinitos. Lo importante es poder encontrar patrones en ese caótico mundo de información. En minería de datos, los datos son almacenados electrónicamente y la búsqueda es, realizada automáticamente por un ordenador.
Es importante hacer una investigación exhaustiva con tal de poder esclarecer los factores y facilitar la divulgación y la concienciación sobre esta enfermedad. Para ello utiliza como técnica la minería de datos. A diferencia de la lenta y tediosa, puede que incompleta, investigación manual la minería de datos permitirá un análisis rápido y directo después de reunir y combinar diferentes fuentes de información.
Con esta técnica se pueden analizar un gran volumen de datos de manera rápida, sencilla y eficaz. Trabajaremos principalmente en PubMed, ya que posee una gran

cantidad de literatura (Jyoti *et al.*; 2015), y utilizaremos como herramienta Pubmed.mineR por su velocidad y flexibilidad.

La minería de datos se define como el proceso de descubrir automáticamente patrones en los datos, trata de resolver problemas mediante el análisis de bases de datos existentes. Estos patrones descubiertos deben ser significativos, ayudar a explicar algo sobre los datos analizados. También se puede utilizar como herramienta predictiva, para predecir valores futuros o faltantes de otra variable.

2.1.7 Aprendizaje automático

La minería de datos utiliza técnicas de aprendizaje automático (*machine learning*), que crea una inteligencia artificial que permite al ordenador aprender utilizando ejemplos. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos.

Este gran campo del aprendizaje automático se puede diferenciar en dos grandes grupos:

- Aprendizaje Supervisado: Es el más utilizado y requiere de intervención humana para la creación de etiquetas en el histórico de datos de manera que la máquina pueda predecir un resultado probable a partir de las mismas.

- Aprendizaje No Supervisado: El aprendizaje no supervisado es menos común y utiliza datos históricos que no han sido etiquetados. El objetivo es encontrar patrones a partir del propio análisis de datos.

Dentro del *Machine Learning* existen tres métodos de aplicación diferenciados:

- Método de Regresión: Este método se utiliza para predecir el valor de un atributo continuo. Consiste en encontrar la mejor ecuación que atraviese de forma óptima de un conjunto de puntos.

- Método de Agrupación: Este método se utiliza cuando se necesita clasificar las instancias de datos pero no se conocen previamente las categorías. Esta agrupación permite construir grupos (*cluster*) coherentes de instancias teniendo en cuenta las variables de la data.

- Método de Clasificación: Método utilizado para predecir un resultado de un atributo con valor discreto dadas unas características.

El principal vínculo con este trabajo es la aplicación de métodos de agrupación, con la capacidad para identificar *clusters*.

2.2 - Materiales y métodos

La minería de texto *(text mining*) utilizada en este trabajo está dividido en diferentes pasos descritos a continuación (Cohen & Hunter, 2013).

2.2.1 Recopilación de información

Los abstracts para este estudio se obtienen de la base de datos *PubMed*, los resúmenes de los artículos se guardan localmente utilizando su motor de búsqueda como aparece en la figura 1.

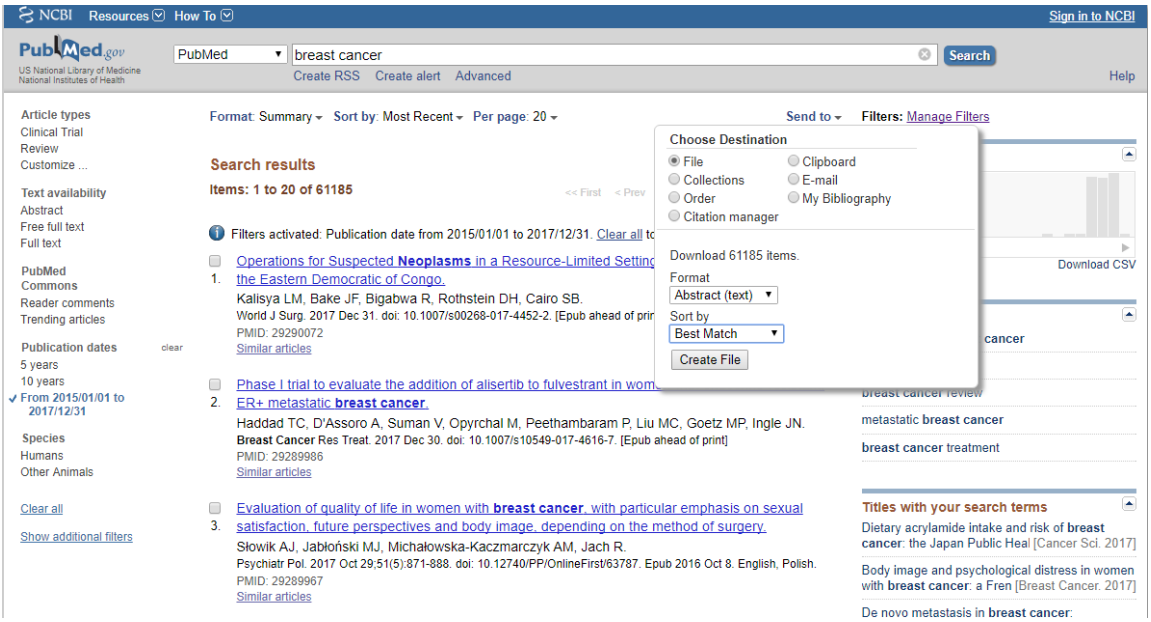


Figura 1. Ejemplo de descargar de los abstracts

En el caso de este estudio en concreto, la selección de abstracts se ha hecho utilizando diferentes palabras clave, filtrando los artículos desde el año 2005 y ordenados por *best match*.

- Breast cancer (n=199968)
- Breast cancer gene (n=47582)
- Breast cancer genetics (n=52113)
- Breast cancer environmental (n=7622)
- Breast cancer risk (n=44070)
- Breast cancer risk factor (n=25310)

Una vez obtenidos los resúmenes, se han ejecutado funciones del paquete de R *pubmed.mineR.* Estos análisis se han llevado a cabo mediante Rstudio.

El primer paso es instalar el nuevo paquete: *pubmed.mineR.*

```
install.packages("pubmed.mineR")
library(pubmed.mineR)
```

Una vez activada la nueva librería se procede a leer el fichero (.txt) obtenido desde *PubMed.*

2.2.2 Clasificación de documentos

La tarea de clasificar documentos en categorías sirve como el primer paso hacia la recopilación sistemática de datos, la conservación, la anotación y para extraer patrones que satisfagan las consultas de los usuarios.

- Para ello se utiliza la función *readabs* del paquete.

```
ejemplo=readabs("ejemplo.txt")
```

Esta función creara un nuevo archivo en el directorio llamado *newabs.txt,* un objeto S4

2.2.3 Resumen

El resumen de documentos es el proceso de construir resúmenes automáticos de los documentos para permitir a los usuarios obtener la esencia en poco tiempo.

2.2.4 Estructura lingüística

La estructura lingüística en este paquete se ha abordado en dos niveles:

(1) tokenización de oraciones y (2) tokenización de palabras. La función word_atomizations () convierte el texto en palabras y las informa en orden descendente de sus frecuencias de ocurrencia. Las palabras comunes en inglés se eliminan automáticamente. Las palabras de mayor rango son términos de alta frecuencia de ocurrencia. Un término se considera importante en función de su frecuencia de ocurrencia (Feinerer et al., 2008). El principio utilizado aquí es que la entidad más pequeña del lenguaje es una 'palabra', y varias palabras están conectadas para construir una oración.

- Se utiliza la función *word_atomizations*.

```
##Tokenización
words<- word_atomizations(ejemplo)
```

De esta función obtenemos el listado palabras contenidos en los abstracts y la frecuencia con la que aparecen. Este listado se crea en un archivo llamado *word_table.txt* donde los signos de puntuación, los espacios y las palabras más frecuentes en ingles son eliminadas.

2.2.5 Reconocimiento y normalización de entidades nombradas

La entidad nombrada más popular es el gen, está escrito como un símbolo genético y su nombre. Se ha incluido en Pubmed.mineR la lista oficial de símbolos genéticos HGNC (Human Gene Nomenclature Committee; Gray *et al*., 2015) para el reconocimiento automático. También incluye los mapeos a Uniprot y Entrez

- La función *gene_atomization* tiene esa función.

```
#Genes
genes= gene_atomization(ejemplo)
```

Esta función se utiliza para encontrar los genes en el texto, usando los datos HGNC.

Se crea un nuevo archivo, *table.txt*, con el nombre de los genes y su frecuencia.

2.2.6 Relaciones

La relación más buscada es la asociación entre genes o la asociación cruzada entre genes y términos, lo que podría implicar su participación en vías vinculadas. También podría implicar factores de riesgo en el caso de enfermedades. Las asociaciones entre genes y términos pueden revelar conexiones entre genotipo y fenotipo.

Se pueden buscar términos concretos y restringir la búsqueda.

- En este caso se utiliza la función *serachabsL*.

```
##Buscar palabras clave
busqueda= searchabsL(ejemplo, include = "termino1", restrict = "termino2")
```

Include para añadir las palabras que quieras buscar, y *restrict* para acotar la búsqueda. También se puede incluir el comando *yr* para filtrar por año.

Para el siguiente paso primero se debe crear un vector con los términos más frecuentes y más relevantes para el estudio.

- La función *tdm_for_lsa* ayuda a crear una matriz que contiene la frecuencia de cada palabra en particular para cada abstract.

```
#Terminos por abstract:
vector = c("")
tdm = tdm_for_lsa(ejemplo, vector)
```

Para acabar se encuentra la asociación entre términos usando el paquete *lsa*.

```
#Paquete lsa
install.packages("lsa")
library(lsa)
```

- El primer paso es crear un espacio semántico latente y después se muestra como una matriz de texto nuevo.

```
#Asosiación entre terminos
lsa = lsa(tdm, dims=dimcalc_share())
matrix = as.textmatrix(lsa)
```

- Para finalizar se encuentra la asociación entre términos

```
associated_words = lapply(tdm, function(x){ associate(matrix, x, measure = "cosine", threshold = "0.7")})
names(associated_words) = tdm
```

En cuanto a las palabras clave de los factores de riesgo ambientales, de han determinado los 7 factores más importantes y que más frecuentemente aparecen tanto en la bibliografía como en los abstracts utilizados.

Existe otro método, basado en el articulo "*pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts*" (Ramachandran *et al.*; 2015) con los pasos siguientes:

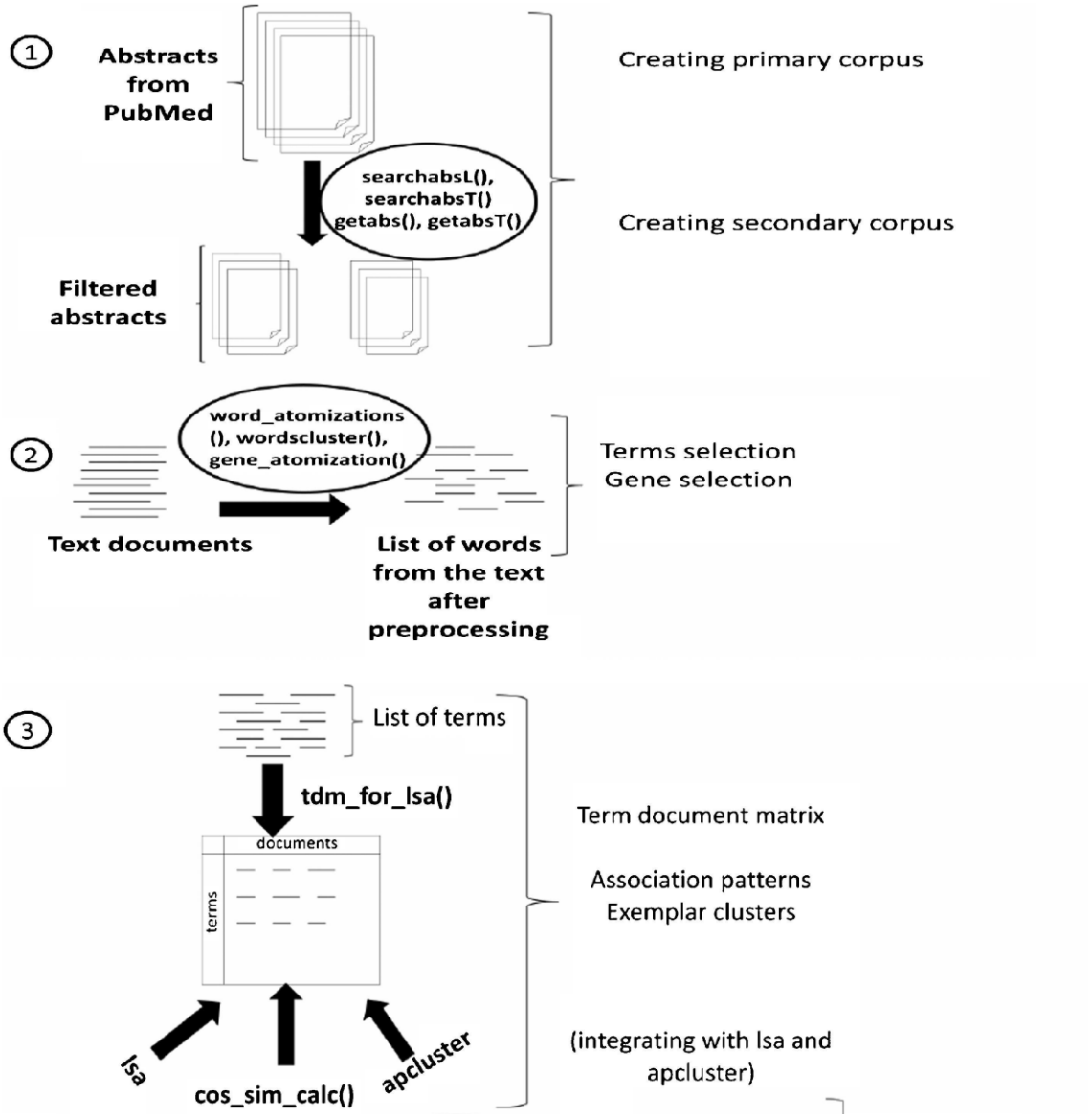


Figura 1. Imagen resumen del artículo mencionado (Ramachandran *et al.:* 2015)

En primer lugar obtenemos el corpus primario al descargar los abstracts con las palabras clave *breast cancer*(n=199968), filtramos desde el año 2005 hasta la actualidad, igual que en el procedimiento anterior. Una vez descargados los resúmenes utilizamos R para leer el archivo.

```
##Leemos el documento descargado
cancer=readabs("breastcancer.txt")
```

Una vez conseguido, se utiliza la función *searchabsL* para filtrar los abstracts por los términos de nuestra elección y obtener el ***corpus secundario***. Crearemos los mismos grupos, con las mismas palabras clave, que en el método anterior:

```
gene= searchabsL(cancer, include = "gene")
genetic= searchabsL(cancer, include = "genetic")
enviro= searchabsL(cancer, include = "environmental")
risk= searchabsL(cancer, include = "risk")
riskfac= searchabsL(cancer, include = "risk factor")
```

De estos comandos obtendremos una tabla (tabla 15) con la el número de abstracts recogidos en cada grupo.

En tercer lugar utilizamos las funciones *word_atomizations y gene_atomization* en cada grupo de resúmenes. La primera función para tokenizar el texto y poder encontrar términos clave de factores de riesgo ambientales (Tabla 17) y la segunda para obtener la lista (Tabla 16) de genes más frecuentes. Para encontrar la frecuencia de factores de riesgo ambientales utilizaremos los mismos términos que en método anterior.

```
##Atomization
wgen<- word_atomizations(gene)
wggen= gene_atomization(gene)

wgenetic<- word_atomizations(genetic)
wggenetic= gene_atomization(genetic)


wenviro<- word_atomizations(enviro)
wgenviro= gene_atomization(enviro)


wrisk<- word_atomizations(risk)
wgrisk= gene_atomization(risk)


wriskfac<- word_atomizations(riskfac)
wgriskfac= gene_atomization(riskfac)
```

El siguiente paso es crear una matriz con los grupos:

```
tdm_cancer = c("gene", "genetic", "environmental", "risk", "risk factor")
tdmcancer = tdm_for_lsa(cancer, tdm_cancer)

lsacancer = lsa(tdmcancer, dims=dimcalc_share())
matrixcancer = as.textmatrix(lsacancer)
```

Y aplicar la función *cos_sim_calc* calcula la medida del coseno de similitud entre pares de términos del corpus (Tabla 18).

```
cos_sim_calc(matrixcancer)
```

Para finalizar, con los dos grupos con mayor numero de abstracts, *gene y risk*, se crea una matriz con los genes y factores ambientales más frecuentes para intentar encontrar relaciones entre ellos utilizando la función *cos_sim_calc y associated_words.*

```
#Grupo gene
tdm_gene = c("age", "estrogen", "family", "radiation","weight", "BRCA1","BRCA2", "HR", "TP53")
tdmgene = tdm_for_lsa(gene, tdm_gene)

lsagene = lsa(tdmgene, dims=dimcalc_share())
matrixgene = as.textmatrix(lsagene)

associated_wordsg = lapply(tdm_gene, function(x){ associate(matrixgene, x, measure = "cosine", threshold = "0.7")
names(associated_wordsg) = tdm_gene
associated_wordsg

cos_sim_calc(matrixgene)

#Grupo risk
tdm_risk = c("age", "estrogen", "family", "radiation","weight", "BRCA1","BRCA2", "HR", "TP53")
tdmrisk = tdm_for_lsa(risk, tdm_risk)

lsarisk = lsa(tdmrisk, dims=dimcalc_share())
matrixrisk = as.textmatrix(lsarisk)

associated_wordsr = lapply(tdm_risk, function(x){ associate(matrixrisk, x, measure = "cosine", threshold = "0.7")
names(associated_wordsg) = tdm_risk
associated_wordsr

cos_sim_calc(matrixrisk)
```

2.3 - Resultados

Aplicando en Rstudio el script de R con los pasos anteriormente descritos para los seis archivos obtenidos desde PubMed, estos son los resultados.

En primer lugar se hace el análisis para los abstracts obtenidos en la búsqueda de *breast cancer*, para conseguir una visión amplia y general.

2.3.1 *Breast cancer*

Factores de riesgo genéticos:

Gene symbol	Genes	Freq
BRCA1	breast cancer 1, early onset	852
BRCA2	breast cancer 2, early onset	458
HER2	epidermal growth factor receptor 2	340
EGFR	epidermal growth factor receptor	252
T	T, brachyury homolog (mouse)	231

Tabla 1: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer*.

Factores de riesgo ambientales:

Factor	Freq
Age	2911
Family	1132
Alcohol	202
Smoking	260
Radiation	837
Estrogens	1639
Weight	323

Tabla 2: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer*.

```
> associated_words
$age
radiation  exercise    family   smoking    weight   alcohol
0.9916299 0.9842792 0.9725164 0.9549051 0.9454637 0.7529441

$family
 exercise   smoking    weight       age radiation   alcohol
0.9983508 0.9977920 0.9953195 0.9725164 0.9343145 0.8854751

$alcohol
      weight       smoking        family progesterone     exercise     estrogen          age
   0.9262375    0.9143828    0.8854751    0.8793020    0.8573378    0.7669743    0.7529441

$smoking
   weight    family  exercise       age   alcohol radiation
0.9995402 0.9977920 0.9923335 0.9549051 0.9143828 0.9085774

$radiation
      age  exercise    family   smoking    weight
0.9916299 0.9532368 0.9343145 0.9085774 0.8954945

$progesterone
 estrogen   alcohol
0.9800105 0.8793020

$estrogen
progesterone      alcohol
   0.9800105    0.7669743

$exercise
   family   smoking    weight       age radiation   alcohol
0.9983508 0.9923335 0.9881301 0.9842792 0.9532368 0.8573378

$weight
  smoking    family  exercise       age   alcohol radiation
0.9995402 0.9953195 0.9881301 0.9454637 0.9262375 0.8954945
```

2.3.2 *Breast cancer gene*

Factores de riesgo genéticos:

Gene symbol	Genes	Freq
BRCA1	breast cancer 1, early onset	5467
BRCA2	breast cancer 2, early onset	2981
TP53	tumor protein p53	732
EGFR	epidermal growth factor receptor	564
T	T, brachyury homolog (mouse)	474

Tabla 3: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer gene*.

Factores de riesgo ambientales:

Factor	Freq
Age	1690
Family	1965
Alcohol	88
Smoking	94
Radiation	416
Estrogens	2383
Weight	71

Tabla 4: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer gene*.

```
> associated_words
$age
  smoking    family radiation    weight   alcohol  exercise
0.9999155 0.9975558 0.9946819 0.9859247 0.9374025 0.7268228

$family
radiation   smoking       age    weight   alcohol  exercise
0.9994474 0.9983799 0.9975558 0.9951972 0.9594450 0.7730379

$alcohol
   weight radiation    family   smoking       age  exercise
0.9824319 0.9682850 0.9594450 0.9418509 0.9374025 0.9205108

$smoking
      age    family radiation    weight   alcohol  exercise
0.9999155 0.9983799 0.9959369 0.9880150 0.9418509 0.7356908

$radiation
   family    weight   smoking       age   alcohol  exercise
0.9994474 0.9979011 0.9959369 0.9946819 0.9682850 0.7936969

$progesterone
 estrogen  exercise
0.9567915 0.9155677

$estrogen
progesterone     exercise
   0.9567915    0.7590681

$exercise
     alcohol progesterone       weight    radiation       family     estrogen      smoking          age
   0.9205108    0.9155677    0.8314228    0.7936969    0.7730379    0.7590681    0.7356908    0.7268228

$weight
radiation    family   smoking       age   alcohol  exercise
0.9979011 0.9951972 0.9880150 0.9859247 0.9824319 0.8314228
```

2.3.3 *Breast cancer genetics*

Factores de riesgo genéticos:

Gene symbol	Genes	Freq
BRCA1	breast cancer 1, early onset	2307
BRCA2	breast cancer 2, early onset	1252
TP53	tumor protein p53	366
EGFR	epidermal growth factor receptor	355
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	286

Tabla 5: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer genetics*.

Factores de riesgo ambientales:

Factor	Freq
Age	1644
Family	2045
Alcohol	90
Smoking	137
Radiation	379
Estrogens	1819
Weight	127

Tabla 6: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer genetics*.

```
> associated_words
$age
   family    weight   smoking   alcohol radiation  exercise
0.9994152 0.9958950 0.9829806 0.9611277 0.9590823 0.9386212

$family
      age    weight   smoking   alcohol radiation  exercise
0.9994152 0.9922176 0.9761242 0.9511248 0.9488405 0.9262775

$alcohol
radiation  exercise   smoking    weight       age    family
0.9999732 0.9973769 0.9954928 0.9821741 0.9611277 0.9511248

$smoking
   weight   alcohol radiation  exercise       age    family
0.9955741 0.9954928 0.9947725 0.9860169 0.9829806 0.9761242

$radiation
  alcohol  exercise   smoking    weight       age    family
0.9999732 0.9978797 0.9947725 0.9807728 0.9590823 0.9488405

$progesterone
 estrogen
0.9708621

$estrogen
progesterone
   0.9708621

$exercise
radiation   alcohol   smoking    weight       age    family
0.9978797 0.9973769 0.9860169 0.9659916 0.9386212 0.9262775

$weight
      age   smoking    family   alcohol radiation  exercise
0.9958950 0.9955741 0.9922176 0.9821741 0.9807728 0.9659916
```

2.3.4 *Breast cancer environmental*

Factores de riesgo genéticos:

Gene symbol	Genes	Freq
BRCA1	breast cancer 1, early onset	511
T	T, brachyury homolog (mouse)	415
BRCA2	breast cancer 2, early onset	238
HER2	epidermal growth factor receptor 2	157
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	116

Tabla 7: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer environmental*.

Factores de riesgo ambientales:

Factor	Freq
Age	1601
Family	906
Alcohol	372
Smoking	362
Radiation	840
Estrogens	1358
Weight	455

Tabla 8: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer environmental*.

```
> associated_words
$age
radiation    family  exercise   smoking    weight   alcohol
0.9979814 0.9932923 0.9870934 0.9860698 0.9439566 0.9343125

$family
 exercise   smoking radiation       age    weight   alcohol
0.9989900 0.9986886 0.9986305 0.9932923 0.9757910 0.9692626

$alcohol
      weight       smoking     exercise       family    radiation          age progesterone
   0.9996054    0.9805872    0.9793384    0.9692626    0.9550638    0.9343125    0.7899384

$smoking
 exercise    family radiation       age    weight   alcohol
0.9999803 0.9986886 0.9946425 0.9860698 0.9857082 0.9805872

$radiation
   family       age  exercise   smoking    weight   alcohol
0.9986305 0.9979814 0.9952712 0.9946425 0.9630128 0.9550638

$progesterone
 estrogen   alcohol    weight
0.9007427 0.7899384 0.7724024

$estrogen
progesterone
   0.9007427

$exercise
  smoking    family radiation       age    weight   alcohol
0.9999803 0.9989900 0.9952712 0.9870934 0.9846325 0.9793384

$weight
     alcohol       smoking     exercise       family    radiation          age progesterone
   0.9996054    0.9857082    0.9846325    0.9757910    0.9630128    0.9439566    0.7724024
```

2.3.5 *Breast cancer risk*

Factores de riesgo genéticos:

Gene symbol	Gene	Freq
BRCA1	breast cancer 1, early onset	1268
BRCA2	breast cancer 2, early onset	951
HER2	epidermal growth factor receptor 2	606
T	T, brachyury homolog (mouse)	298
CPM	carboxypeptidase M	172

Tabla 9: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer risk*.

Factores de riesgo ambientales:

Factor	Freq
Age	5904
Family	2066
Alcohol	820
Smoking	818
Radiation	1238
Estrogens	1926
Weight	1054

Tabla 10: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer risk*.

```
> associated_words
$age
   family radiation   smoking   alcohol
0.9962657 0.8608231 0.7772138 0.7186637

$family
      age radiation   smoking   alcohol
0.9962657 0.8936870 0.7470553 0.7042752

$alcohol
     smoking     exercise progesterone          age       family
   0.9725849    0.9247103    0.8129679    0.7186637    0.7042752

$smoking
  alcohol  exercise       age    family
0.9725849 0.9550600 0.7772138 0.7470553

$radiation
   family       age
0.8936870 0.8608231

$progesterone
 estrogen   alcohol
0.9750714 0.8129679

$estrogen
progesterone
   0.9750714

$exercise
  smoking   alcohol    weight
0.9550600 0.9247103 0.8036605

$weight
 exercise
0.8036605
```

2.3.6 *Breast cancer risk factor*

Factores de riesgo genéticos:

Gene symbol	Genes	Freq.
BRCA1	breast cancer 1, early onset	1194
BRCA2	breast cancer 2, early onset	910
HER2	epidermal growth factor receptor 2	608
T	T, brachyury homolog (mouse)	330
SLN	sarcolipin	167

Tabla 11: Tabla con los 5 genes más frecuentes en los abstracts *breast cancer risk factor*.

Factores de riesgo ambientales:

Factor	Freq
Age	6042
Family	1965
Alcohol	934
Smoking	938
Radiation	1199
Estrogens	1853
Weight	1191

Tabla 12: Tabla con los 7 factores ambientales más frecuentes en los abstracts *breast cancer risk factor*.

```
> associated_words
$age
   family radiation   smoking   alcohol
0.9989279 0.9430020 0.8685864 0.8039816

$family
      age radiation   smoking   alcohol
0.9989279 0.9459241 0.8722789 0.8096855

$alcohol
     smoking     exercise       family          age progesterone
   0.9927896    0.9322758    0.8096855    0.8039816    0.7918763

$smoking
     alcohol     exercise       family          age progesterone
   0.9927896    0.9173075    0.8722789    0.8685864    0.7262622

$radiation
   family       age
0.9459241 0.9430020

$progesterone
 estrogen   alcohol   smoking
0.9741835 0.7918763 0.7262622

$estrogen
progesterone
   0.9741835

$exercise
  alcohol   smoking    weight
0.9322758 0.9173075 0.7282286

$weight
 exercise
0.7282286
```

Tabla resumen de genes

	1	2	3	4	5
Breast Cancer	BRCA1	BRCA2	HER2	EGFR	T
Breast cancer gene	BRCA1	BRCA2	TP53	EGFR	T
Breast cancer genetics	BRCA1	BRCA2	TP53	EGFR	PIK3CA
Breast cancer environmental	BRCA1	T	BRCA2	HER2	CYP1A1
Breast cancer risk	BRCA1	BRCA2	HER2	T	CPM
Breast cancer risk factor	BRCA1	BRCA2	HER2	T	SLN

Tabla 13: Tabla resumen de los cinco genes principales en todos los abstracts.

Tabla resumen de factores ambientales

	Age	Family	Alcohol	Smoking	Radiation	Estrogens	Weight
Breast Cancer	2911	1132	202	260	837	1639	323
Breast cancer gene	1690	1965	88	94	416	2383	71
Breast cancer genetics	1644	2045	90	137	379	1819	127
Breast cancer environmental	1601	906	372	362	840	1358	455
Breast cancer risk	5904	2066	820	818	1238	1926	1054
Breast cancer risk factor	6042	1965	934	938	1199	1853	1191

Tabla 14: Tabla resumen de los 7 factores ambientales más frecuentes en los abstracts.

2.3.7 Análisis del Corpus secundario

En primer lugar obtenemos la n de los diferentes grupos. Como podemos ver, los dos grupos más grandes son el de *Gene y Risk*.

Grupo	Número de abstracts
Gene	4012
Genetic	1128
Environmental	164
Risk	2924
Risk factor	709

Tabla 15: Corpus secundario, obtenida de la función *searchabsL*, número de resúmenes por grupo.

Después de aplicar la función *gene_atomization*, esta es la tabla con los 5 genes más frecuentes en cada grupo y sus frecuencias.

Grupo	1	2	3	4	5
gene	BRCA1 (687)	BRCA2 (368)	EGFR (147)	PIK3CA (115)	TP53 (115)
genetics	BRCA1 (409)	BRCA2 (250)	TP53 (59)	PIK3CA (54)	CHEK2 (46)
environmental	BRCA1 (20)	CRIP1 (14)	GSTM1 (13)	CDH1 (10)	APC (10)
risk	BRCA1 (462)	BRCA2 (323)	HER2 (176)	CPM (122)	T (70)
risk factor	BRCA1 (92)	BRCA2 (56)	HER2 (40)	CPM (37)	ESR1 (14)

Tabla 16: Corpus secundario. Tabla resumen de frecuencia de genes.

Después de aplicar la función *word_atomizations*, esta es la tabla con los 7 factores ambientales elegidos y sus frecuencias.

Grupo	Age	Family	Alcohol	Smoking	Radiation	Estrogens	Weight
gene	1051	598	101	115	211	805	90
genetics	383	275	45	51	60	194	14
environmental	63	36	36	28	13	40	7
risk	1535	692	183	818	288	626	216
risk factor	593	273	85	108	56	174	88

Tabla 17: Corpus secundario. Tabla resumen de frecuencia de genes.

Resultados de aplicar la función *cos_sim_calc* a los grupos.

	Gen	Genetic	Environmental	Risk	Risk factor
Gen	Null	0.99	0.77	0.22	0.21
Genetic	0.99	Null	0.84	0.34	0.33
Environmental	0.77	0.84	Null	0.79	0.78
Risk	0.22	0.34	0.79	Null	0.99
Risk factor	0.21	0.33	0.78	0.99	Null

Tabla 18. Resultados de la función *cos_sim_calc* aplicadas a la matriz de grupos.

Al aplicar la función *associated_words* al grupo *gene* obtenemos estos resultados:

```
> associated_wordsg
$age
radiation    weight    family
0.9936883 0.8221374 0.7376879

$estrogen
       HR    weight
0.9393536 0.8625254

$family
     TP53 radiation     BRCA2     BRCA1       age
0.9520672 0.7822123 0.7808752 0.7725765 0.7376879

$radiation
      age    weight    family
0.9936883 0.8615108 0.7822123

$weight
 estrogen radiation       age
0.8625254 0.8615108 0.8221374

$BRCA1
    BRCA2      TP53    family
0.9975278 0.9186405 0.7725765

$BRCA2
    BRCA1      TP53    family
0.9975278 0.9155276 0.7808752

$HR
 estrogen
0.9393536

$TP53
   family     BRCA1     BRCA2
0.9520672 0.9186405 0.9155276
```

Al aplicar la función *associated_words* al grupo *risk* obtenemos estos resultados:

```
> associated_wordsr
$age
radiation    family    weight  estrogen
0.9997592 0.8541885 0.7267299 0.7214900

$estrogen
   weight        HR radiation       age
0.9902745 0.8191038 0.7280707 0.7214900

$family
radiation       age      TP53
0.8630634 0.8541885 0.8140066

$radiation
      age    family    weight  estrogen
0.9997592 0.8630634 0.7304067 0.7280707

$weight
 estrogen        HR radiation       age
0.9902745 0.7997651 0.7304067 0.7267299

$BRCA1
    BRCA2      TP53
0.9998519 0.9633939

$BRCA2
    BRCA1      TP53
0.9998519 0.9666269

$HR
 estrogen    weight
0.8191038 0.7997651

$TP53
    BRCA2     BRCA1    family
0.9666269 0.9633939 0.8140066
```

Cuando aplicamos la función *cos_sim_calc* a los dos grupos obtenomes estas tablas.

Gene -Factor 1	Gene-Factor 2	Sim.
age	BRCA1	0,15
age	BRCA2	0,18
age	HR	0,15
age	TP53	0,50
estrogen	family	0,49
estrogen	radiation	0,56
estrogen	weight	0,86
estrogen	BRCA1	0,17
estrogen	BRCA2	0,12
estrogen	HR	0,94
estrogen	TP53	0,45
family	radiation	0,78
family	weight	0,57
family	BRCA1	0,77
family	BRCA2	0,78
family	HR	0,22
family	TP53	0,95
radiation	weight	0,86
radiation	BRCA1	0,21
radiation	BRCA2	0,23
radiation	HR	0,24
radiation	TP53	0,56
weight	BRCA1	-0,00
weight	BRCA2	-0,02
weight	HR	0,67
weight	TP53	0,39
BRCA1	BRCA2	1,00
BRCA1	HR	0,05
BRCA1	TP53	0,92
BRCA2	HR	-0,01
BRCA2	TP53	0,92
HR	TP53	0,25

Tabla 19. Función cos_sim_calc en grupo gene.

Risk - Factor 1	Risk - Factor 2	Sim
age	estrogen	0,72
age	family	0,85
age	radiation	1,00
age	weight	0,73
age	BRCA1	0,15
age	BRCA2	0,16
age	HR	0,19
age	TP53	0,41
estrogen	family	0,55
estrogen	radiation	0,73
estrogen	weight	0,99
estrogen	BRCA1	0,11
estrogen	BRCA2	0,11
estrogen	HR	0,82
estrogen	TP53	0,31
family	radiation	0,86
family	weight	0,48
family	BRCA1	0,63
family	BRCA2	0,64
family	HR	0,07
family	TP53	0,81
radiation	weight	0,73
radiation	BRCA1	0,17
radiation	BRCA2	0,18
radiation	HR	0,20
radiation	TP53	0,43
weight	BRCA1	-0,03
weight	BRCA2	-0,03
weight	HR	0,80
weight	TP53	0,18
BRCA1	BRCA2	1,00
BRCA1	HR	0,03
BRCA1	TP53	0,96
BRCA2	HR	0,03
BRCA2	TP53	0,97
HR	TP53	0,10

Tabla 20. Tabla 19. Función cos_sim_calc en grupo risk.

2.4 - Discusión

En primer lugar haremos un análisis de los factores de riesgo genéticos partiendo de los datos obtenidos en el apartado anterior. Como se puede ver, no existen prácticamente diferencias en los resultados causados por los diferentes métodos.

2.4.1 Factores de riesgo genéticos

En todos los resultados exceptuando el de las palabras clave *breast cancer environmental*, el gen que ha salido siempre en primera posición es *breast cancer 1* (BRCA1), seguido de *breast cancer 2* (BRCA2).

Estos genes humanos son responsables de codificar proteínas. Estas proteínas ayudan a reparar el material genético y evitan la formación de tumores (Kumar *et al*.; 2017). Aunque sus funciones específicas aún no están del todo definidas, la función de la proteína codificada por BRCA 1 forma complejo con otras proteínas implicadas en la integridad del genoma, reparando errores en la doble cadena de ADN. En cambio BRCA 2 está más centrada en la regulación de la expresión génica (Duarte et al.; 2002). Se han descrito cerca de 460 mutaciones germinales diferentes en el gen BRCA1 y cerca de 200 para el gen BRCA2. La mayoría de estas mutaciones (80%) originan un codón de parada (stop) que lleva a la síntesis de una proteína truncada (Duarte et al.; 2002). Estos genes tienen un problema añadido, son los responsable del 90% de los casos de cáncer de mama hereditario (Takahashi et al.; 1996), que representar cerca del 10% de los casos totales (Claus *et al.,* 1991).

El siguiente gen que aparece con mucha frecuencia, y en el tercer lugar de las palabras claves relacionadas con genética específicamente, *breast cancer gen* y *breast cancer genetics* es el gen *TP53. TP53* participa en diversas funciones fisiológicas de la célula para mantener la integridad y estabilidad genómica, referido así como un gen maestro *(Lane; 1992). TP53 es un gen supresor de tumores y* es uno de los genes más importantes y estudiados en la genética del cáncer, ya que se encuentra mutado en más del 50% de todos los tipos de cáncer humano y codifica para una proteína multifuncional cuya deficiencia contribuye a la inestabilidad genómica que conduce a la acumulación de mutaciones y a la aceleración en el desarrollo del tumor (Rangel; 2006). Lemoine (1994) señala que la mutación de la p53, proteína sintetizada a partir de TP53, caracteriza una forma altamente agresiva de cáncer de mama, asociada con un pobre pronóstico tanto en los casos con ganglios axilares positivos o negativos (Li; 1982).

Otro de los genes con más frecuencia de aparición en los abstracts es el gen EGFR. EGFR forma parte de una red de señalización que es componente central de varios procesos celulares críticos, como el crecimiento, la proliferación y la motilidad celulares (Shepherd.; 2005). *EGFR* codifica para una proteína localizada en la membrana de las células que actúa como receptor del factor de crecimiento epidérmico. Cuando este factor está presente, se une a la proteína EGFR y se inicia una cascada de respuestas celulares que llevan a la proliferación celular. Numerosos cánceres presentan mutaciones en el gen *EGFR* que inducen una sobreproducción o sobre activación del receptor y por tanto, llevan a que las células se dividan de forma excesiva (Kohsaka *et al*.; 2017). Este gen esta mayormente relacionado al cáncer de pulmón. Aunque puede influir en la aparición de cáncer de mama, probablemente aparece tan referenciado en los abstracts ya que la inhibición del receptor producido por este gen puede ser un tratamiento contra los canceres de mama triple negativos. La actividad de proteína quinasa activada por AMP (AMPK) regula la actividad de EGFR en células de cáncer de mama, por lo tanto, podría proporcionar un beneficio terapéutico en tales cánceres (Jhaveri *et al*.; 2015).

El siguiente en la lista es un gen con un comportamiento muy parecido al explicado anteriormente, el gen HER2. El gen *HER2* produce las proteínas HER2. La función normal de estas proteínas es la de receptores en las células mamarias. En los casos en que los receptores HER2 funcionan correctamente, ayudan a controlar la proliferación de las células (Ménard *et al*.; 2000). Pero en el 25% de los casos de cáncer de mama, el gen *HER2* presenta una mutación en el dominio transmembrana (Colomer *et al*.; 2001). Esta mutación provoca una sobreexpresión de este gen (amplificación), que produce una proliferación incontrolada de células, dando lugar al tumor.

Para finalizar, el gen T aparece en gran frecuencia en todas las búsquedas. Este gen no es específico para el cáncer de mama. el factor de transcripción T-box Brachyury induce en las células tumorales la transición epitelio-mesenquimal (EMT), un paso importante en la progresión de los tumores primarios hacia la metástasis (Romaine *et al.;* 2010).

En cuanto a los demás genes, sus frecuencias de aparición no son demasiado altas y no se repiten en las diferentes búsquedas. La gran mayoría de ellos son reguladores de algún paso en la división celular, por lo que mutaciones podrían provocar una proliferación celular errónea.

2.4.2 Factores de riesgo ambientales

En segundo lugar haremos un análisis de los factores de riesgo ambientales partiendo de los datos obtenidos en el apartado anterior.

Partiendo de la tabla 14, se pueden ver dos grandes grupos. En el primer grupo vemos como las búsquedas relacionadas con factores genéticos, tiene una frecuencia baja de términos relacionados con factores de riesgo ambientales, como es lógico. El segundo grupo donde la búsqueda se centra específicamente en factores de riesgo ambientales, vemos una frecuencia de aparición mucho más alta de las palabras clave. Aun y así, las proporciones en ambos grupos son muy similares.

Sin duda, el factor de riesgo que con más frecuencia aparece en todos los casos es la edad. El riesgo de desarrollar cáncer de mama, como la mayoría de canceres, aumenta con la edad. El riesgo empieza a ser significativo en mujeres de 50 años o más (DeSantis *et al.; 2017*). Vemos que la edad siempre aparece muy relacionada con la palabra clave *family* y *radiation*.

Vemos que el segundo factor de mucho peso es el relacionado con la palabra *family*, el cáncer de mama familiar. Como ya hemos visto anteriormente, este tipo de cáncer de mama, aproximadamente el 10%, es el que se hereda a través de los genes BRCA1 y BRCA2 (Cao et al.; 2013). Estos dos genes no son los únicos causantes, aunque representan aproximadamente el 80% de los casos. Otros genes que pueden ser causantes son el CHEK2, el PTEN y el TP53, que aparecen en la lista de factores de riesgo genéticos. El cáncer familiar suele aparecer a edades más tempranas.

Otra palabra clave que aparece con mucha frecuencia es la de *estrogens.* Es una hormona sexual esteroidea, producida generalmente por los ovarios (Samavat & Kurzer; 2014). El estrógeno es un estimulador de las células mamarias, por lo tanto, una exposición prolongada puede aumentar el riesgo de aparición del cáncer.

El sobrepeso puede ser un desencadenante, entre otras cosas, por su relación con el estrógeno. En las mujeres postmenopáusicas, el tejido graso es la principal fuente de estrógeno, una vez los ovarios dejan de producir la hormona. Una mayor cantidad de tejido graso, implica niveles de estrógeno más elevados.

En cuanto al alcohol, un consumo abundante, sobretodo en la adolescencia, implica un aumento de probabilidades de sufrir cáncer. Además puede incrementar los niveles de

estrógenos, y dañar el ADN. El tabaco, por otro lado, da resultados diversos en su relación con el cáncer de mama (Gaudet *et al*.; 2013). Existe una relación en mujeres que fumaban antes de tener su primer hijo y la propensión a padecer el cáncer, así como adelantar la edad de aparición. Vemos que estos términos siempre aparecen asociados al ejercicio, ya que al practicarlo, reduces el impacto de estos factores.

Para finalizar, la radiación es otro factor importante, aunque no exclusivo del cáncer de seno. La radiación ionizante provoca mutaciones en el ADN, dañándolo, que pueden alterar oncogenes que acaben produciendo cáncer. El cáncer de mama aumenta al recibir esta radiación en el pecho, sobre todo a edades tempranas.

2.4.3 Relación entre factores

Antes de discutir sobre la relación entre los factores genéticos y ambientales, hacer un pequeño comentario a la tabla 18, donde vemos que los grupos *gene y gene factor* están muy relacionados entre ellos, igual que los grupos *risk y risk factor*. En cambio entre los grupos de genes y los de factores la relación es muy pequeña. Por otro lado el grupo de *environmental* se encuentra relacionado en menor medida con todos los demás.

Analizamos en primer lugar el grupo *gene*.

Vemos que tanto en la asociación de palabras como en la similaridad hay factores que se encuentran relacionados. Vemos que los genes BRAC1 y BRCA2, así como el TP53 se encuentran relacionados con *family*. Posiblemente por la heredabilidad de estos genes. Vemos una relación también entre el peso y los estrógenos, como ya se había comentado en el apartado anterior. Estos están relacionados con el gen HER2. Estas serian las relaciones más directas para este grupo, superiores al 85%, hay otras que, aunque menores, también son importantes.

En el caso del grupo *risk,* encontramos exactamente, aunque con pequeñas diferencias, los mismos resultados que en el caso del grupo *gene*. Se añade, además, una nueva relación entre el termino *family* y la radiación. Asociación que encontremos, menos prenunciada en el grupo anterior.

2.4.4 Ranking de factores de riesgo

Al tener en cuenta la frecuencia de aparición de los factores en los abstracts, este sería el ranking de factores de riesgo.

1 - BRCA1/ BRCA2

2 - Edad

3 - Cáncer de mama familiar

4 – Estrógenos

5- Peso

6 - HER2/ TP53

7 - Radiación

8 - EFGR

9 - Alcohol / Tabaco

3. Conclusiones

Gracias a este estudio se pueden extraer diversas conclusiones. El primero es que los genes BRCA1 y BRCA2 mutados son el componente genético más importante en la aparición de cáncer de mama. Son los causantes de la mayoría de canceres de mama familiares. Las mutaciones en estos genes pueden deberse a factores ambientales, entre ellos la edad o la radiación tiene un papel muy importante. También se ha visto que el sobrepeso es un factor de riesgo destacable, sobre todo si se combina con la sobreexpresión de la hormona estrógeno.

Aunque sí que se han cumplido los objetivos marcados para el trabajo, el análisis de los datos y los resultados obtenidos han dejado un poco que desear, el estudio se ha realizado utilizando un paquete de R poco extendido, por lo que la información referente a él es limitada.

Las futuras líneas de trabajo pasan, a la vez, en los avances en el estudio del cáncer. Se detectaran nuevos genes, nuevos factores de riesgo, tal vez más importantes que los mencionados en este TFM. Con la nueva información y una amplia utilización del *data mining*, las posibilidades de nuevos estudios son inimaginables.

4 - Bibliografía

Weinberg, RA, **The biology of cancer**, 2nd ed., Garland Sciencien, New York, 2014.

Hoboken, NJ & Wiley-Blackwell, **The Molecular biology of cancer : a bridge from bench to bedside**, 2nd ed., 2013.

(1) http://www.breastcancer.org/symptoms/understand_bc/statistics

Margolese, RG., Hortobagyi, GN., Buchholz, TA., **Holland-Frei Cancer Medicine,** Kufe DW, Pollock RE, Weichselbaum RR, et al., editors , 6th edition. Hamilton (ON): BC Decker; 2003.

Sun H, Zou J, Chen L, Zu X, Wen G, Zhong J., **Triple-negative breast cancer and its association with obesity,** Mol Clin Oncol, 935-942, 2017.

(2) https://www.aecc.es/SobreElCancer/CancerPorLocalizacion/CancerMama/Paginas/sintomas.aspx

(3) http://www.nationalbreastcancer.org/breast-cancer-stages

Witten IH., Frank E., Mark A., Data Mining: Practical Machine Learning Tools and Techniques**, Elsevier, 2nd ed., San Francisco, 2005.**

Jyoti,R., Shah, A. R., Ramachandran, S. **Pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts.** Biosci, 40(4):671-82, 2015.

Cohen KB, Hunter LE, Chapter 16: text mining for translational bioinformatics**, PLoS Comput Biol. Apr;9(4):e1003044. doi: 10.1371/journal.pcbi.1003044, 2013.**

Gray KA., Yates B., Seal RL., Wright MW., Bruford EA., **Genenames.org: the HGNC resources in 2015.** Nucleic Acids Res.;43(Database issue):D1079-85. doi: 10.1093/nar/gku1071, 2014.

Kumar M, Sahu RK, Goyal A, Sharma S, Kaur N, Mehrotra R, Singh UR, Hedau S. **BRCA1 Promoter Methylation and Expression - Associations with ER+, PR+ and HER2+ Subtypes of Breast Carcinoma,** Asian Pac J Cancer Prev.;18(12):3293-3299, 2017.

Duarte[a], J F., Cameselle-Teijeiro[b], R. Soares[a], C., Seixas, M E.,Cortizo-Torres, F., **Analysis of mutations in genes BRCA1 and BRCA2 among patients with breast and ovarian cancer in northern Portugal and Galicia,** Rev Clin Esp.;202(5):259-63, 2002.

Takahashi H,Chiu H-C,Bandera CA,Behbakht K,Liu PC,Couch FJ,et al. **Mutations of the BRCA2 gene in ovarian carcinomas.** Cancer Res, 56, pp. 2738-41, (1996).

Claus EB,Risch N,Thompson WD. **Genetic analysis of breast cancer in the cancer and steroid hormone study.** Am J Hum Genet, 48 (1991), pp. 232-42

Lane DP. **P53, guardian of the genome**. *Nature*; 358: 15–16, 1992.

Rangel, A., Piña, P., Salcedo, M., **Genetic variations of the tumor suppressor TP53: Outstanding and strategies of analysis,** Rev Invest Clin, 58(3):254-64, 2006.

Li FP,Fraumeni JF. **Prospective study of a family cancer syndrome.** JAMA, 247 ;pp. 2692-4, 1982.

Shepherd,J. Rodrigues Pereira,T. Ciuleanu,EH Tan,V Hirsh,S Thongprasert, **Erlotinib in previously treated non-small-cell lung cancer**, N Engl J Med, 353, pp. 123-132, 2005

Kohsaka S, et al. *A **method of high-throughput functional evaluation of EGFR gene variants of unknown significance in cancer*.** Sci transl Med. 2017.

Jhaveri TZ, Woo J, Shang X, Park BH, Gabrielson E**, AMP-activated kinase (AMPK) regulates activity of HER2 and EGFR in breast cancer.** Oncotarget. 20;6(17):14754-65, 2015.

Ménard S, Tagliabue E, Campiglio M, Pupa SM. **Role of HER2 gene overexpression in breast carcinoma.** J Cell Physiol.182(2):150-62, 2000.

Colomer, S. Montero, S. Ropero, JA. Menéndez, H. Cortés Funes, M. Solanas[b], E. Escrich **El oncogén HER2 como ejemplo del progreso diagnóstico y terapéutico en cáncer de mama**, Rev Senol Patol Mamar;14:8-19, 2001

Romaine I. Fernando, Mary Litzinger, Paola Trono, Duane H. Hamilton, Jeffrey Schlom, and Claudia Palena, **The T-box transcription factor Brachyury promotes epithelial-mesenchymal transition in human tumor cells,** J Clin Invest, 1; 120(2): 533–544, 2010.

DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. **Breast cancer statistics, 2017, racial disparity in mortality by state.** CA Cancer J Clin. 2017

Cao, W., Wang, X., Li, JC. **Hereditary Breast Cancer in the Han Chinese Population,** J Epidemiol.; 23(2): 75–84. 2013

Samavat, H, Kurzer, M., **Estrogen Metabolism and Breast Cancer,** Cancer Lett.;356(2 Pt A):231-43. doi: 10.1016/j.canlet.2014.04.018. 2014

Gaudet MM1, Gapstur SM, Sun J, Diver WR, Hannan LM, Thun MJ. Active Smoking and Breast Cancer Risk: Original Cohort Data and Meta-Analysis**, J Natl Cancer Inst.;105(8):515-25. 2013.**