Estudio combinado de marcadores fenotípicos y genotípicos para la evaluación del riesgo individualizado a desarrollar cáncer de mama

Tesis doctoral presentada por

Juan Carlos Triviño Pardo

Dirigida por

Dr. D. Javier Benítez Ortiz

Dr. D. David Moratal Pérez

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Valencia, Marzo 2024

Resumen

El cáncer de mama es uno de los tumores más frecuentes en la población femenina, estableciéndose una prevalencia global en los países occidentales alrededor del 12,5%, es decir, aproximadamente una de cada ocho mujeres padecerá cáncer de mama a lo largo de su vida.

En la lucha por la reducción del impacto de esta enfermedad en la sociedad y en las entidades sanitarias públicas, en los últimos años se han desarrollado diferentes propuestas, siendo la estratificación de la población general uno de dichos métodos. Estos sistemas de estratificación permiten clasificar a la población general en función del riesgo a padecer cáncer de mama de manera individual usando diferentes variables genéticas, fenotípicas o ambientales.

En la presente tesis, partiendo de una cohorte caso – control retrospectivo, basado en población española femenina, se aplicarán diferentes modelos matemáticos utilizando variables genéticas, fenotípicas y combinando ambas. Estos modelos permitirán la estratificación de la población general femenina en función del riesgo individual a padecer cáncer de mama esporádico.

En un primer estudio se evaluaron las variables genéticas como factor discriminante en la cohorte caso – control estudiado. Estas variables son variantes genéticas puntuales denominados SNPs (*single nucleotide polymorphism*). Estas variantes fueron seleccionadas mediante estudios GWAs (*genome-wide association study***)** basados en cohortes casos-control, asociados a cáncer de mama esporádico y usando miles de mujeres de población general Caucásica. Estas alteraciones, estadísticamente significativas y asociadas a fenotipo, tienen escaso efecto funcional, poca penetrancia, son relativamente frecuentes en la población y con asociación individual a un riesgo bajo o moderado a padecer cáncer de mama.

En la presente tesis, se utilizaron inicialmente 121 de estos SNPs descritos anteriormente y combinados mediante modelos *"polygenic risk score"* (PRS) para evaluar su poder discriminante en función del riesgo individual a padecer cáncer de mama esporádico en una cohorte caso-control representativa de la población española

En un segundo estudio se evaluaron las variables fenotípicas como factores discriminantes de la cohorte caso- control estudiado. Los factores de riesgo fenotípicos seleccionados se obtuvieron a partir de bibliografía científica mayoritariamente de población española y son; la densidad mamaria, el estatus de menopausia, la edad de la mujer, la edad de menarquia, edad el primer hijo nacido vivo y los posibles antecedentes familiares.

En un tercer estudio se evaluó la significancia y el poder discriminante utilizando un modelo logístico que combina las variables genéticas agrupadas mediante un PRS, las variables fenotípicas, y la interacción de algunas de estas variables. Se evaluó la posible utilidad del modelo para la clasificación de la cohorte representativa de la población española en función del riesgo individual a padecer cáncer de mama esporádico.

En la presente tesis se presenta un modelo matemático que combina factores genéticos y fenotípicos de riesgo a padecer cáncer de mama esporádico. En este estudio se demuestra la sinergia y la ausencia de colinealidad entre este tipo de variables, permitiendo obtener una capacidad discriminante entre las mujeres que desarrollaron cáncer de mama esporádico durante los siguientes cinco años después de la toma de los datos y mujeres sanas que no lo desarrollaron. Esta capacidad discriminante obtenida mediante estos modelos combinados de los factores de riesgo es mayor que la obtenida usando modelos basados solo en los factores de riesgo individuales. También permite, como prueba de concepto, valorar su posible utilidad en un sistema de cribado mediante la clasificación de la población general en diferentes categorías en función de su riesgo a padecer cáncer de mama esporádico.

Summary

Breast cancer is one of the most common tumours in the female population, with a global worldwide incidence around 12.5%, that is, this means there is a 1 in 8 chance will suffer breast cancer throughout the live.

In the investigation to reduce the impact of this disease on society and public health entities, in last years, different proposals have been developed, where the stratification of the general population being one of these methods. These stratification systems allow the general population to be classified based on the risk of suffering from breast cancer individually using different genetic, phenotypic or environmental variables.

Was started from a retrospective case-control cohort based on the Spanish female population, different mathematical models will be applied using genetic and phenotypic variables and combining both of them. These models will allow the stratification of the general female population based on the individual risk of suffering sporadic breast cancer.

In a first study, genetic variables were evaluated as a discriminating factor. These variables are genetic punctual variants called SNPs (single nucleotide polymorphism). These variants were selected using GWA (genome-wide association study) studies based on case-control cohorts, associated with sporadic breast cancer and using thousands of women from the general Caucasian population. These alterations, statistically significant and associated with phenotype, present modest functional effect, low penetrance, modest risk to development breast cancer and are relatively frequent in the general population.

In this study, 121 of these previously described SNPs were initially used and combined using “polygenic risk score” (PRS) models to evaluate their discriminating capacity based on the individual risk of suffering sporadic breast cancer in a representative case-control cohort of the Spanish population.

In a second part of the study, the phenotypic variables are evaluated as discriminating factors of the case-control cohort studied. The selected phenotypic risk factors were obtained from scientific literature, mainly from the

Spanish population, and are: breast density, menopause status, age of the woman, age of menarche, age of first live birth and family history.

In a third study, the significance and discriminant capacity were evaluated using a logistic model that combines the genetic variables grouped through a PRS score, the phenotypic variables, and the interaction of some of these variables. The possible usefulness of the model was evaluated for the classification of the representative cohort of the Spanish population based on the individual risk of suffering sporadic breast cancer.

In this thesis, a mathematical model is presented where combines genetic and phenotypic risk factors for sporadic breast cancer. This study demonstrates the synergy and absence of collinearity between this type of variables, allowing obtaining a discriminant capacity between women who developed sporadic breast cancer during the following five years after data collection and healthy women. This discriminant capacity obtained using these combined models are greater than that obtained using models based only on the individual risk factors. It also allows, as a proof of concept, to assess its possible usefulness in a screening system by classifying the general population into different categories based on their risk of suffering sporadic breast cancer.

Resum

El càncer de mama és un dels tumors més freqüents en la població femenina, establint-se una prevalença global als països occidentals al voltant del 12,5%, és a dir, aproximadament una de cada huit dones patirà càncer de mama al llarg de la seua vida.

En la lluita per la reducció de l'impacte d'esta malaltia en la societat i en les entitats sanitàries públiques, en els últims anys s'han desenvolupat diferents propostes, sent l'estratificació de la població general un d'estos mètodes. Estos sistemes d'estratificació permeten classificar a la població general en funció del risc a patir càncer de mama de manera individual usant diferents variables genètiques, fenotípiques o ambientals.

En la present tesi, partint d'una cohort cas – control retrospectiu, basat en població espanyola femenina, s'aplicaran diferents models matemàtics utilitzant variables genètiques, fenotípiques i combinant ambdues. Estos models permetran l'estratificació de la població general femenina en funció del risc individual a patir càncer de mama esporàdic.

En un primer estudi es van avaluar les variables genètiques com a factor discriminant en la cohort case – control estudiat. Estes variables són variants genètiques puntuals denominats SNPs (single nucleotide polymorphism). Estes variants van ser seleccionades mitjançant estudis GWAs (genome-wide association study) basats en cohorts casos-control, associats a càncer de mama esporàdic i usant milers de dones de població general Caucàsica. Estes alteracions, estadísticament significatives i associades a fenotip, tenen escàs efecte funcional, poca penetrancia, són relativament freqüents en la població i amb associació individual a un risc baix o moderat a patir càncer de mama.

En la present tesi, es van utilitzar inicialment 121 d'estos SNPs descrits anteriorment i combinats mitjançant models “polygenic risk score” (PRS) per a avaluar el seu poder discriminant en funció del risc individual a patir càncer de mama esporàdic en una cohort cas-control representativa de la població espanyola

En un segon estudi es van avaluar les variables fenotípiques com a factors discriminants de la cohort case- control estudiat. Els factors de risc fenotípics seleccionats es van obtindre a partir de bibliografia científica majoritàriament de població espanyola i són; la densitat mamària, l'estatus de menopausa, l'edat de la dona, l'edat de menarquia, edat el primer fill nascut viu i els possibles antecedents familiars.

En un tercer estudi es va avaluar la significança i el poder discriminant utilitzant un model logístic que combina les variables genètiques agrupades mitjançant un PRS, les variables fenotípiques, i la interacció d'algunes d'estes variables. Es va avaluar la possible utilitat del model per a la classificació de la cohort representativa de la població espanyola en funció del risc individual a patir càncer de mama esporàdic.

En la present tesi es presenta un model matemàtic que combina factors genètics i fenotípics de risc a patir càncer de mama esporàdic. En este estudi es demostra la sinergia i l'absència de colinealidad entre esta mena de variables, permetent obtindre una capacitat discriminant entre les dones que van desenvolupar càncer de mama esporàdic durant els següents cinc anys després de la presa de les dades i dones sanes que no ho van desenvolupar. Esta capacitat discriminant obtinguda mitjançant estos models combinats dels factors de risc és major que l'obtinguda usant models basats només en els factors de risc individuals. També permet, com a prova de concepte, valorar la seua possible utilitat en un sistema de cribratge mitjançant la classificació de la població general en diferents categories en funció del seu risc a patir càncer de mama esporàdic.

Abreviaturas

- SNPs: Single nucleotide polymorphism.
- PRS: Polygenic risk score.
- ADN: Ácido desoxirribonucleico
- GWAs: Genome-wide association study.
- OR: Odds Ratio.
- MAFF: Frecuencia alélica mínima
- DM: Densidad mamaria.
- ROC: Receiver Operating Characteristic.
- AUC: area under curve
- COGS: European Collaborative Oncological Gene Environment Study.
- CNIO: Centro Nacional de Investigación Oncológicas.
- PRS: Polygenic risk score.
- DM: Mammographic density.
- COGS: European Collaborative Oncological Gene Environment Study.
- LRT: Likelihood ratio test.
- CI: Confidence interval.
- SD: Standard deviation.
- SEER: Surveillance, Epidemiology, and End Results.
- *NCI: National Cancer Institute.*
- CEGEN: Centro Nacional de Genotipado.
- COGS*:* Collaborative Oncological Gene-environment Study.
- VEP: Ensembl Variant Effect Predictor
- SEOM: Sociedad española de oncología médica.

Agradecimientos

Agradezco al laboratorio del Centro Nacional de Genotipado (CEGEN) por el soporte técnico. A rebeca Miñambres, Estrella Rubio y Andrea Ceba por la gestión del proyecto durante su estancia en Sistemas Genómicos. A Antonio Cano, del servicio de ginecología y a Ana Lluch del servicio de oncología, ambos del Hospital Clínico Universitario de Valencia por el soporte clínico. A Lola Salas del CIBERESP por el soporte epidemiológico.

A mis directores Dr Javier Benítez y Dr David Moratal por su paciencia, motivación y orientación en todo momento.

Financiación

Este estudio fue financiado por Sistemas Genomicos S.L. Además, contó con el apoyo de dos organizaciones que aportaron fondos para la ejecución del proyecto: IVACE (Instituto Valenciano de competitividad empresarial) y AVI (Agencia Valenciana de Innovación). Números de becas IVACE: IMIDTA/2016/75 e IMIDTA/2018/7; número de subvención del programa AVI; INNTAL21/197002.

Índice

Capítulo 1. Introducción

1.1 Introducción al cáncer de mama

Según los datos de la Asociación Española Contra el Cáncer, en España se diagnostica alrededor de 35.000 casos de cáncer de mama al año. Este valor, según el informe anual de la sociedad española de oncología médica (SEOM) en 2022, representa casi el 30% de todos los tumores diagnosticados en población femenina. La media nacional de incidencias para esta enfermedad se eleva a 50,9 casos cada 100.000 habitantes. A nivel global, se estima que en todo el mundo se diagnostican 1,3 millones de casos al año, con lo que una de cada doce mujeres padecerá un cáncer a lo largo de su vida. Esta prevalencia que representa alrededor del 12,5% en los países occidentales genera una gran problemática, no sólo desde el punto de vista emocional de las pacientes sino también en el económico, asociado al diagnóstico, seguimiento y tratamiento de la enfermedad por las entidades públicas.

El cáncer de mama es una enfermedad compleja y multifactorial donde se ven involucrados diferentes tipos de factores [1], entre los más importantes destacan los factores de riesgos genéticos y fenotípicos. De todos los casos de cáncer de mama se calcula que aproximadamente entre un 5% y un 10% son hereditarios [2]. Este tipo de cáncer aparece cuando se transmiten cambios genéticos de naturaleza patológica dentro de una misma familia. Los principales genes asociados al cáncer de mama hereditario por su importancia son BRCA1 y BRCA2 [2]. El resto son esporádicos o casos con agregación familiar no asociados a mutación en los genes BRCAs. En estos casos el cáncer parece ser debido a otros múltiples factores, tanto de naturaleza genética como de naturaleza fenotípica, que combinados puede generar la alteración del ADN (á*cido desoxirribonucleico*) de una única célula, adquiriendo las capacidades propias de los tumores, como la proliferación y la invasión [1]. Este tipo de cáncer, que se puede desarrollar a lo largo de la vida de la mujer, no tienen factores de riesgo obvios ni antecedentes familiares concordantes con un modelo de cáncer hereditario, por lo tanto, es necesario estudios o revisiones médicas periódicas como mamografías. Este estudio radiológico del seno permite a menudo detectar tumores demasiado pequeños para ser

palpados, permitiendo la posible identificación del tumor en un estadio temprano.

El diagnóstico precoz tiene una influencia positiva en la probabilidad de supervivencia de las pacientes [3]. Diferentes estudios sugieren que la detección temprana junto con un correcto tratamiento podría reducir la mortalidad significativamente a largo plazo [4]. La tasa de supervivencia es la probabilidad de que una mujer con cáncer de mama pueda sobrevivir durante los 5 años posteriores al diagnóstico. La tasa de supervivencias es controladas y contenidas en bases de datos como SEER (*Surveillance, Epidemiology, and End Results*) [5]. Esta base de datos estratifica los tumores de mama:

- Localizado: cuando el cáncer no se ha propagado fuera de la mama
- regional: cuando se disemina fuera del seno a estructuras cercanas o ganglios linfáticos
- distante: cuando se propaga a otras partes del cuerpo, como el hígado, los pulmones o los huesos

Según las estadísticas proporcionadas por *National Cancer Institute* (NCI) (https://www.cancer.gov/) la tasa de supervivencia media a 5 años es del 90,3%. Este dato es considerando todas las estratificaciones mencionadas anteriormente. Estos datos desglosados por los grupos definidos anteriormente presentan diferencias significativas. La tasa de supervivencia a 5 años para las mujeres con cáncer de mama localizado es del 99%, mientras que este estadístico se reduce al 86% para canceres regionales y 29% para distantes.

Según los datos, existe una alta correlación entre la tasa de supervivencia a 5 años y el diagnóstico temprano del cáncer de mama. Por esta correlación, actualmente y para controlar y disminuir el desarrollo e impacto de esta enfermedad, se aplican diferentes estrategias preventivas como los programas de cribado. En estos programas se trata de identificar, en un estadio temprano el cáncer de mama mediante técnicas de imagen (mamografías) o los tratamientos endocrinos para la reducción del riesgo. Por ejemplo, se ha identificado que los programas de cribado basados en densidad mamaria podrían reducir la mortalidad en un rango que va desde el 20% al 25% en mujeres mayores de 50 años [6]. Estos procedimientos presentan desventajas

relacionadas con posibles efectos secundarios, elevados costes económicos de estos programas o basados solo en pautas generales de actuación sin considerar el perfil de riesgo individualizado de cada mujer estudiada [7].

En este sentido, el Consejo de la Unión Europea publicó en el año 2003 unas recomendaciones sobre la detección precoz de cáncer de mama aplicado a mujeres de rango de edad desde 50 hasta los 69 años empleando mamografías cada dos años. Dichas recomendaciones han sido revisadas y confirmadas por la Agencia Internacional de investigación sobre el Cáncer. Esta prueba de cribado adaptada por la mayoría de los países europeos presenta una serie de limitaciones:

- Fasos positivos y/o sobrediagnóstico: Se ha establecido que la tasa de falsos positivos está definida en un rango desde el 20% hasta el 50% [8]. El sobrediagnóstico hace referencia a la detección de tumores asintomáticos que no tienen un impacto en la salud de la mujer
- Falsos negativos:

El estándar actual de atención para la detección del cáncer de mama ofrece a las mujeres de la población objetivo una estrategia homogénea basada únicamente en la edad, en este sentido, esta estrategia puede resultar insuficiente para una eficiente estratificación de la población en cuanto al riesgo de padecer cáncer de mama. En este punto, diferentes autores proponen definir estrategias personalizadas donde la calidad de los hallazgos es muy importante.

La comunidad científica no ha validado la eficacia real de estas estrategias basadas en riesgos personalizados, por lo que se debe ampliar el conocimiento con estudios de cohortes amplias para calcular el impacto y la aceptación en rutina clínica.

1.2 Factores de riesgo del cáncer de mama.

La incidencia o la probabilidad de padecer cáncer de mama viene condicionadas por diversos factores, englobándose en tres tipos; genéticos, ambientales y fenotípicos.

Según la comunidad científica (*breastcancer.org)* los factores de riesgos más reconocidos e importantes son:

- Genero. La mujer tiene asociado un mayor riesgo de cáncer de mama respecto al varón.
- Edad. El riesgo de padecer cáncer de mama aumenta a medida que la edad de la mujer avanza. Casi dos terceras partes los cánceres de mama invasivos ocurren en mujeres mayores de 55 años.
- Densidad mamaria. Es uno de los factores más estudiados. Se ha establecido que los senos densos tienen el doble de probabilidad de padecer cáncer de mama respecto a los no densos [9].
- Lactancia. La lactancia maternal está definida como un factor de protección frente al fenotipo [10].
- Menopausia. Es uno de los factores más documentados e influyentes [11].
- Historia Familiar. Las mujeres cuyos parientes cercanos presentan cáncer de mama tienen un mayor riesgo de padecer la enfermedad. Este riesgo se incrementa sensiblemente si la pariente es de primer grado. Es uno de los factores más estudiados por la comunidad científica [12]. En este punto hay que diferenciarlo del cáncer de mama hereditario, porque a pesar de que esté presente en la familia, incluso en parientes de primer grado, no presenta mutación en los genes de alta o moderada penetrancia o la distribución el fenotipo dentro de la familia no concuerda exactamente con el cáncer hereditario.
- Componente Genético. Este factor de riesgo está asociado tanto a cáncer de mama hereditario como esporádico. El primer caso está asociado a mutaciones en genes de alta o moderada penetrancia mientras que los esporádicos están más asociados a genes de baja susceptibilidad y a modelos poligénicos.
- Historia Personal. Las mujeres que ya han sido diagnosticada con cáncer de mama previamente tienen un riesgo entre 3 y 4 veces mayor que una mujer de la misma condición sin histórico positivo.

- Afecciones mamarias previas. Las afecciones mamarias, aunque sean benignas puede ser un factor de riesgo.
- Etnia. Aunque no existan grandes diferencias, las mujeres de raza caucásica podrían tener un mayor riesgo respecto a mujeres de otra etnia, como la hispánica o la asiática. Por otro lado, las mujeres afroamericanas podrían tener un mayor riesgo a padecer cáncer de mama más agresivo.
- Sobrepeso. Las mujeres con sobrepeso podrían tener una mayor probabilidad de desarrollar el tumor, especialmente después de la menopausia.
- Edad del primer hijo. Se ha observado que los factores reproductivos son factores importantes para la estimación de los riesgos de la mujer. Las mujeres que no han tenido hijos o que han tenido el primer hijo después de los 30 años presentan un mayor riesgo.
- Historia menstrual. Las mujeres con menstruación antes de los 12 años tienen un mayor riego a padecer cáncer de mama.

Además de los factores de riesgos descritos anteriormente, de manera periódica, aparecen otros nuevos asociados al cáncer de mama esporádico en la literatura científica. Sin embargo, normalmente estos factores no presentan un peso importante ya que aún las evidencias científicas no están muy consolidadas bien por la validación en un número elevado de mujeres o bien gestionados por consorcios importantes [13].

Por otro lado, existen otro tipo de factores como los ambientales donde existe un nivel alto de incertidumbre en la precisión de la toma de los datos de las mujeres estudiadas. Esto genera una enrome variabilidad e incertidumbre en el cálculo cuantitativos de estos riesgos. Algunos ejemplos de estos factores son la exposición a químicos, la presencia de bajos niveles de vitamina D o el tabaquismo [14], [15].

El componente genético es uno de los factores más estudiados cuyo riesgo se describe en la literatura científica. Se estima que entre el 5 % y el 10 % de los

cánceres de mama son hereditarios, es decir, causados por mutaciones germinales en genes importantes para el fenotipo y transmitidos de padres a hijos. Por ejemplo, se estima que el riesgo de enfermedad en mujeres con antecedentes familiares es el doble respecto al de mujeres sin historia familiar [16]. La mayoría de los cánceres de mama hereditarios están ocasionados por mutaciones en los genes BRCA1 y BRCA2. Estos genes son considerados como genes supresores de tumores cuya función es reparar el daño del DNA y controlar el ciclo celular [17]. Mutaciones en este tipo de genes aumentan el riesgo a padecer cáncer ya que se ve alterada la capacidad de la célula de reparar el daño producido al DNA [18], pero no se conoce aún en profundidad el motivo por el cual las alteraciones en estos genes generan cáncer de mama y ovario de manera preferente. En los últimos años, juntos con los genes BRCA1 y BRCA2 se han ido identificando nuevos genes funcionalmente relacionados con la ruta de BRCA cuya mutación también genera un factor de riesgo importante a padecer cáncer de mama [19]. Algunos de estos genes cuya relevancia clínica es mayor son ATM, PALB2 y CHECK2 [20], tal y como se puede observar en la **Figura1** [19], son genes reguladores de procesos asociados con los mecanismos patológicos del cáncer.

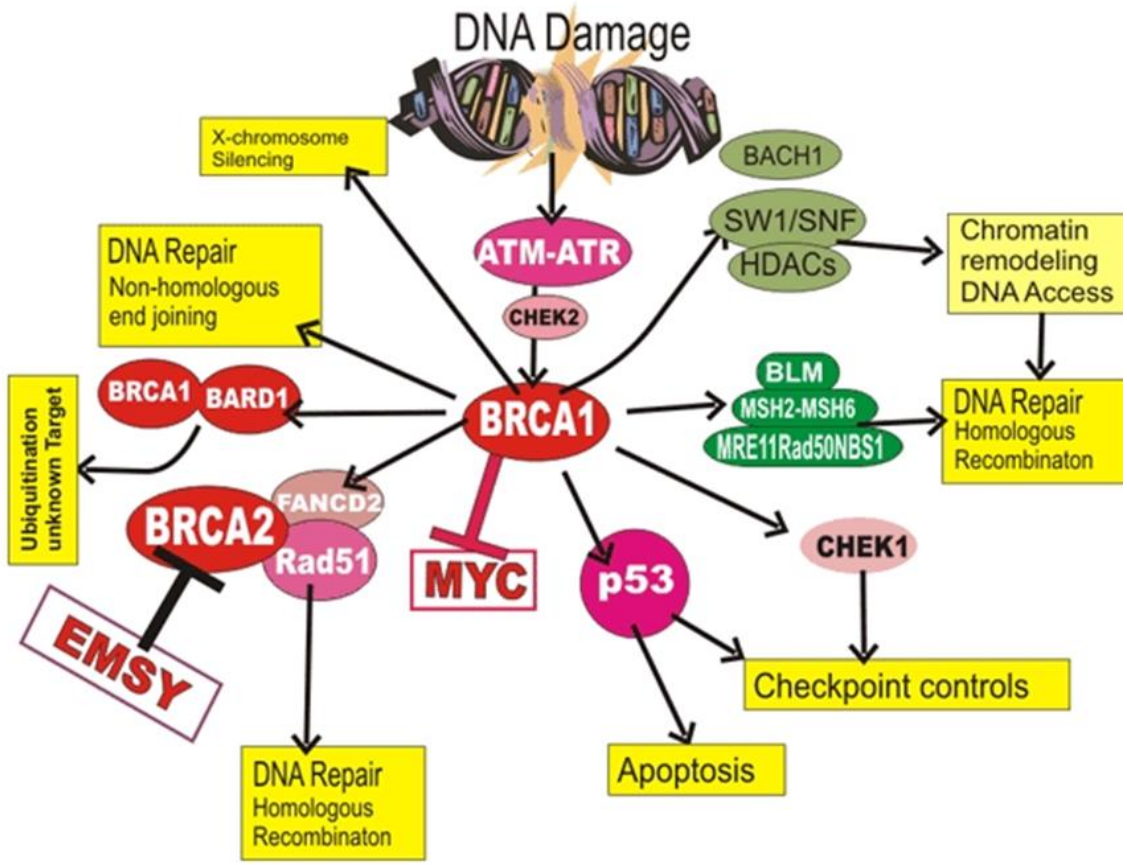

Figura1. Ruta de los genes BRCAs [19].

Por ejemplo, mutaciones patológicas en el gen *ATM* (*ataxia-telangiectasia mutated*) está descrito que puede aumentar el riesgo a padecer cáncer de mama en una media de 2,8 [21] y podría estar mutado en torno al 8,5 % en los cánceres descritos en edad temprana [22]. Otro gen interesante es el CHECK2, donde se describe que podría estar mutado en un 5% de las mujeres con una fuerte historia familiar de cáncer de mama y ovario y sin embargo fueron negativos en mutaciones BRCA1 y BRCA2 [23]. La importancia de estas nuevas dianas ha crecido en los últimos años y se han utilizado en diferentes métodos de estratificación basados en el componente genético [24], [25].

Según lo comentado anteriormente, entre el 5 y el 10 % de los cánceres de mama descritos son hereditarios, por lo que el 90% restante se denominan cánceres esporádicos, sin mutación descrita que explique el mayor riesgo de la mujer. En este tipo de cáncer también está descrito que existen factores de riesgo genético, pero en este caso, normalmente están asociados a modelos poligénicos [26]. El modelo poligénico puede ser descrito como un valor que resume o aglutina el efecto acumulativo de múltiples variantes genéticas de

prevalencia alta o moderada en la población general y que presentan un riesgo relativo bajo pero que estadísticamente están asociado al fenotipo [27], [28], [29], tal y como puede observarse en la **Figura2**. Este modelo se ha estudiado ampliamente y también se asocia a otras enfermedades complejas [30]. La identificación y el cálculo del riesgo de los SNPs son obtenidos mediante estudios masivos basados en asociación genómica (GWAS).

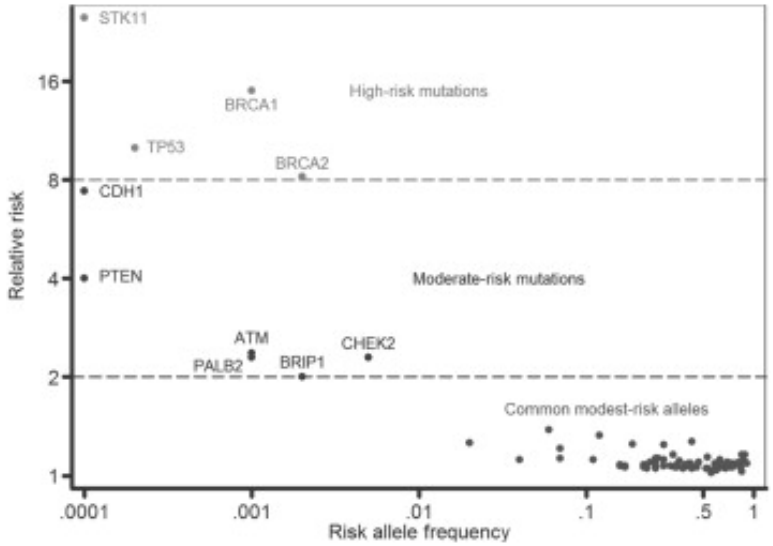

Figura2: Relación entre frecuencia alélica dentro de la población y el riesgo relativo de los SNPs presentes en los diferentes genes dianas asociados a cáncer de mama [6].

La agrupación de los riesgos de estos SNPs en un único valor denominado PRS ha sido validada como una variable informativa importante del riesgo de padecer cáncer de mama esporádico en diferentes poblaciones [31], [32]. En este sentido, estudios basados en PRS compuesto de 76 SNPs estadísticamente significativos y basados en la población española demuestran que las mujeres clasificadas como de bajo riesgo presentan la mitad de probabilidad de padecer cáncer de mama respecto a la población general. Por otro lado, las mujeres clasificadas como de alto riesgo presentan 2,5 veces más probabilidad de padecer cáncer de mama respecto a la población general [33].

En los últimos años, la comunidad científica está realizando cada vez más experimentos de este tipo, permitiendo la identificación de mayor número de SNPs significativos asociados a cáncer de mama esporádico y generando valores PRS más precisos para su uso a través de nuevos modelos. En el mismo sentido, cada vez son más las poblaciones estudiadas con estos métodos, identificando las posibles similitudes y diferencias, acerca el uso de estas estrategias en una rutina clínica [34], [35].

Además de los factores genéticos existen diferentes factores fenotípicos que son descritos y consolidados por la comunidad científica como factores de riesgo de cáncer de mama esporádico. Uno de los más importantes es la edad [36]. Según la literatura científica, el 80% de los cánceres esporádicos se diagnostican en mujeres mayores de 50 años, sobre todo en los rangos de edad entre los 50 y los 69 años. Este intervalo de edad podría ser importante ya que, como media, la mujer entra en menopausia en torno a los 55 años [37]. Por otro lado, la incidencia en mujeres menores de 40 años es significativamente más baja respecto a mujeres de otro rango de edad [38] [39].

Otro grupo importante de factores fenotípicos de riesgo a padecer cáncer de mama esporádico son los factores reproductivos, especialmente la edad de la menopausia, la menarquia y la edad del primer hijo nacido con vida. Estos factores miden el número de años que la mujer ha estado expuesto a estrógenos a lo largo de su vida [40].

Otro de los factores fenotípicos cuya asociación a cáncer de mama es ampliamente estudiado es la densidad mamaria (DM). Normalmente, la densidad mamaria se clasifica en 4 categorías que van desde el tejido prácticamente graso hasta el tejido extremadamente denso y con muy poca grasa. El radiólogo decide de manera relativamente subjetiva cuál de las 4 categorías se asocia cada seno. Esta clasificación se denomina BI-RADS [41] .Las mujeres que tienen tejido mamario denso tienen un mayor riesgo de padecer cáncer de mama en comparación con las mujeres con tejido mamario menos denso.

Otro factor de riesgo fenotípico considerado relevante son las biopsias de senos hechas previamente por la mujer.

1.3 Otros modelos de predicción del riesgo individual de padecer cáncer de mama

El cribado de cáncer de mama se incluye en la cartera de servicios comunes del Sistema Nacional de Salud desde el 2006 [42]. Actualmente todas las comunidades autónomas (CCAA) cuentan con un programa de cribado de cáncer de mama (PCCM) coordinado a través de la Red de Programas de Cribado de Cáncer y por el Ministerio de Sanidad [43]. En el 2008 el parlamento europeo lanza una iniciativa para el desarrollo de nuevas guías sobre el cribado y diagnóstico del cáncer de mama a nivel europeo. Esta iniciativa, basada en un comité de expertos multidisciplinar apoyado por revisiones sistemáticas de evidencias científicas, proporciona seguridad y robustez a las guías propuestas. Actualmente, las recomendaciones incluyen usar la mamografía de las mujeres en los programas de cribado y, para mujeres con alta densidad mamográfica, realizar una mamografía digital [44]. A pesar de que la mamografía cuenta con altos niveles de sensibilidad (70-96%) y especificidad (90-95%) [45], existen diferentes problemas asociados que ponen de manifiesto las limitaciones del uso exclusivo de la mamografía como variable discriminatoria del riesgo a padecer cáncer de mama en los sistemas de cribado [46]. En este punto, en los últimos años, se han desarrollado diferentes modelos utilizando otras fuentes de origen genético y fenotípico como variables de predicción y estratificación, entre los más importantes están BOADICEA, modelo multifactorial basado en factores genéticos y fenotípicos o el modelo de Gail, modelo que incluyen la etnia de las mujeres [47].

Capítulo2. Objetivos:

2.1 Objetivos generales

El principal objetivo de la presente tesis es el diseño, desarrollo y evaluación de un modelo matemático que permita estratificar a la población general española femenina en función del riesgo individual a padecer cáncer de mama esporádico (BRCA1 /2 negativo). Dicho modelo está basado en factores de riesgos genéticos, fenotípicos y la interacción entre ambos.

Para alcanzar este objetivo principal se definen los siguientes objetivos específicos

1. Evaluación del poder discriminante del estimador PRS basado en 121 SNPs significativos aplicado a una cohorte caso –control de población general española femenina (BRCA1/BRCA2 negativo).
2. Evaluación del poder discriminante de los diferentes factores de riesgo fenotípicos definidos en el estudio aplicado a una cohorte caso –control de población general española femenina (BRCA1/BRCA2 negativo.). Los factores fenotípicos seleccionados fueron basados en la literatura científica mayoritariamente de población española, y son la edad de la mujer, antecedentes familiares, densidad mamaria, edad de la menopausia, primer hijo y menarquia y la presencia de biopsia.
3. Estudio de las posibles interacciones entre variables genéticas y fenotípicas y su influencia en el aumento de la capacidad discriminante del modelo.
4. Definición y evaluación del modelo, y evaluar su posible incorporación a la rutina de cribado en población general española femenina.

2.2 Contribución al conocimiento

Esta tesis presenta la definición de un novedoso modelo donde se integra en una misma estructura factores de riesgos de diferente naturaleza como son los fenotípicos y los genéticos. Adicionalmente, el trabajo estudia las posibles interacciones entre variables que permite obtener valores de discriminación mayor, entre otras, la interacción entre la edad y la densidad mamaria y la edad con la menopausia.

Adicionalmente, tanto la definición del método como su evaluación están basadas en una población pobremente estudiada por la comunidad científica como es la población general española femenina, definiendo población general como aquella que no tiene ninguna mutación portadora en los genes de alta penetrancia para el cáncer de mama, es decir, para los genes BRCA1 y BRCA2. Esta cohorte permite una evaluación realista de la posible capacidad del método descrito y su posible interés en la incorporación en diferentes programas de cribado dentro de la rutina clínica.

Capítulo 3. Materiales y Métodos:

3.1 Descripción de la cohorte

El estudio descrito en la presente tesis fue revisado y aprobado por el Comité Ético de Investigación Clínica (CEIC) del Hospital Clínico Universitario de Valencia y fue acorde con la declaración de Helsinki [48]. La cohorte se basa en un estudio caso/control, donde los casos son mujeres con diagnóstico de cáncer de mama en el momento de la toma de los datos y los controles son mujeres que no desarrollan cáncer de mama hasta al menos 5 años después de la toma de los datos.

El reclutamiento fue llevado a cabo desde enero del 2017 hasta diciembre del 2018 procedentes de dos localizaciones, el Hospital Clínico Universitario de Valencia y el programa de cribado de la Comunidad Valenciana.

Un total de 867 mujeres sanas, denominadas controles, y 640 mujeres con diagnóstico de cáncer de mama, denominadas casos fueron reclutadas. Los rangos de edad fueron desde los 30 hasta los 70 años. Las mujeres que presentaron datos genéticos o fenotípicos incompletos fueron eliminadas de la cohorte. Dentro de los casos, el 45% fueron tumores tipo Luminal A, 20% Luminal B, 20% Her-2 positivos y 15% como triple negativo.

La información clínica de las mujeres se obtuvo en el reclutamiento. Esta información fue la historia familiar respecto al cáncer de mama, edad de la mujer, edad del primer hijo (si los tuviera), edad de la menarquia, edad de la menopausia y la densidad mamográfica. Respecto a la densidad mamaria, se practicó en todas las mujeres una mamografía bilateral en doble proyección craneocaudal y oblicuo medio lateral llevado a cabo por radiólogos con más de 10 años de experiencia. Dichos radiólogos utilizaron el formato de imagen médica DICOM (General Electric GIMD) como visualizador y para la clasificación utilizaron la escala semicuantitativa Boyd's [49].

El tamaño muestral se calculó con un nivel de confianza del 95 % (prueba de dos colas), un poder estadístico del 80% (valor beta), una relación de casos y

controles de 1:3 y una prevalencia inicial de cáncer de mama del 12%. Según estos datos, el número total de mujeres necesarias para que los resultados fueran estadísticamente significativos fue 1138. Este valor fue similar a la cohorte de casos y controles usados en esta tesis, que finalmente fue de 1126 mujeres.

3.2 Procesamiento de muestras y genotipado

El genotipado de las mujeres que forman parte de la cohorte estudiada se basó en una muestra de sangre obtenida en el momento de la toma de los datos.

Posteriormente, el ADN se extrajo de manera automática utilizando un extractor de ADN (MagNA Pure, Roche, Mannheim, Alemania). Posteriormente se cuantificó el ADN mediante la técnica fluorimétrica Picogreen y finalmente fueron genotipificadas en el Centro Nacional de Genotipado (CEGEN) del CNIO, en Madrid, usando un array de genotipificación de Openarray que contiene 121 SNPs asociados a cáncer de mama.

La selección de los primeros 76 SNPs se basó en estudios GWAs, casos-control asociados a cáncer de mama esporádico procedentes del estudio *Collaborative Oncological Gene-environment Study* (COGS) [50]. Este consorcio es una propuesta europea cuyo principal objetivo es identificar a personas con mayor riesgo de padecer diferentes tipos de cánceres, entre ellos, el cáncer de mama, ovario o próstata. Para ello toma en consideración variables genéticas.

Los 76 primeros SNPs seleccionados se habían encontrado asociados a cáncer de mama esporádico en dichos estudios con un valor de pValue $< 5 \times 10^{-8}$, con una frecuencia alélica mínima (MAF) > 1% y unos OR > 1,05 o < 0,95.

Posteriormente se llevó a cabo una ampliación de los SNPS significativamente asociados a cáncer de mama esporádico desde las 76 iniciales hasta los 121. Estos SNPs adicionales fueron obtenidos del proyecto OncoArray [51] con las mismas características técnicas y poblacionales descritas anteriormente, y añadiendo factores fenotípicos y ambientales. Todos los SNPs con sus OR y MAF están descritos en la **tabla1**. Como control de calidad se estableció que aquellas muestras que presentaran una tasa de éxito en el genotipado inferior

al 95%, es decir, que el genotipado tenga más de 6 SNPs clasificado como indefinido serían excluidas del estudio.

En los análisis posteriores al tratarse exclusivamente de muestras femeninas y por tanto presentar todos los SNPs diploidía, el genotipo de cada variante se representó por el número de copias del alelo de riesgo 0, 1, 2, es decir, como ausente, heterocigoto y homocigoto para la variante, respectivamente.

SNP	MAF España	OR	Alelo May	Alelo Min
rs11196174	0,332	1,13	G	C
rs11196175	0,322	1,12	C	G
rs2290854	0,318	1,13	A	T
rs6682208	0,341	1,12	T	A
rs2736108	0,294	0,92	T	A
rs2420946	0,402	1,27	T	A
rs27633	0,449	1,14	G	C
rs16886113	0,084	1,24	G	C
rs9348512	0,336	0,85	A	T
rs2253407	0,472	0,92	T	A
rs4733664	0,444	1,10	T	A
rs10965163	0,089	0,84	T	A
rs183211	0,299	1,25	A	T
rs9303542	0,262	1,12	G	C
rs344008	0,047	1,21	A	T
rs4691139	0,472	1,20	G	C
rs10088218	0,107	0,86	A	T
rs3814113	0,308	0,77	C	G
rs67397200	0,266	1,16	G	C
rs184577	0,276	0,85	A	T
rs13039229	0,168	0,90	C	G
rs619373	0,439	1,30	A	C
rs13387042	0,463	1,20	A	C
rs10941679	0,248	1,15	G	C
rs2046210	0,397	1,11	A	T
rs8170	0,140	1,26	A	T
rs10771399	0,145	0,79	G	C
rs10069690	0,276	1,06	T	A
rs4973768	0,449	1,11	T	A
rs614367	0,127	1,15	T	A
rs3803662	0,252	1,20	A	T
rs1292011	0,416	0,92	G	C
rs865686	0,388	0,89	G	C
rs10995190	0,192	0,86	A	T
rs2981579	0,397	1,43	A	T
rs11075995	0,178	1,03	T	A
rs909116	0,486	1,17	T	G
rs11249433	0,500	1,11	G	C
rs58847541	0,159	1,08	A	T
rs889312	0,248	1,13	C	G
rs6001930	0,103	1,12	C	G

rs941764	0,383	1,03	G	C
rs12443621	0,495	1,11	G	C
rs6472903	0,145	0,94	G	C
rs12710696	0,308	1,03	T	A
rs9790879	0,388	1,10	C	G
rs17356907	0,285	0,91	G	C
rs6828523	0,084	0,91	A	T
rs1045485	0,150	0,88	C	G
rs8009944	0,262	0,88	C	G
rs10931936	0,395	0,88	T	A
rs2363956	0,463	1,19	T	A
rs4849887	0,150	0,91	T	A
rs7072776	0,308	1,05	A	T
rs11552449	0,210	1,04	T	A
rs6762644	0,313	1,05	G	C
rs17631303	0,234	1,27	G	C
rs3760982	0,495	1,05	A	T
rs9790517	0,192	1,04	T	A
rs10472076	0,360	1,03	C	G
rs204247	0,444	1,04	G	C
rs1436904	0,411	0,95	G	C
rs11199914	0,299	0,96	T	A
rs11820646	0,402	0,96	T	A
rs527616	0,369	0,97	C	G
rs3903072	0,435	0,97	T	A
rs6504950	0,252	0,95	A	T
rs1550623	0,187	0,94	G	C
rs6815814	0,444	1,06	C	G
rs720475	0,304	0,96	A	T
rs10995201	0,187	0,90	G	C
rs17817449	0,374	0,95	G	C
rs1353747	0,075	0,96	G	C
rs2304277	0,182	1,00	A	T
rs1466785	0,416	1,00	T	A
rs616488	0,322	0,94	G	C
rs72755295	0,028	1,15	G	C
rs4442975	0,481	0,89	T	A
rs16857609	0,248	1,06	T	A
rs1432679	0,491	1,08	C	G
rs3757322	0,369	1,08	G	C
rs9397437	0,089	1,17	A	T
rs2747652	0,477	0,94	T	A
rs11977670	0,355	1,06	A	T
rs9693444	0,364	1,06	A	T
rs13365225	0,145	0,91	G	C
rs2943559	0,093	1,10	G	C
rs13281615	0,486	1,11	G	C
rs10816625	0,103	1,11	G	C
rs13294895	0,178	1,06	T	A
rs11814448	0,028	1,12	C	G
rs704010	0,453	1,07	T	A
rs554219	-	1,21	G	C

rs75915166	0,061	1,28	A	T
rs12422552	0,313	1,06	C	G
rs7297051	0,234	0,89	T	A
rs11571833	0,014	1,35	T	A
rs2236007	0,252	0,93	A	T
rs2588809	0,243	1,06	T	A
rs4784227	0,238	1,23	T	A
rs13329835	0,229	1,07	G	C
rs6507583	0,084	0,92	G	C
rs4808801	0,336	0,93	G	C
rs2823093	0,243	0,94	A	T
rs10474352	0,159	0,94	T	A
rs2290203	0,215	0,94	A	T
rs2992756	0,481	1,06	T	A
rs7529522	0,313	1,06	C	G
rs13066793	0,107	0,94	G	C
rs9833888	0,215	1,06	T	A
rs72749841	0,243	0,93	C	G
rs1895062	0,481	0,94	G	C
rs72826962	0,014	1,20	T	A
rs78269692	0,056	1,09	C	G
rs16991615	0,079	1,10	A	T
rs73161324	0,033	1,06	T	A
rs676256	0,397	0,91	C	G
rs113577745	0,093	1,08	G	C
rs58058861	0,117	1,06	A	T
rs45631563	0,042	0,81	T	A

Tabla1: SNPs utilizados para el cálculo de PRS. Cada SNPs tienen asociado una frecuencia en la población española, un OR o riesgo relativo y un alelo mayoritario y minoritario.

3.3 Cálculo del "PolyGenic Risk Score", PRS

El valor PRS se basó en el efecto combinado de los 121 SNPs estadísticamente asociados a cáncer de mama esporádico descritos en la **tabla1**. Esta estrategia considera un efecto independiente de cada SNPs ignorando un posible modelo multiplicativo [52]. El valor de PRS utilizado en el estudio tiene la siguiente formula:

$$PRS = \beta 1x1 + \beta 2x2 + \ldots + \beta kxk$$

Donde los valores ***xk*** corresponden a los alelos de riesgos (0, 1 o 2) según la cigosidad de cada SNPs. Los valores ***βk*** son los ORs de los alelos de riesgo obtenidos de su asociación a cáncer de mama. Esta estrategia ha sido utilizada en otros estudios similares. El valor obtenido es normalizado usando la mediana del valor PRS de las mujeres sanas de la cohorte.

3.4 Estudio Estadístico

En el análisis fenotípico, las categorías fueron transformadas en variables cuantitativas usando los OR descritos en el estudio de Pollan [53], excepto la historia familiar cuyos OR fueron basados en el estudio de Pharoa [12]. Además, la edad de la mujer (el diagnóstico en casos y la de toma de los datos en controles) se agruparon en 5 años, como se ha propuesto en otros estudios similares [54].

Para el estudio univariable, se aplicó una regresión logística para cada factor de riesgo ajustada por la edad y en el caso de la variable genética PRS fue ajustada también por las 5 primeras componentes principales. Los coeficientes de los modelos fueron estandarizados mediante la librearía de R reghelper [55]. La interacción entre términos fue evaluada mediante el test LTR. Todos los análisis fueron bajo la premisa de dos colas y con un umbral de pValue de 0,05.

Para confirmar la independencia entre las variables genéticas y fenotípicas se empleó la correlación de Spearman usando solo los controles de la cohorte estudiada.

Para el estudio multivariable se realizó una regresión logística utilizando todas las variables estadísticamente significativas obtenidas en el apartado anterior, incluida las interacciones entre términos. La historia familiar y la edad de la menarquia también se incorporaron al análisis, aunque no fueron significativas, por la fuerte relación con el fenotipo descrito tras una exhaustiva revisión bibliográfica. La significación estadística del modelo multivariable fue evaluada mediante Wald Test [56].

La precisión de los modelos logísticos fue evaluada mediante el test de ajuste de bondad *Homer-Lemeshow* usando deciles [57]. Para la evaluación de la capacidad discriminante entre casos y controles de los diferentes modelos se evaluó la curva ROC [58]. Este cálculo se realizó usando la librería de R pROC [59]. Para evitar la sobreestimación de los modelos se calcularon los intervalos de confianza al 95% (CI) mediante una estrategia de validación cruzada [60]. Este paso se realizó mediante el uso de 1000 permutaciones con una selección aleatoria del 90% de la cohorte para la formación del modelo y el 10% restante para su evaluación como cohorte de testeo.

Finalmente, los valores de riesgos obtenidos del modelo multivariable para cada mujer de la cohorte caso-control fue ordenado por deciles y se calculó el OR de los deciles extremos frente al rango central del 40-60%.

3.5 Interpretación biológica y funcional de los SNPs significativos

Los SNPs seleccionados como estadísticamente significativos para el estudio de PRS fueron funcionalmente analizados mediante el método "*Ensembl Variant Effect Predictor*" (VEP), versión 103 [61]. Mediante este método se obtiene información funcional asociada a las variantes. Entre los puntos más importantes está el gen afectado, la consecuencia, el impacto en función de las diferentes isoformas que conforman el gen o el dominio funcional de la proteína que puede verse afectada. En este último punto, la información proporcionada fue basada en las bases de datos **"PHATHER"** [62] y **"PFAM"** [63]**.**

Otras fuentes de información funcional utilizadas en este análisis fueron las frecuencias descritas en diferentes poblacionales obtenidas del consorcio **gnomeAD** [64]**,** la relevancia clínica utilizando **ClinVar** [65], predictores de patología como **SIFT** [66] y **POLYPHEN** [67]. Además del posible efecto de las variantes en la estabilidad de la proteína, también se evaluó la posible afectación en zonas regulatorias, entre otras, los sitios de unión de factores de transcripción.

Las funciones biológicas de los genes afectados fueron analizadas mediante la ontología **Gene Ontology** [68] y las rutas metabólicas a la cual pertenecen

fueron analizadas mediante las bases de datos **KEGG** [69] y **Reactome** [70]. El estudio de enriquecimiento funcional fuera realizado mediante **TopFunc** [71] y el estudio de redes mediante **Cytoscape** [72].

Capítulo 4. Resultados:

4.1 Descripción de la cohorte y edad como posible variable cofundadora

La cohorte final estudiada la componen 1126 mujeres; 463 fueron casos y 663 se clasificaron como controles según las especificaciones definidas en la sección de métodos. La cohorte caso control se define con una proporción 1:1,4 aproximadamente.

La edad es uno de los factores de riesgo más importante para la estratificación de la población, por eso esta variable puede desviarse en la cohorte entre las mujeres clasificadas como controles y como casos, pudiendo tener un efecto cofundador. La cohorte presenta una distribución de edades que van desde 25 hasta 78 años teniendo una mediana de 51 años. Debido a que las edades se van a agrupar en diferentes categorías (en lustros), las colas 1% y 99% percentil de edades son eliminadas del estudio. Este método de agrupamiento tiene como finalidad definir grupos cuya representatividad sea homogénea. Este caso, las mujeres menores de 25 y mayores de 70 años fueros eliminados del estudio. La distribución acumulativa y los umbrales se observan en la **Figura 3**.

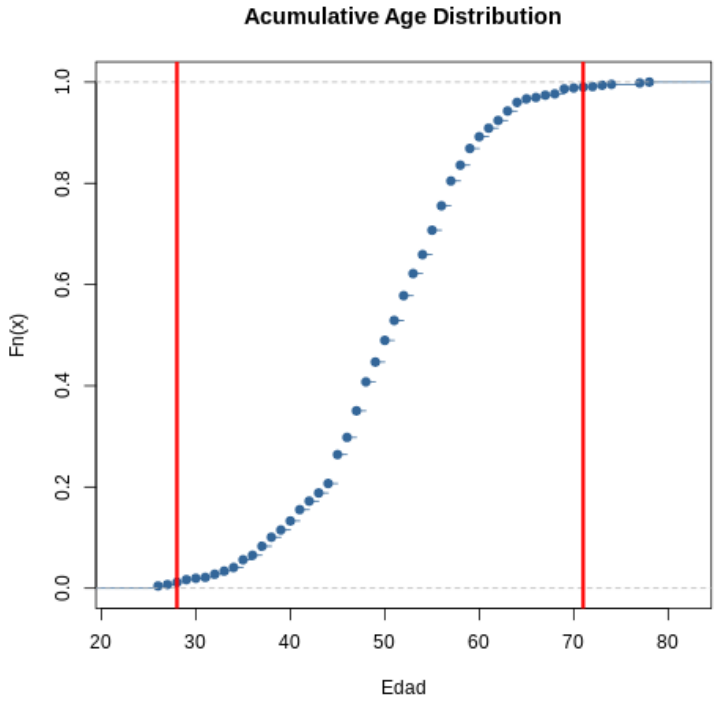

Figura3. Distribución de edades en años en la cohorte. Las líneas verticales marcan los umbrales de edad al 1% y al 99% que serán eliminados del estudio.

Para verificar que no existe desplazamiento respecto la edad en la cohorte estudiada, tanto en controles como en casos, se realizó un test de Wilcoxon utilizando las dos distribuciones numéricas de edades. El **pValue** obtenido fue 0,27, rechazándose la hipótesis alternativa y aceptando la hipótesis nula de que ambas distribuciones de edades, tanto en controles como en casos, presentan la misma distribución (ver **Figura 4**).

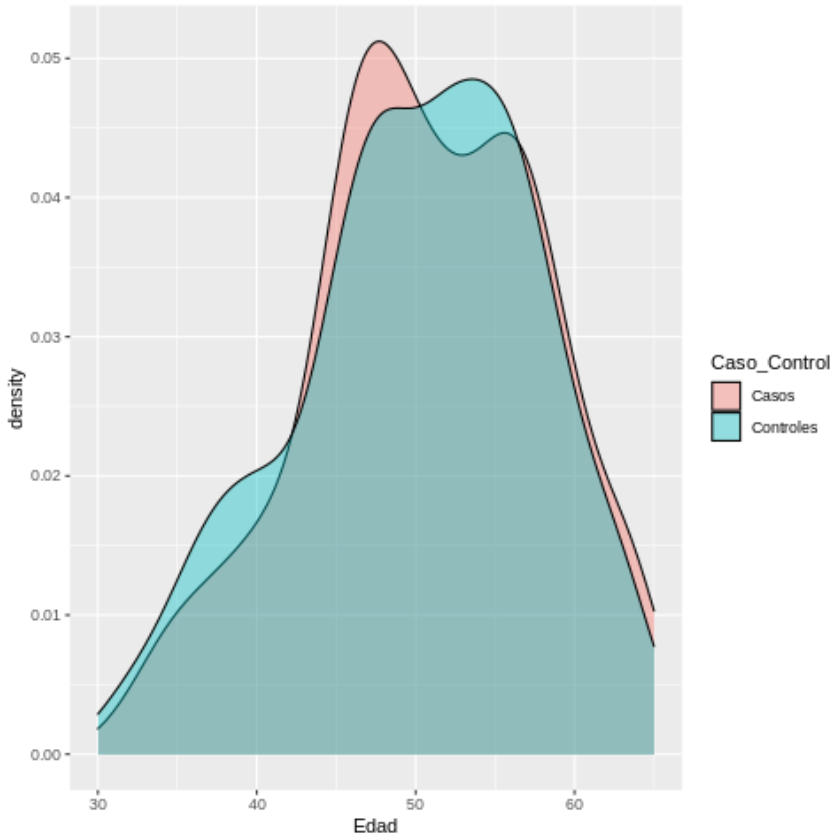

Figura4. Distribuciones de edades en años en la cohorte para Casos y Controles. Mediante un test de Wilcoxon no se observan diferencias estadísticas.

Después, las edades de las mujeres se clasificaron en lustros, de 25-30 años a 65-70. Las categorías adquieren un valor numérico creciente desde el valor 0 para las mujeres en rangos de edad entre 25-30 años, hasta el valor 8 en rangos 65-70. A partir de este punto, las edades trabajadas en el modelo se basarán es esta variable numérica.

4.2 Descripción genotípica de la cohorte de validación.

Los resultados sobre la frecuencia de los SNPS analizados están descritos en la **tabla 2**. En dicha tabla se puede observar la frecuencia del alelo mayoritario (o más frecuente) dentro de la cohorte estudiada y el equilibrio de *Hardy-Weinberg**.*** Este estimador nos indica que la composición genética de una población permanece en equilibrio mientras no actúe la selección natural ni ningún otro factor y no se produzca ninguna mutación. Si esto se cumple, las frecuencias de las diferentes variantes estarán en un valor de equilibrio. Normalmente, en los estudios GWAs, las variantes que no siguen esta ley (pValue <=0,001) son eliminadas del estudio. En esta cohorte, ningún SNPs fue eliminado debido a esta ley.

Adicionalmente se valoró el porcentaje de error del genotipado para los diferentes alelos en la cohorte. Se observó que 10 SNPs dieron errores en todas las muestras analizadas y fueron eliminados del estudio.

ID	Alelos	Mayor_Freq	HWE	%Error
rs10816625	A/G	92,3	0,202955	0,8
rs10931936	C/T	72,3	0,231427	1
rs10941679	A/G	78,9	0,718949	0,7
rs10965163	C/T	91	0,140429	0,8
rs10995190	G/A	86,8	0,602154	0,5
rs10995201	A/G	86,3	0,899537	0,6
rs11075995	NA	NA	NA	100
rs11196174	A/G	67	0,588263	1,1
rs11196175	T/C	68,2	0,534998	1
rs11199914	C/T	69,6	0,066396	0,7
rs11249433	A/G	56,9	0,465251	0,3
rs113577745	C/G	88,4	0	1
rs11552449	C/T	83,8	1	0,7
rs11571833	A/T	99	1	1,1
rs11814448	A/C	97,2	0,205525	0,5
rs11820646	C/T	63,1	0,519719	1,4
rs11977670	G/A	60,9	1	0,5
rs12422552	G/C	68	1	0,6
rs12443621	G/A	51,9	0,904724	0,8
rs12710696	C/T	64,9	0,049295	0,5
rs1292011	A/G	64,5	0,153844	8,3
rs13039229	A/C	81,1	0,062225	2,1
rs13066793	A/G	91,3	0,571431	0,9
rs13281615	A/G	52,5	0,308789	0,4
rs13294895	C/T	82,5	0,408708	0,5
rs13329835	A/G	74,8	0,936793	0,6
rs13365225	A/G	81,1	0,120338	0,1
rs13387042	A/G	56,8	0,426456	1,4
rs1353747	T/G	91,3	0,022829	0,4
rs1432679	C/T	50,4	0,082697	0,7
rs1436904	NA	NA	NA	100
rs1466785	C/T	58,1	0,295479	1,1
rs1550623	A/G	85,6	0,627699	1,2
rs16857609	C/T	73,9	0,351917	0,6
rs16886113	T/G	91,6	0,115621	1
rs16991615	G/A	90,8	1	0,5
rs17356907	A/G	71,4	0,826796	0,2
rs17631303	NA	NA	NA	100
rs17817449	T/G	59,5	0,710082	0,4
rs183211	NA	NA	NA	100
rs184577	G/A	72,2	0,501828	0,8
rs1895062	A/G	55,6	0,089941	0,6
rs204247	A/G	54	0,904172	0,7
rs2046210	G/A	60,9	0,130299	1,8

rs2236007	G/A	77,7	0,434699	2
rs2253407	G/T	52,8	0,18487	1,5
rs2290203	G/A	80,1	0,091301	0,4
rs2290854	G/A	64,4	0,103045	0,8
rs2304277	G/A	80,1	0,512669	0,4
rs2363956	T/G	51,4	0,168079	0,9
rs2420946	C/T	56,8	0,329483	0,8
rs2588809	C/T	76,6	0,505315	0,5
rs2736108	C/T	73,2	0,170136	0,7
rs2747652	C/T	51,2	0,403077	0,5
rs27633	T/G	59	1	2,3
rs2823093	G/A	77,7	0,931144	0,6
rs2943559	A/G	90,2	0,017751	0,7
rs2981579	G/A	56,1	0,145255	0,5
rs2992756	C/T	50,6	0,629625	2,2
rs344008	NA	NA	NA	100
rs3757322	T/G	62,3	0,277976	1,1
rs3760982	G/A	53,6	0,0629	0,4
rs3803662	G/A	67	0,31122	0,3
rs3814113	NA	NA	NA	100
rs3903072	G/T	53,4	0,763551	1,1
rs4442975	G/T	54,3	1	0,6
rs45631563	A/T	94,1	0,424291	0,6
rs4691139	NA	NA	NA	100
rs4733664	C/T	51,1	0,857504	0,8
rs4784227	C/T	71	0,109957	0,7
rs4808801	A/G	66,3	0,0268	1,2
rs4849887	C/T	87,6	0,784784	0,8
rs4973768	T/C	53	0,163862	2,3
rs527616	G/C	61,5	0,613379	0,6
rs554219	C/G	89,8	0,512606	1
rs58058861	G/A	83,6	0,299998	10,4
rs58847541	G/A	88,3	0,248154	0,4
rs6001930	T/C	90,1	0,044223	0,9
rs614367	C/T	88,4	0,186909	0,5
rs616488	A/G	71,3	0,37584	2,8
rs619373	G/A	89	0,064838	0,1
rs6472903	T/G	87,2	0,593968	0,4
rs6504950	G/A	74	0,074293	0,4
rs6507583	A/G	91,2	0,452048	1,6
rs6682208	C/T	61,2	0,115162	0,7
rs67397200	NA	NA	NA	100
rs676256	NA	NA	NA	100
rs6762644	A/G	66,1	0,894228	0,3
rs6815814	A/C	60,2	0,150061	1,2

rs6828523	C/A	88,8	0,366973	0,4
rs704010	C/T	58,7	1	0,2
rs7072776	G/A	70,5	0,666493	0,5
rs720475	G/A	72,3	0,822017	1,5
rs72749841	T/C	74,9	0,025742	0,4
rs72755295	A/G	97,6	1	0,2
rs72826962	C/T	98,6	1	1,2
rs7297051	C/T	77,4	0,087652	0,5
rs73161324	C/T	95,1	0,747372	1
rs7529522	T/C	73,3	0,126049	0,5
rs75915166	C/A	95,4	0,509239	0,1
rs78269692	T/C	97,4	0,175233	0,3
rs8009944	A/C	72,7	0,019443	0,9
rs8170	G/A	82,3	0,182619	0,3
rs865686	T/G	62,7	0,002838	11,5
rs889312	A/C	67,8	0,836598	1,4
rs909116	T/C	54	0,047404	0,3
rs9303542	NA	NA	NA	100
rs9348512	C/A	70,9	0,612337	0,8
rs9397437	G/A	91,7	0,692953	0,3
rs941764	A/G	66,8	0,590312	0,5
rs9693444	C/A	64,6	0,557365	0,4
rs9790517	C/T	81	0,62627	0,7
rs9790879	T/C	64,7	0,359586	0,6
rs9833888	G/T	81,9	0,26904	0,4

Tabla 2. Principales estimadores globales de los SNPs analizados en el PRS,

Adicionalmente, para evaluar la posible diferencia en la prevalencia de los diferentes alelos entre los casos y los controles se realizó un estudio de asociación Caso/Control bajo diferentes modelos de herencia, como puede observarse en la **tabla 3**.

ID	codominant	dominant	recessive	overdominant	log-additive
rs10069690	0,4121215	0,1852462	0,7755006	0,2348987	0,2495184
rs10088218	NA	NA	NA	NA	NA
rs1045485	NA	NA	NA	NA	NA
rs10472076	0,4436773	0,3034684	0,7092241	0,2047964	0,5923837
rs10474352	0,6771622	0,7607581	0,3797319	0,9882573	0,5977921
rs10771399	0,00428685	0,00096359	0,77403537	0,00103007	0,00109939
rs10816625	0,15511539	0,32561046	0,06531543	0,61442921	0,17735933
rs10931936	0,5180367	0,259363	0,9061104	0,2785554	0,33831
rs10941679	0,9958649	0,9924885	0,9284776	0,977335	0,9685604
rs10965163	0,2374645	0,5345024	0,1647113	0,3012066	0,8301625
rs10995190	0,9875185	0,8798123	0,9890379	0,8740889	0,8931304
rs10995201	0,6290834	0,3664113	0,9059617	0,3361879	0,4344632
rs11075995	NA	NA	NA	NA	NA
rs11196174	0,8762319	0,6406614	0,9505463	0,6140023	0,7467475
rs11196175	0,6404629	0,3538111	0,9114511	0,3876894	0,4481424
rs11199914	0,6455564	0,5538703	0,6288515	0,3661125	0,8235699
rs11249433	0,7976407	0,6198467	0,7864646	0,5024836	0,8543327
rs113577745	0,4387977	0,3466225	0,7462255	0,2055632	0,5715001
rs11552449	0,07575157	0,50786548	0,05141785	0,17639016	0,99483272
rs11571833	0,1260976	NA	NA	NA	NA
rs11814448	0,5224881	0,254944	0,8028651	0,2673492	0,2588885
rs11820646	0,03192458	0,36132947	0,00868898	0,36336443	0,04978379
rs11977670	0,09154845	0,03033362	0,67814078	0,07354319	0,08392105
rs12422552	0,9077514	0,665798	0,8302351	0,7613733	0,6705912
rs12443621	0,6069376	0,7927495	0,4111874	0,3544334	0,7443175
rs12710696	0,19668944	0,45082023	0,07391261	0,7008605	0,1546009
rs1292011	0,540892	0,6842492	0,2677648	0,7195232	0,4012719
rs13039229	0,04573643	0,02417766	0,75352674	0,01300238	0,07961665
rs13066793	0,4391009	0,8798924	0,2257787	0,6365164	0,883139
rs13281615	0,8512644	0,6220515	0,9258378	0,6010253	0,7990695
rs13294895	0,5407261	0,4690163	0,3223931	0,6968137	0,346474
rs13329835	0,11870176	0,24417643	0,19836173	0,06557991	0,66569897
rs13365225	0,8316134	0,7403449	0,675906	0,6266582	0,8693738
rs13387042	0,01155227	0,00373778	0,09536713	0,15405305	0,00430574
rs1353747	0,2451791	0,2866028	0,3378962	0,1553486	0,4980067
rs1432679	0,01337568	0,08758605	0,00512157	0,37775991	0,0055392
rs1436904	NA	NA	NA	NA	NA
rs1466785	0,0453235	0,02920633	0,71601099	0,0179691	0,20462246
rs1550623	0,9564775	0,9630725	0,7845078	0,886778	0,9678177
rs16857609	0,5160024	0,2992275	0,8820894	0,2519569	0,4482632
rs16886113	0,42728	0,4592074	0,2218617	0,687278	0,323954

rs16991615	0,7883433	0,4921905	0,8455606	0,5136996	0,493878
rs17356907	0,09198229	0,05747714	0,10794968	0,29754803	0,02978004
rs17631303	NA	NA	NA	NA	NA
rs17817449	0,7366269	0,9779843	0,4555301	0,5967671	0,6771022
rs183211	NA	NA	NA	NA	NA
rs184577	0,8566518	0,9060978	0,5792052	0,8638276	0,7465709
rs1895062	0,9033986	0,8609197	0,7442043	0,6671601	0,943616
rs204247	0,4051194	0,2526959	0,2946597	0,858306	0,1790275
rs2046210	0,5680499	0,2982319	0,5972634	0,5277366	0,3100464
rs2236007	0,8098458	0,7625036	0,6351907	0,6070061	0,9315807
rs2253407	0,2613226	0,3359817	0,1141305	0,639567	0,1249389
rs2290203	0,5201966	0,6588528	0,3581089	0,4282841	0,9249922
rs2290854	0,0827388	0,57555558	0,05707635	0,07944055	0,61980803
rs2304277	0,4645729	0,2704446	0,8183456	0,2160477	0,3989891
rs2363956	0,2116731	0,7378901	0,1339463	0,1103891	0,4900391
rs2420946	9,05E-05	0,00011882	0,0016236	0,2779773	1,73E-05
rs2588809	0,9036838	0,8259646	0,6614527	0,9802521	0,7294032
rs2736108	0,9976248	0,987408	0,9450797	0,985541	0,9678897
rs2747652	0,2186702	0,766773	0,1251546	0,1216714	0,4607756
rs27633	0,9747423	0,8239843	0,9109578	0,8979217	0,8317055
rs2823093	0,5561932	0,3644221	0,4088432	0,5808451	0,2897438
rs2943559	0,20537662	0,12038014	0,49783189	0,09619529	0,15778391
rs2981579	0,00189554	0,00138121	0,01101288	0,35464068	0,00043916
rs2992756	0,6407846	0,5857541	0,3629724	0,7567552	0,3747918
rs344008	NA	NA	NA	NA	NA
rs3757322	0,8917316	0,8471637	0,6338984	0,8914548	0,7044059
rs3760982	0,6508658	0,5503715	0,3814083	0,8524982	0,3760443
rs3803662	0,00376528	0,57237634	0,00099478	0,11401996	0,0469107
rs3814113	NA	NA	NA	NA	NA
rs3903072	0,17511538	0,70599376	0,06427855	0,23584533	0,18411457
rs4442975	0,02796762	0,02611758	0,03258564	0,7563553	0,00749571
rs45631563	0,2543313	0,1617539	0,5156997	0,2113369	0,2543313
rs4691139	NA	NA	NA	NA	NA
rs4733664	0,13123319	0,04422969	0,43003702	0,2703152	0,08393439
rs4784227	0,00887257	0,14138946	0,00240245	0,74556712	0,01335619
rs4808801	0,4669194	0,2782547	0,3514502	0,6415157	0,2173498
rs4849887	0,04183502	0,03169957	0,44232588	0,01445939	0,08362183
rs4973768	0,01039465	0,00287474	0,63737415	0,02283687	0,02742427
rs527616	0,09089373	0,29323233	0,03121841	0,6315307	0,06469653
rs554219	0,00021901	0,00015652	0,02457654	0,00086033	5,75E-05
rs58058861	0,3485545	0,3496249	0,3845583	0,2152754	0,5548427
rs58847541	0,00788216	0,00310706	0,13352533	0,0082097	0,00195527
rs6001930	0,4111423	0,3206046	0,3116953	0,4027424	0,2640827
rs614367	0,00072886	0,00032882	0,06344519	0,00128629	0,00016298
rs616488	0,23718508	0,199726	0,51710876	0,09283433	0,47971633

rs619373	0,6411363	0,4903576	0,4210114	0,6610242	0,3992088
rs6472903	0,09403209	0,06450234	0,47356925	0,03614009	0,13147673
rs6504950	0,8688076	0,624455	0,9738593	0,6005183	0,7125736
rs6507583	0,18467843	0,06664486	0,67646706	0,07577782	0,06781983
rs6682208	0,12166583	0,09847583	0,50565102	0,04080335	0,4016838
rs67397200	NA	NA	NA	NA	NA
rs676256	NA	NA	NA	NA	NA
rs6762644	0,6594531	0,9194996	0,3719179	0,6404678	0,6182946
rs6815814	0,2658005	0,5825017	0,1037175	0,5259548	0,2134396
rs6828523	0,2805605	0,3441073	0,3061633	0,1999507	0,5615635
rs704010	0,9248823	0,6952014	0,9366043	0,7546545	0,756235
rs7072776	0,7681271	0,6162678	0,7242761	0,4810455	0,8112194
rs720475	0,5634928	0,2841402	0,7535422	0,3561496	0,3263781
rs72749841	0,08427487	0,43755384	0,02686688	0,83739482	0,14039735
rs72755295	0,225525	NA	NA	NA	NA
rs72826962	0,1559148	NA	NA	NA	NA
rs7297051	0,10730767	0,22857249	0,04784914	0,67214198	0,08798103
rs73161324	0,9607762	0,9761888	0,7847778	0,9376072	0,9865193
rs7529522	0,01936122	0,45485419	0,01579596	0,05609616	0,7501886
rs75915166	0,6129726	0,4498202	1	0,3979656	0,6129726
rs78269692	0,670487	0,6878494	0,515269	0,8599899	0,670487
rs8009944	0,2790718	0,2089924	0,1751983	0,621463	0,1206815
rs8170	0,9291733	0,7405136	0,928963	0,703263	0,8033721
rs865686	0,5359875	0,3348126	0,3787138	0,7604248	0,2642749
rs889312	0,19160084	0,10720644	0,19373419	0,41248733	0,06917039
rs909116	0,4868494	0,294257	0,3813745	0,806084	0,2315649
rs9303542	NA	NA	NA	NA	NA
rs9348512	0,9387827	0,7968581	0,7562247	0,9290591	0,7375225
rs9397437	0,7739317	0,6320193	0,6576209	0,5646535	0,7108719
rs941764	0,7647693	0,8217891	0,5529733	0,5529251	0,9138823
rs9693444	0,19510332	0,28270908	0,31005958	0,07914358	0,78193096
rs9790517	0,12278665	0,06498586	0,69275475	0,04102543	0,14163338
rs9790879	0,5090419	0,2501653	0,8464104	0,3125079	0,3446578
rs9833888	0,4392184	0,4321069	0,2409632	0,7011176	0,2919721

Tabla 3. Valores de significancia estadística de los diferentes SNPs en un estudio caso/control bajo diferentes modelos de herencia,

Utilizando un umbral de significancia estadística de un pValue de 0,05, se obtienen bajo el modelo dominante un total de 14 SNPs como significativos, estos son: **rs10771399, rs11977670, rs13039229, rs13387042, rs1466785,**

rs2420946, rs2981579, rs4442975, rs4733664, rs4849887, rs4973768, rs554219, rs58847541 y rs614367, Por otro lado, usando un modelo recesivo, obtenemos 12 SNPS significativos, 8 exclusivos del modelo recesivo, Estos son: **rs11820646, rs1432679, rs3803662, rs4784227, rs527616, rs72749841, rs7297051, rs7529522.**

Estos resultados se pueden visualizar de forma global en la **Figura 5.**

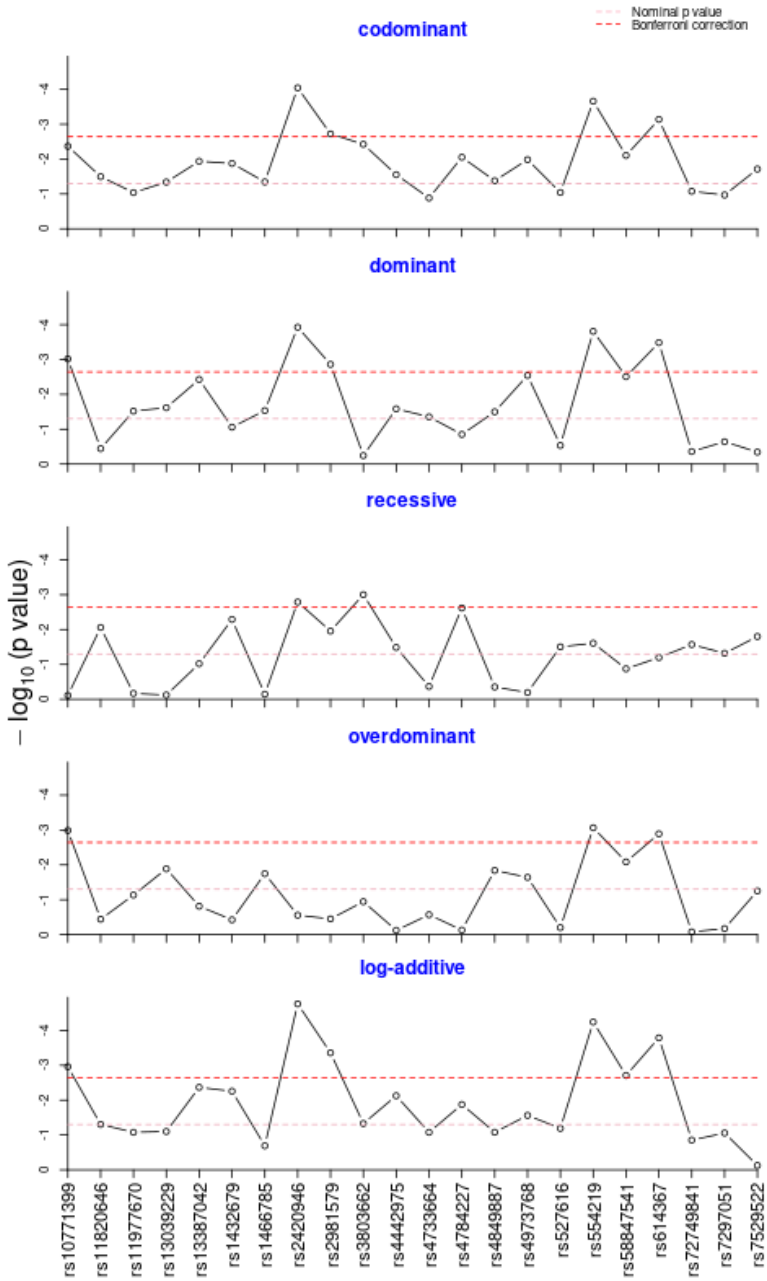

Figura 5. Valores de –log(pValue) frente a diferentes modelos de herencia de los SNPs estadísticamente más significativos.

El valor PRS basado en todos los SNPs estudiados y que están descritos en la **tabla 2** presenta en la cohorte estudiada un OR por 1 desviación estándar (SD) de 1,41, correspondiendo un intervalo de confianza al 95% de 1,24-161 y un pValue en la regresión de 6,30 x10-8, Desglosando el valor de OR del PRS en diferentes franjas del 5%, se observa que mujeres en el decil más bajo (5%), la distribución PRS presenta un OR de media de 0,38 (95 CI 0,22-0,63 y pValue de 0,0026) con respecto a las mujeres de los deciles medios (40%-60%), Este resultado indica que las mujeres cuyo valor de PRS esté en el decil más bajo presenta, como media, 2,5 veces menos probabilidad de desarrollar cáncer de mama respecto a las mujeres cuyo PRS están en los deciles centrales 40-60%,

Por otro lado, las mujeres con valores de PRS en el decil más alto (95%) exhiben un valor medio de OR de 1,87 (95% CI: 1,16 - 3,08, pValue 0,036), **Figura 6.** Es decir, las mujeres cuyo PRS están en el decil más alto presenta una media de 1,87 veces más probabilidad de desarrollar cáncer de mama respecto a las mujeres cuyo PRS están en los deciles intermedios 40-60%,

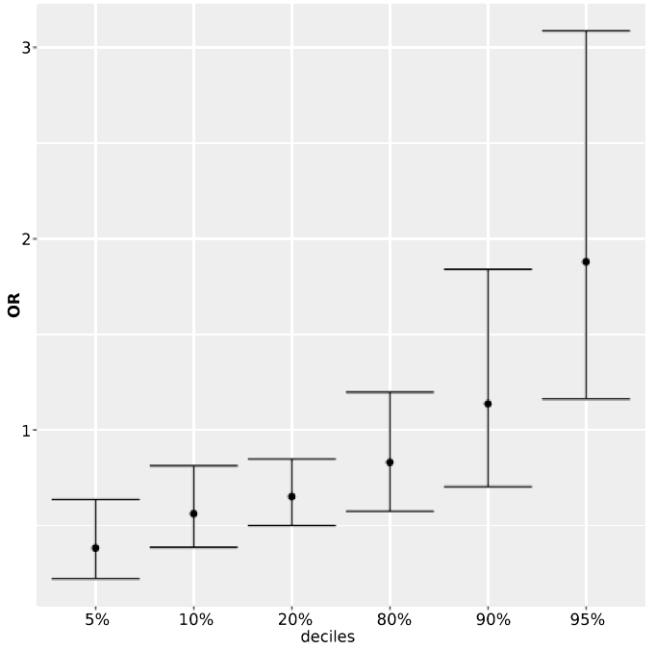

Figura 6. OR basado en PRS dividido por deciles de 5%. Los valores de referencia fueron los obtenidos en el intervalo 40-60%. Las barras indican los intervalos de confianza a los 95% estimados mediante regresión logística.

Adicionalmente, el valor PRS presentó una capacidad discriminarte mediante regresión logística y curva ROC de 0,62, con un intervalo de confianza al 95% de 0,56-0,66.

Respecto al análisis funcional de los SNPs que forman parte del cálculo de PRS, el enriquecimiento funcional de los genes en los cuales afecta dichos SNPs utilizando *Gene Ontology* , *KEGG* y *Reactome* como base de datos de referencia esta descrito en la **Figura 7**.

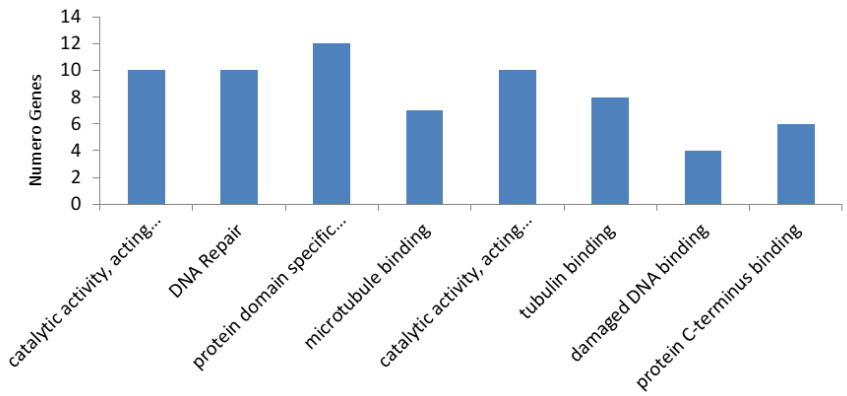

Figura 7. Número de genes observados asociados a los términos de *Gene Ontoloy, KEGG y Reactome* más significativos estadísticamente basados en la lista de genes que afecta los SNPs seleccionados para el cálculo de PRS.

Las categorías funcionales más significativas son la reparación de DNA, actividad catalítica y unión a microtúbulos. Estas funciones están directamente relacionadas con el cáncer de mama. Respecto a la categoría “reparación de DNA”, es observada en 11 de los 84 genes que hay en total, lo que corresponde con un 13% del total del grupo. Es la categoría más representativa y pretende reparar el daño de DNA que puede ser por causa de motivos endógenos y exógenos.

La relación funcional entre los genes afectados por los SNPs estadísticamente significativos seleccionados en este trabajo se corrobora mediante el estudio de redes de interacción entre proteínas. Estudiando la posible relación entre los genes seleccionados mediante su posible interacción física y su relación con rutas metabólicas mediante la teoría de redes se observa una alta correlación física y funcional, **Figura 8.**

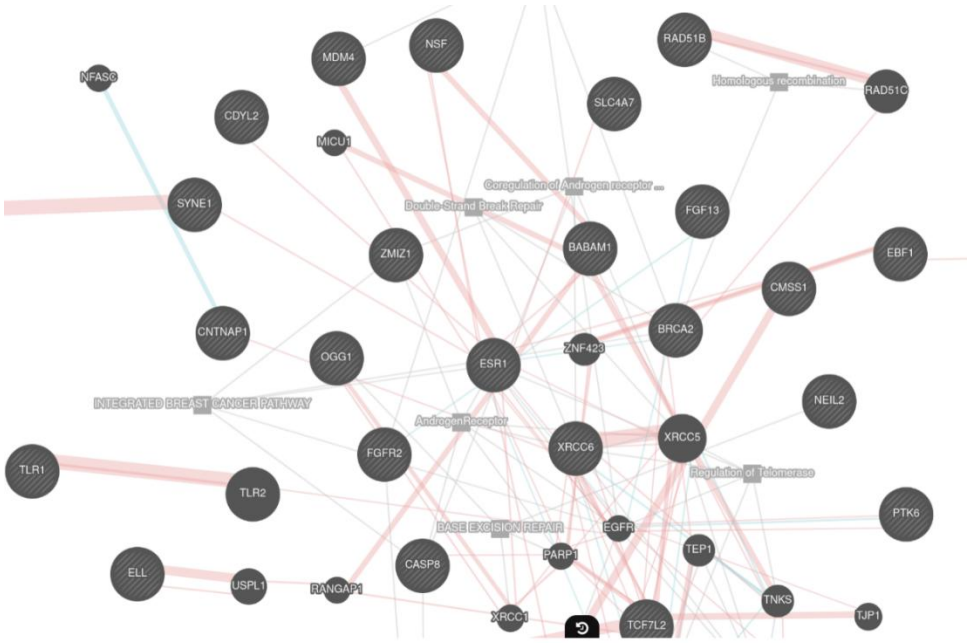

Figura 8. Red de interacción entre los genes afectados por los SNPs seleccionados. Las líneas rosas hacen referencia a la interacción proteína - proteína, Las líneas gris relación proteína – ruta metabólica.

En este estudio se puede observar una alta densidad de interacción física y funcional entre genes y rutas metabólicas. La distribución del número de conexiones obtenida en la red esta descrita en la **Figura 9.**

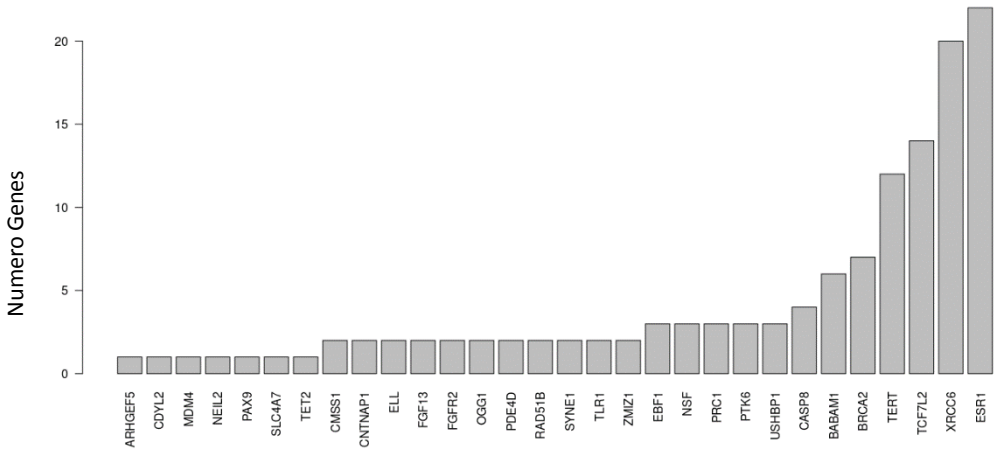

Figura 9. Distribución del número de conexiones por gen de interés.

Se puede observar en la distribución como la mayoría de los genes tiene un número bajo de conexiones mientras que un grupo pequeño de genes presenta mayor conectividad, estos genes son los denominados "**Hubs**". Esta distribución presenta una posible relación exponencial que coincide con las distribuciones descritas en la literatura para redes biológicas mediante la "*ley de poder*" [73]. Según dicha ley, se podría considerar estos genes con alta conectividad o "**Hubs**" como genes de alta importancia funcional y, por tanto, con poca tolerancia a pérdida de función mediante alteraciones patológicas.

Los genes que presentan mayor conectividad son TERT, BRCA2, ESR1 y XRCC6, Estos genes presentan una fuerte relación funcional entre sí, estando asociadas a la ruta de cáncer de mama y a la regulación de telómeros, otro proceso relacionado con el desarrollo tumoral, **Figura 10.**

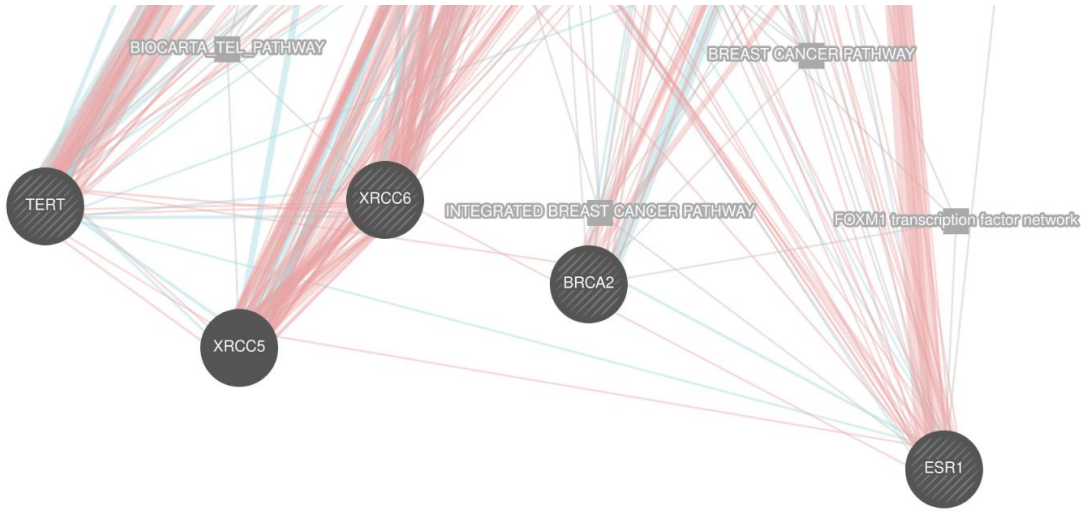

Figura 10. Red de interacción entre los genes con mayor conectividad, Líneas rosas hacen referencia a la interacción proteína - proteína, Línea gris relación proteína – ruta metabólica.

Respecto al estudio de las consecuencias sobre la proteínas de los SNPs seleccionados, se puede observar como la gran mayoría de estas consecuencias, utilizando todos los transcritos descritos en **Ensembl** versión 103, afectan principalmente a regiones no codificantes, es decir, regiones intrónicas (50%), no codificantes (15%), regiones regulatorias "*up/down stream*"

(18%) o posibles sitios de unión a factores de transcripción (5%) o incluso a regiones no asociadas a ningún gen o intergénicas (3%),**Figura10**.

El porcentaje de las variantes que afectan a la región codificante usando cualquiera de los transcritos descritos en ensembl es menor al 2%. Estos resultados son coherentes a lo esperado en SNPs asociado a susceptibilidad, ya que debe tener una afectación ligera o no altamente penetrante. Este tipo de alteración se asocia a zonas regulatorias o a leves alteraciones sobre la estructura de la proteína, **Figura11**.

El 30 % fueron sinónimas en la distribución de las variantes que afectan a la parte codificante, es decir, no hay cambio entre los aminoácidos y, por tanto, es neutral para la estructura de la proteína. Un 58% fueron variantes que modificaron el aminoácido original de la proteína y por tanto tiene modificación de secuencia de la proteína original y un posible efecto estructural en la misma. Finalmente, un 12% tuvieron un efecto aparentemente patológico de alto impacto al generar una posible proteína truncada.

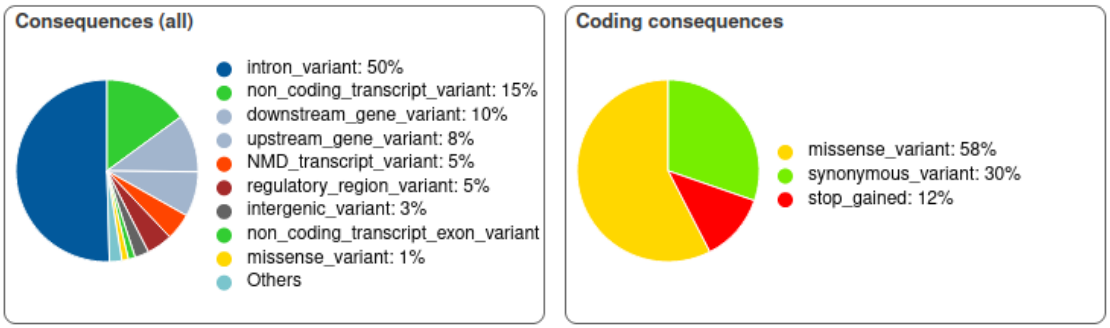

Figura11, Distribución de consecuencias asociadas a los SNP seleccionados en el cálculo de PRS,

4.3 Descripción fenotípica de la cohorte de validación

Los resultados de significancia estadística para las variables fenotípicas usando la cohorte de validación son descritos en la **tabla 4.**

Factor de Riesgo	Categoría	Descripción	Número controles	% Controles	Número casos	% Casos	OR	OR CI 95%	p-value
Edad	0	30-35 years	28	4,36	16	3,52	1,05	0,79-1,13	0,27
	1	35-40 years	58	9,03	28	6,15			
	2	40-45 years	84	13,08	63	13,85			
	3	45-50 years	138	21,50	115	25,27			
	4	50-55 years	158	24,61	86	18,90			
	5	55-60 years	113	17,60	94	20,66			
	6	60-65 years	47	7,32	37	8,13			
	7	>65 years	16	2,49	16	3,52			
Densidad Mamaria	0	From 0 to 10%	99	15,42	51	11,21	1,46	1,21-1,71	1,64E-07
	1	From 11 to 25%	116	18,07	53	11,65			
	2	From 26 to 50%	185	28,82	116	25,49			
	3	From 51 to 75%	181	28,19	133	29,23			
	4	Greater than 75%	61	9,50	102	22,42			
Edad al primer embarazo	0	Less than 20 years	33	5,14	23	5,05	1,15	1,02-1,31	0,03
	1	From 20 to 24 years	165	25,70	104	22,86			
	2	From 25 to 29 years	203	31,62	107	23,52			
	3	From 30 to 34 years	106	16,51	101	22,20			
	4	Greater than 34 years	56	8,72	50	10,99			
	5	Nulliparous	79	12,31	70	15,38			
	0	Less than 46 years	97	15,11	47	10,33			

Edad de menopausia	1	From 46 to 50 years	147	22,90	102	22,42	1,96	1,72-2,24	2,20E-16
	2	Greater than 50 years	110	17,13	71	15,60			
	3	Premenopause	87	13,55	212	46,59			
	4	Menstruating	201	31,31	23	4,97			
Edad Menarquia	0	Equal to or greater than 15 years	34	5,30	34	7,47	0,89	0,78-1,04	0,061
	1	14 years	115	17,91	85	18,68			
	2	13 years	178	27,73	100	21,98			
	3	12 years	140	21,81	110	24,18			
	4	Less than 12 years	175	27,26	123	27,03			
	5	Null	0	0,00	3	0,66			
Ant. Familiar	0	No affected relative	468	72,90	308	67,69	1,05	0,93-1,19	0,34
	1	A first-degree relative diagnosed with breast cancer at age 50 years or older	52	8,10	43	9,45			
	2	A first-degree relative diagnosed with breast cancer younger than 50 years	25	3,89	18	3,96			
	3	1 affected second-degree relative	90	14,02	79	17,36			
	4	2 affected first-degree relatives	4	0,62	5	1,10			
	5	2 affected second-degree relatives	1	0,16	2	0,44			

	6	3 or more affected relatives	2	0,31	0	0,00			

Tabla 4. Datos de prevalencia, OR y significancias estadísticas de cada variable fenotípica utilizada, Todos los riesgos fenotípicos se desglosaron en sus categorías definidas.

Desglosando los resultados para cada categoría se observa que no hay diferencias estadísticas significativas respecto a la edad entre controles y casos obteniendo un pValue utilizando la regresión logística de 0,27. Estos datos corroboran el test Wilcoxon realizado con anterioridad y excluye la edad como variable de confusión.

Respecto a la densidad mamaria, se obtuvo un valor muy significativo entre casos y controles con un pValue de 1 e-07 y un OR de 1,46, con un intervalo de confianza del 95% de 1,21-1,71. Desglosado en las diferentes categorías descritas se observa que la densidad mamaria baja (categorías 0 y 1) presenta una mayor prevalencia en controles frente a los casos, en torno al 11%. Esa tendencia cambia en las categorías de mayor densidad mamaría cuyo máximo se alcanza en las mujeres cuya densidad mamaria se encuentra al 75%, donde la prevalencia en casos es del 22,42% respecto al 9,50 en controles, es decir, la densidad mamaria de más alta categoría (categoría 4), tienen una prevalencia de 2,36 veces mayor en mujeres que han desarrollado cáncer frente a las mujeres que no lo han desarrollado.

Respecto a la edad de la mujer al primer embarazo de un hijo vivo se obtuvo una modesta significancia con un pValue asociado de 0,03 y un OR de 1,15 con intervalo de confianza al 95% de 1,02-1,31. En cuanto a las categorías, hubo mayor diferencia en controles respecto a casos en mujeres con su primer hijo a temprana edad (de 20 a 24 años y 25 a 29) con una diferencia de 3 % y 8 % respectivamente. Estos datos confirman que las mujeres que tienen hijos a edades tempranas tienen menos probabilidad de padecer cáncer de mama. Esa tendencia cambia según aumenta la edad, por ejemplo, en las mujeres que tuvieron el primer hijo de 30 a 34 años, presentaron un 6 % más de casos que controles. En la misma tendencia, aquellas mujeres que no han tenido hijos

presentan un 3% más de prevalencia de cáncer de mama respecto a las que han tenido hijos.

Otra de las variables fenotípicas más significativas junto con la densidad mamaria fue la edad de la menopausia con un pValue de 2,2 e-16 y un OR de 1,96, intervalo de confianza al 95% de 1,71-2,24. En esta categoría se observa que las diferencias son en las mujeres menstruantes con un 33,31 % de los controles y solo el 4,97 % de los casos y en las con premenopausia, donde los casos presentan una prevalencia del 46,59 % y los controles el 13,55 %. Como la menopausia depende de la edad y estas prevalencias pueden afectar la distribución de edad de las mujeres en la cohorte en validación, se estudió la posible interacción de esta variable con la edad para corregir cualquier desviación en este aspecto,

Otra variable reproductiva analizada fue la edad de la menarquia. En esta variable se obtuvo una modesta significancia de 0,061 con un OR de 0,89 e intervalos de confianza al 95% de 0,78-1,04. Según el OR obtenido, esta variable es protectora, encontrándose la mayor diferencia en 13 años, donde la prevalencia de los controles fue en torno a un 6%. Considerando que la edad media de menarquia en la población es a los 12 años, las mujeres que desarrollaron la menarquia un año después presentaron menos probabilidad de desarrollar cáncer de mama, confirmando el factor protectivo. Estas diferencias no se observaron a edades de 14 ,15 años o superiores. Estos últimos datos no concuerdan con lo esperado y puede ser debido a algún sesgo de la cohorte de validación por lo que estudios independientes con nuevas cohortes deben ser realizados en el futuro para corroborar dicha relación.

Respecto a los antecedentes familiares no se encontró ninguna significancia estadística, pValue de 0,34 y OR de 1,05, intervalos de confianza al 95% de 0,93-1,19. Estos resultados no concuerdan con lo esperado en la literatura científica. Observando las prevalencias se puede indicar como las mujeres que presentan un familiar afecto de segundo grado representan un 17,36% de casos respecto al 14,02 % en controles. En el resto de las categorías no se observaron diferencias significativas, sin embargo, este hecho puede ser debido al bajo número de mujeres representativas de cada categoría en nuestra cohorte de validación, siendo necesario ampliar la cohorte para tratar

de obtener mayor número de mujeres y tratar de verificar la relación lo que está descrita en la literatura científica.

Todas las variables significativas fueron seleccionadas para su inclusión en el modelo multivariable. Respecto a los antecedentes familiares, a pesar de no ser significativa, debido a las evidencias científicas observadas y después de un consenso clínico, se decidió añadirla también al modelo.

4.3 Capacidad discriminante de las variables analizadas

Adicionalmente, para cada una de las variantes fenotípicas y para la variable genética PRS se calculó el poder discriminante mediante la evaluación del área bajo la curva ROC (AUC). En este estudio, cada variable, fue ajustada a la edad de la mujer, lo que da una estimación más precisa. Los resultados se describen en la **tabla 5.**

Model	**Median AUC**	**95% CI AUC**	**p-value**
Densidad mamaria	0,60	0,54-0,66	2,17E-03
Edad primer hijo	0,54	0,48-0,60	1,49E-01
Edad menopausia	0,64	0,58-0,70	5,40E-09
Antec, familiares	0,52	0,47-0,58	6,45E-01
Edad menarquia	0,53	0,48-0,59	2,80E-01
PRS	0,62	0,56-0,66	3,64E-03

Tabla5, Valores de significancia estadística, intervalos de confianza y área bajo la curva ROC (AUC) de los modelos univariables para cada factor de riesgo ajustado por la edad de la mujer,

Según estos resultados, las variables utilizadas presentan un importante poder discriminante, principalmente la edad de la menopausia, PRS y densidad mamaria con valores de 0,64, 0,62 y 0,60 respectivamente. Por el contrario, tal y como se ha observado en la sección anterior, las variables antecedentes familiares y edad de la menarquia presentan una modesta capacidad de discriminación entre casos y controles, siendo los valores de 0,52 y 0,53 respectivamente.

4.4 Modelo multivariable para la estratificación de la población general en función de la probabilidad de sufrir cáncer de mama esporádico

Como se ha descrito en la sección 3,2, finalmente, se decidió incorporar todas las variables significativas estadísticamente. En el caso de antecedentes familiares, a pesar de que no fue significativa, se decidió incorporarlo al modelo por recomendación clínica al ser un factor de riesgo ampliamente asociado a cáncer de mama esporádico en la literatura científica.

Para la identificación de posibles colinealidades entre variables se realizó una correlación de Spearman utilizando los controles de la cohorte a estudiada. Los resultados se describen en la **Figura 12.**

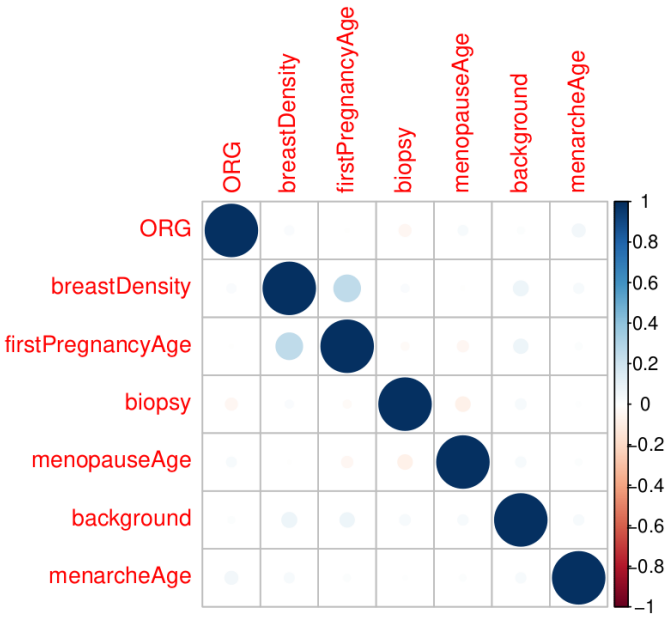

Figura 12. Valores de correlación de Spearman entre las diferentes variables introducidas en el modelo utilizando toda la cohorte de validación.

Según los resultados observados no se encontró ninguna colinealidad evidente entre las variables analizadas utilizando un umbral de +/- 0,50 de correlación de Spearman. Por estos resultados se incorporaron todas las variables al modelo multivariable. Los resultados del área bajo la curva ROC (AUC) para el

modelo multivariable con y sin interacciones pueden ser observados en la **Figura 13**.

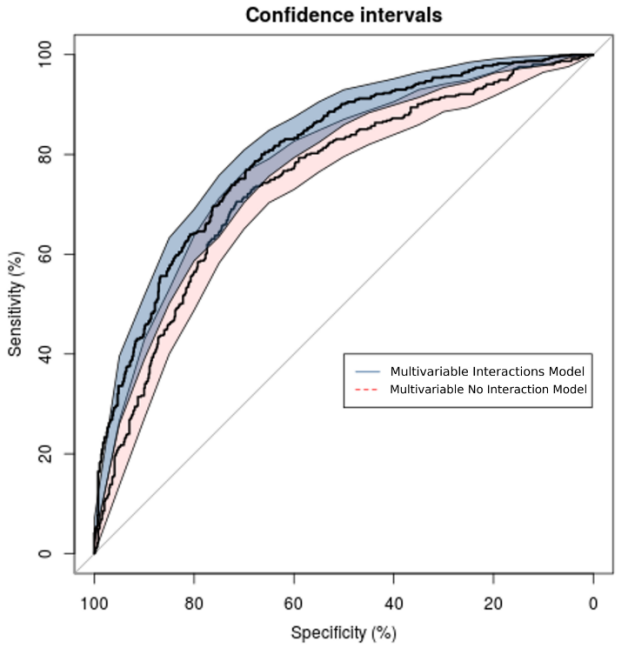

Figura 13. Valores de la curva ROC para el modelo multivariable con interacciones (azul) y sin interacciones (rojo).

Los resultados obtenidos para el modelo multivariable sin interacciones son de una media en la curva ROC (AUC) de 0,74, con un intervalo de confianza al 95% de 0,71-0,77 y un pValue inferior a 2,26 e-16. Por otro lado, el modelo multivariable usando las diferentes interacciones genera un valor medio de AUC de 0,80, con un intervalo de confianza al 95% de 0,77-0,83 y un pValue inferior a 2,26 e-16.

Según los resultados obtenidos se puede indicar que el modelo multivariable tanto con interacciones como sin interacciones tiene una capacidad discriminante superior a todos los modelos univariables analizados. Adicionalmente, el modelo multivariable con interacciones presenta una capacidad discriminante superior al modelo multivariable sin interacciones, con una diferencia media del 6%. Ambos modelos presentaron una diferencia significativa con un pValue de 5,375 e-09. Estos resultados presentan mayor

capacidad discriminante que otros modelos publicados con anterioridad [74], [75].

Adicionalmente se investigó el riesgo relativo tanto en los casos como en los controles utilizando el modelo multivariable con interacciones descrito en el apartado anterior. Los valores de OR y porcentajes desglosados por deciles son descritos en la **Figura 14** y **tabla 6.**

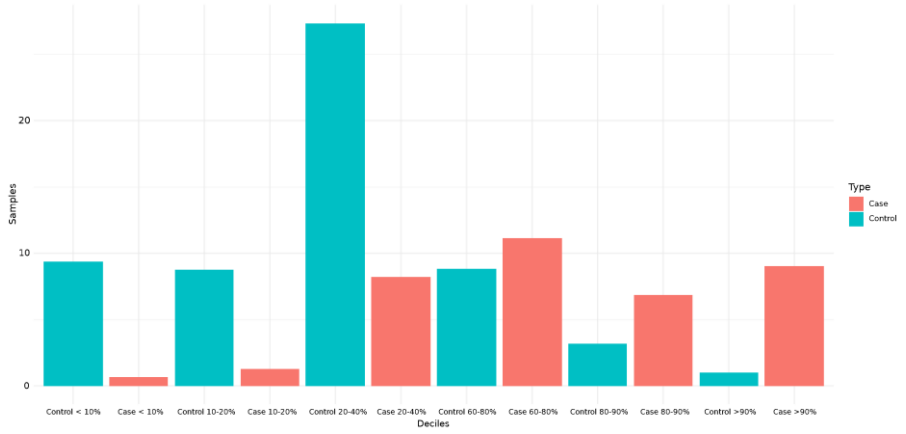

Figura 14. Distribución de casos y controles para cada decil del riesgo obtenido utilizando el modelo multivariable con interacciones.

Deciles	OR	OR 5%	OR95%	p-value	% Controls	%Cases
<10%	0,097	0,046	0,184	1,86E-08	9,39	0,64
10-20%	0,209	0,121	0,345	8,12E-07	8,75	1,28
20-40%	0,402	0,282	0,570	1,99E-05	27,35	8,20
60-80%	1,803	1,313	2,481	2,30E-03	8,84	11,12
80-90%	3,071	2,057	4,634	5,31E-06	3,19	6,84
>90%	12,900	5,098	23,332	3,43E-07	1,00	9,02

Tabla 6. OR y frecuencias entre casos y controles para los diferentes deciles de las muestras de la cohorte basados en el riesgo relativo obtenido del modelo multivariable con interacciones,

Según los resultados, se puede observar como las mujeres que presentan un riesgo dentro del decil más bajo (<10%) tiene una proporción mayor de controles respecto a casos, con una frecuencia del 9,39% y 0,64% respectivamente, es decir, las mujeres que presentan un riesgo en este decil presentan 14 veces más probabilidad de no padecer cáncer de mama. Esta tendencia se mantiene en el siguiente decil (entre el 10%-20%) con un 8,75% de controles respecto al 1,28% de los casos, es decir, la proporción de controles es 6,8 veces superior a los casos.

Por el contrario, la tendencia cambia para los deciles mayores, para el decil >=90% se observa una frecuencia del 9% de los casos respecto al 1% de los controles, presentando un OR de 12,9. El decil 80%-90% también presenta la misma tendencia con un OR de 3,071, es decir, las mujeres que presentan un riesgo dentro de este decil presentan 3 veces más probabilidad de padecer cáncer de mama en los siguientes 5 años respecto a la población general.

Estos resultados demuestran la capacidad discriminante del modelo multivariable con interacciones dentro de la cohorte a estudio. Todos los deciles estudiados fueron estadísticamente significativos.

Capítulo 5. Discusión:

En los últimos años ha habido diferentes propuestas donde utilizan modelos matemáticos multivariables para la estratificación de las mujeres de la población general en función de un riesgo individual a padecer cáncer de mama esporádico. Este tipo de pruebas ha generado un alto interés por parte de la comunidad científica ya que permite acercarnos al concepto de medicina personalizada dentro del ámbito de la prevención.

Estos modelos utilizan diferentes marcadores cuya capacidad discriminante respecto al fenotipo estudiado ha sido respaldada por la comunidad científica a lo largo de los años en diferentes consorcios y en diferentes poblaciones. Estos marcadores o factores de riesgo generalmente se engloban en diferentes categorías; factores fenotípicos, ambientales y genéticos. Los factores genéticos son muy importantes en cánceres de mama hereditarios y esporádicos, aunque los modelos genéticos en los que se basan cambian sustancialmente. En el caso del cáncer de mama esporádico, la base genética parece basarse en un modelo poligénico. Este modelo explicaría la naturaleza compleja y multivariable del fenotipo donde no existe alteración genética de riesgo elevado, sino una acumulación de alteraciones de riesgo bajo o moderado independientes entre sí que genera un riesgo global. Este tipo de modelos parece adaptarse mejor a lo observado en los pacientes, ya que, por regla general, el cáncer de mama esporádico suele aparecer a partir de los 55 años, lo que significa que la mayor parte de la vida de la mujer está libre de cáncer.

Este tipo de modelo poligénicos han sido validados en diferentes tipos de enfermedades como la diabetes, el Parkinson y diferentes tipos de cáncer, entre los más estudiados están el cáncer de colon, próstata y mama.

Para profundizar en este tipo de aproximación es necesaria la identificación de nuevos SNPs o alteraciones genéticas estadísticamente significativas y asociadas al fenotipo estudiado, además del cálculo robusto y preciso de su riesgo individual (OR), que normalmente depende de la etnia estudiada, Es decir, es necesario la formación de grandes consorcios donde se pueda

incorporar el análisis de GWAS de miles de pacientes teniendo en cuenta todas las variables que puedan influir en este tipo de estudio, como la etnia, la edad etc, Actualmente, existen consorcios importantes como “European Collaborative Oncological Gene Environment Study (COGS)” en el ámbito oncológico que permite tener un conocimiento cada vez más profundo de este tipo de información.

Dentro del ámbito del cáncer de mama esporádico cada vez hay más consorcios y estudios independientes que permiten corroborar la información existente e identificar nuevas alteraciones genéticas de interés, no solo en la población caucásica, sino en otras poblaciones menos estudiada como la población afroamericana, asiática, latinoamericana o como en este estudio, población española [76] y [77].

En este contexto, la selección de 121 SNPs estadísticamente significativos obtenidos de los consorcios COGS y oncoArray y su validación en una cohorte independiente basada en población española tiene un interés clínico, ya que cada vez hay más evidencias científicas de las mejoras que conlleva la incorporación de este tipo de información en los programas de cribado. En esta validación, analizando individualmente cada SNPs, se puede observar como la gran mayoría de estas alteraciones tienen una frecuencia alta del alelo mayoritario, corroborando el hecho de que son alteraciones relativamente frecuentes en ambas poblaciones, la caucásica y nuestra cohorte estudiada representativa de la población española. Además, no encontrar desplazamientos Hardy-Weinberg significativo también corrobora la reproducibilidad de estas alteraciones entre ambas poblaciones, Adicionalmente se observó problemas tecnológicos en la identificación de 4 de los SNPs, quedando finalmente 115 SNPs.

El estudio funcional de los 115 SNPs identifica genes asociados a daño y reparación de DNA, Este tipo de mecanismo molecular, a los cuales pertenece BRCA1 y BRCA2, está altamente asociado a cáncer de mama. El fallo de dicha maquinaría puede ocasionar la aparición de errores en el proceso de replicación celular aumentando la probabilidad de mutaciones en

protooncogenes o genes protectores de tumores. Este hecho podría explicar la mayor probabilidad de tener cáncer.

Otro tipo de funciones observadas, como la modulación de la tubulina, está asociado a otros aspectos del cáncer como la regulación de la mitosis y el arresto celular. Estudios más completos utilizando la biología de sistemas e integrando diferentes capas de regulación como la genética y la interacción física entre proteínas, permite identificar genes importantes asociados al cáncer de mama, organización de los telómeros o proliferación celular, estos son TERT, ESR1 y XRCC6.

Por otro lado, el estudio de consecuencias sobre la proteína de los SNPs seleccionados no identifica variantes claramente patológicas, sino que principalmente están asociados a elementos regulatorios e intrónicos. Este tipo de hallazgos son concordantes con lo obtenido en otros tipos de GWAs, donde se espera alteraciones que generen modificaciones estructurales de bajo impacto en la estructura y absorbibles por el organismo, ya que son variantes relativamente frecuentes en la población general y que de manera individual no debe generar un riesgo alto sobre el fenotipo [78].

La única variante con posible consecuencia patológica generando una posible proteína truncada tiene como id rs11571833. Esta variante es un cambio A/T que afecta al último exón de los 4 transcritos descritos en el gene BRCA2, gen asociado con alta penetrancia a cáncer de mama hereditario además de ser uno de los genes con mayor conectividad en la red de regulación. Según la base de datos clínica *ClinVar*, esa variante es posiblemente benigna, categorización que coincide con las guías ACMG de diagnóstico clínico [79], donde las variantes truncantes que afecta al último exón de los transcritos no suelen ocasionar patología, posiblemente debido a la poca afectación estructural.

Respecto a las otras variantes con posibles efectos patológicos en sus consecuencias, es decir, aquellas descritas como “missense” en algunos de los transcritos descritos en la base de datos Ensembl, fueron identificadas un número total de 3 variantes. Estas variantes afectan a los genes DCLRE1B,

ANKLE1 y MCM8. Estos genes están asociados a una actividad catalítica (GO: 0140097, “catalytic activity, acting on DNA”), con pvalue ajustado mediante FDR de 1,014E-4. Sin embargo, no parecen afectar a ninguna ruta específica. Según las predictores de patogenicidad según propiedades fisicoquímicas, 2 de las 3 variantes se consideraron benignas, es decir, el cambio de aminoácido no altera significativamente la estructura y/o función de la proteína a nivel global. En la última variante, que afecta al gen ANKLE1, es patogénica según ambos predictores utilizados, no obstante, según el dominio de la proteína a la cual afecta, es una región “ankyrin repeat”, PTHR46427. Este dominio es un dominio repetitivo y poco conservado de unidades de entre 30-40 aminoácidos principalmente asociadas a dominios de unión con otras proteínas. Estos datos parecen indicar que esta variante, funcionalmente, no tiene un efecto relevante en la estructura o la función final de la proteína asociada.

Respecto a variantes que afectan a elementos regulatorios, se encontraron 5 variantes que afectan a sitios de unión de diferentes factores de transcripción, estos factores son ELF1, MAX y E2F8. El estudio de la posible patogenicidad de este tipo de variantes es complejo y aún no bien definido por la comunidad científica. Esto es debido a que los factores de transcripción suelen interaccionar con motivos de DNAs, normalmente denominado logos, que funcionan como una matriz de probabilidad donde algunos cambios son más permitidos que otros [80]. Las variantes que afecta a estos sitios de unión de estos factores de transcripción pueden afectar de diferentes maneras, haciendo menos estable la interacción y, por tanto, reduciendo la regulación entre ambos genes sin eliminarlas completamente [81]. Este hecho podría explicar la susceptibilidad, pero se necesitan estudios más profundos para confirmar dicha hipótesis.

Mediante el estudio funcional se puede obtener una visión global de los posibles efectos de los SNPs estadísticamente significativos utilizados para el cálculo de PRS. Como era de esperar, el efecto funcional y estructural es mínimo, no encontrándose evidencias de una fuerte patogenicidad en ninguna de las variantes consideradas. No obstante, se observaron evidencias de posible modulación funcional en muchas de ellas. Respecto a los genes

afectados, diferentes estudios estadísticos indican como son claramente enriquecidos en rutas asociadas a cáncer y más concretamente a cáncer de mama.

Respecto a la reproducibilidad entre las cohortes de la significación estadística y OR individual de cada SNP, se observó diferencias importantes. Basándose en la cohorte de validación, es decir, en la cohorte representativa de la población española, solo 17 de los 115 (~14%) SNPs analizados presentaron significación estadística en la cohorte bajo modelo dominante y/o recesivo. Este número aumenta a 26 (~22,5%) si consideramos otros modelos de herencia como log-aditivo o codominante. La baja reproducibilidad puede ser debida al bajo número de mujeres analizadas en la cohorte estudiada para este tipo de análisis estadístico, además de posibles sesgos en el criterio de selección de las mujeres, por lo que nuevos estudios incorporando mayor número de mujeres de diferentes localizaciones es necesario para corroborar estos resultados y estudiar en profundidad la posible concordancia entre cohortes en base a los 115 SNPs.

Analizando los 10 SNPs más significativos obtenidos en la cohorte española (pValue mínimo de 0,0016), todos han sido asociados a cáncer de mama en otros estudios independientes [82]. Este hecho podría corroborar que estos SNPs son alteraciones corroboradas. Funcionalmente, 2 de los 10 SNP afectan a diferentes intrones del gen FGFR2, receptor del factor de crecimiento de fibroblastos, gen importante en la regulación del crecimiento y diferenciación celular y estimula la angiogénesis, mecanismos moleculares asociados al desarrollo de diferentes tipos de canceres [83].

A pesar de la baja reproducibilidad obtenida en la cohorte representativa de la población española se decidió incorporar todos los SNPs al valor de PRS manteniendo los OR obtenidos en los consorcios internacionales. El hecho de que cada SNPs tenga un peso relativamente pequeño dentro de este tipo de modelo puede absorber las desviaciones puntuales observadas. Al contrastar este valor de PRS, basado en 115 SNPs, con la cohorte representativa de la población española, se obtuvo un nivel de significancia muy elevado de 6,30 E-8, por lo que, a pesar de las diferencias puntuales observadas en los SNPs, el modelo acumulativo aditivo si fue significativo estadísticamente por lo que se

consideró continuar con el PRS basado en los definitivos 115 SNPs. Este valor genético representa para la cohorte estudiada, un factor de riesgo de 1,41 con confianza al 95% de 1,24-1,61. Dividiendo la cohorte en diferentes segmentos se observa como los valores extremos presentan valores de OR muy diferentes, representando 0,38 (95 CI 0,22-0,63) el 5% con el valor de más bajo frente al 95% que presentó un OR de 1,87 (95% CI: 1,16 - 3,08, pValue 0,036). Estos valores indican que el PRS basado en los 115 SNPs tiene la capacidad de discriminar adecuadamente los extremos 5% de la cohorte estudiada. De la misma manera, pero con significancia más baja (datos no mostrados) para los extremos 10%. Utilizando estimadores globales como AUC, aplicado a un modelo caso control y usando este PRS se obtienen valores significativos de 0,62 (95% CI: 0,56-0,66, pValue 3,64E-03). Este valor es concordante con los resultados obtenidos en la literatura científica basados en población europea (0,58-0,65) y en población no europea (0,53-0,64) [84].

Además, comparado con los resultados obtenidos para otros riesgos fenotípicos, este valor fue superior incluso a factores de riesgo estudiados por la comunidad científica como la edad de la menarquia o la de la mujer al nacimiento del primer hijo vivo. Estos resultados podrían corroborar la relación de estos SNPs seleccionados con el riesgo a padecer cáncer de mama en la cohorte representativa de la población española, cuya capacidad discriminante es por sí sola capaz de discriminar los valores extremos. Sin embargo, debido al bajo número de mujeres de la cohorte estudiada, estos resultados deben ser consideramos como una prueba conceptual que abre la posibilidad de incorporar este tipo de información al seguimiento de la mujer dentro de los programas de cribado, pero es necesario incrementar el número de mujeres y ampliar el número de centros para tener valores más robustos y detallados.

Dentro de los factores fenotípicos de riesgo a padecer cáncer de mama esporádico más importantes son los factores reproductivos, especialmente la edad de la menopausia y de la menarquia. Tener la menopausia después de los 55 años aumenta el riesgo de padecer cáncer de mama y cáncer de endometrio. Probablemente se debe a que la mujer ha estado expuesta durante un mayor número de años a estrógenos endógenos. Durante el ciclo menstrual de una mujer, el estrógeno estimula el útero y el tejido mamario, por

lo tanto, cuantos más períodos menstruales tenga una mujer, más tiempo estarán expuestos estos tejidos al estrógeno endógeno [85]. Los mecanismos moleculares por el cual los estrógenos favorecen el desarrollo del cáncer de mama no están bien definidos aun por la comunidad científica. La teoría más aceptada y más respaldada por evidencias científicas [86], [87] sostiene que los efectos del estradiol (E2), actuando a través del receptor de estrógenos alfa estimula la proliferación celular e induce la posibilidad de incorporar mutaciones durante el proceso replicativo del DNA [88]. Este efecto promocional inducido por E2 apoya el crecimiento de estas células que albergan estas mutaciones que se acumulan hasta que se produce el cáncer.

Otro de los principales factores de riesgo dentro de las variables reproductivas es la edad del primer hijo nacido. Es conocido desde hace tiempo que el embarazo protege contra el cáncer de mama y que el nivel de protección está directamente relacionado con el número de embarazos [89]. Esta protección se observa con la edad de la madre cuando da a luz a su primer hijo [90], [91]. Por ejemplo, las mujeres que dan a luz a su primer hijo antes de los 35 años presentan un menor riesgo de padecer cáncer de mama [92]. También presentan un menor riesgo las mujeres que han dado a luz respecto a las que nunca han sido madre [93].

El motivo biológico de esta relación es complejo y poco conocido, Se cree que durante el embarazo las células mamarias crecen rápidamente. Este factor es importante si existe algún daño genético en dichas células antes del embarazo debido a que el mayor número de copia de células con material genético dañado aumenta la probabilidad de cáncer. Por lo tanto, debido a que la probabilidad de tener daño genético aumenta con la edad, esto puede explicar por qué las mujeres que tienen su primer hijo a una edad más avanzada tienen un mayor riesgo de cáncer de mama que las mujeres que tienen su primer hijo a una edad más temprana. Otra posibilidad es que, a mayor número de embarazos, menor es el número de ciclos menstruales que presenta la mujer a lo largo de su vida y por tanto tiene menor exposición a estrógenos.

Otro de los factores fenotípicos cuya asociación a cáncer de mama es ampliamente estudiado es la densidad mamaria (DM). La radiografía mamaria permite observar variaciones en la composición del tejido mamario. Mientras el

tejido graso aparece con tonalidades más oscuras, el tejido conectivo es más luminoso. El porcentaje de la parte densa (estromal, epitelial y adiposo) de la mama es lo que se conoce como densidad mamaria [94]. Las mujeres con DM alta tienen más células estromales y epiteliales y menos tejido adiposo graso, y tienen más probabilidades de desarrollar cáncer de mama a lo largo de su vida en comparación con las mujeres con DM baja [95]. Normalmente, la densidad mamaria se clasifica en 4 categorías que van desde el tejido prácticamente graso hasta el tejido extremadamente denso y con muy poca grasa. El radiólogo decide de manera relativamente subjetiva cuál de las 4 categorías se asocia cada seno. Esta clasificación se denomina BI-RADS [96]. El nivel uno define un tejido mamario casi completamente graso con una densidad de tejido del 5% al 24%. A esta categoría pertenece el 10% de las mujeres de EEUU [97]. El nivel dos es definido por un tejido mamario compuesto por áreas dispersas cuya densidad oscila entre el 25% y el 49%, sin embargo, la mama aún presenta un tejido principalmente adiposo. A esta categoría pertenece el 40% de las mujeres de EEUU. El tercer nivel, descrito como densidad heterogénea, indica áreas no densas y con un valor global del tejido que va desde el 50% hasta el 75%. A esta categoría pertenece el 40% de las mujeres de EEUU. Finalmente, el nivel cuatro se caracteriza por un tejido extremadamente denso (>=75%) donde presenta muy poco o prácticamente nada de tejido adiposo. A esta categoría pertenece el 10% de las mujeres de EEUU [95].

Las mujeres que tienen tejido mamario denso tienen un mayor riesgo de padecer cáncer de mama en comparación con las mujeres con tejido mamario menos denso. En la actualidad no está claro por qué el tejido mamario denso está relacionado con el riesgo de cáncer de mama. Puede ser que tenga más células que pueden convertirse en células anormales o también presenta mayor dificultad para los radiólogos ver el cáncer en las mamografías [98]. Esto podría significar que las mujeres con senos densos tienen más probabilidades de experimentar falsos positivos y falsos negativos en las interpretaciones radiológicas de las mamografías [99]. Wolfe JN fue el primer investigador en observar y publicar la asociación entre la presencia y proporción de tejido mamario denso y la aparición de cáncer de mama [100], [101]. En diferentes

estudios realizados en grandes consorcios se pudo demostrar que el riesgo relativo de desarrollar cáncer de mama en mujeres cuyos senos estaban en la primera categoría de densidad mamaria era de 1,79, 2,11 los de segunda categoría, 2,92 los de tercera y 4,64 los de categoría cuarta [102]. Estos datos corroboran la existencia de una fuerte asociación positiva entre el aumento de la densidad mamaria y el aumento del riesgo de cáncer de mama.

Según algunos estudios la densidad mamográfica puede ser hereditaria [103], por lo que demostraría la importancia del componente genético en este estimador. Sin embargo, aún se desconoce si este posible efecto hereditario está influenciado por otro tipo de factor, como el ambiental [104]. Otro estudio, demostró la posible relación inversa entre el número de nacimientos o estado de paridad con el porcentaje de colágeno en el tejido mamario, es decir, con la densidad de los senos [105]. Otro factor que podría estar relacionada con la densidad mamaria podría ser la etnia de la mujer [106]. Estos estudios indican que las mujeres asiáticas presentan mayor densidad mamaria respecto al resto de etnias estudiadas, siendo las mujeres afroamericanas las que menor valor presentan.

La densidad mamaria también ha sido asociada al tipo y la capacidad de invasión del tumor indicando que los senos cuyo nivel de densidad es elevado muestran el doble de riesgo de padecer tumores invasivos respecto a mujeres de densidad promedio [107]. En este estudio hubo una asociación positiva entre la densidad mamaria y cánceres de mama tipo ER-HER2- en mujeres menores de 55 años en comparación con mujeres mayores de 55 años. El receptor 2 del factor de crecimiento epidérmico humano (HER2) es un miembro de la familia de receptores del factor de crecimiento epidérmico (EGFR) y se sobre expresa en aproximadamente el 30 % de los cánceres de mama invasivos. La DM alta está fuertemente asociada con tumores grandes, ganglios linfáticos positivos y tumores ER− en mujeres menores de 55 años. Esto sugiere que la DM podría potencialmente jugar un papel en la agresividad de los cánceres de mama.

Otro factor fenotípico descrito como factor de riesgo a padecer cáncer de mama son las biopsias previas que haya podido tener la mujer. La frecuencia de este tipo de procedimiento debido a enfermedades benignas de la mama es

muy elevada, en EEUU se ha determinado que pueden ser hasta 4 veces más frecuentes en el diagnóstico de cáncer de mama invasivo, siendo aproximadamente 1 millón de mujeres anuales en dicho país [100]. Este factor de riesgo se debe a las condiciones fisiológicas que llevaron a ejecutar la biopsia y no a los procedimientos de la biopsia en sí misma. A pesar de que la mayoría de los cambios observados dentro de la mama mediante biopsia no son cancerígenos, algunos pueden incrementar el riesgo a desarrollar cáncer de mama. Dentro de este grupo, los más habituales son la hiperplasia atípica (donde la alteración presenta un número y características celulares anormales), carcinoma lobular in situ (donde se encuentra células anormales en los lobulillos de la mama) y carcinoma ductal in situ (las células anormales se encuentran en el revestimiento de los conductos mamarios).

La hiperplasia atípica comprende entre el 3 %-4% de todas las enfermedades benignas de la mama que provoca la petición de una biopsia y puede incrementar hasta 4 veces el riesgo a padecer cáncer de mama [108]. Las alteraciones lobulares in situ representan entre 0,8% hasta 3,5% de las biopsias [109] y puede incrementar el riesgo entre 3 y 10 veces respecto a la población basal [110], de hecho, las mujeres con este tipo de alteración suelen tener una pauta de 1 mamografía anual y examen clínico de la mama cada 6 meses, en algunos casos, puede incluso aconsejarse la quimio prevención.

Respecto a la verificación usando la cohorte representativa de la población española de estos factores de riesgos fenotípicos de manera puntual, se observó como la edad de la menopausia y la densidad mamaria presentó los valores estadísticamente más significativos con 2,20E-16 y 1,64E-07 respectivamente.

Respecto a la edad de la menopausia se observó un OR de 1,96 donde los desplazamientos más significativos fueron observados en las mujeres premenopáusicas, donde se observó un desplazamiento del ~13% de controles respecto a un ~46% de los casos. Estos datos son concordantes en la orientación y signo del riesgo con los resultados obtenidos por Pollán [53] y en otras referencias [111], donde el estatus de premenopausia es un factor de

riesgo a padecer cáncer de mama, aunque la diferencia entre los casos y controles presentada por Pollán fue sustancialmente menor.

Otra categoría con una amplia diferencia entre casos y controles son las mujeres aún menstruantes. Para este caso, Pollán no describe esta categoría en sus estudios, por lo que no hay valores de referencia, no obstante, la desviación observada en el presente estudio, con un ~31% de los controles y ~5% de los casos, es debido a la incorporación de un subgrupo de mujeres de menor edad (entre 40 y 45 años) dentro de la cohorte. Este punto es de interés ya que permite validar la capacidad discriminante de las variables y de los modelos propuestos no solo en mujeres con edad concordantes a la inclusión de los programas de cribado, sino en mujeres más jóvenes que podrían estar preseleccionadas como de alto riesgo a través de otros cauces, como, por ejemplo, por tener antecedentes familiares o tener biopsias previas.

Las diferencias entre la cohorte analizada y los datos de Pollán para la edad de la menopausia podrían ser debidas al bajo número de mujeres de la cohorte estudiada o a sesgos en la selección de las mujeres en ambas categorías, casos y controles. No obstante, los valores medios de los riesgos para la edad de la menopausia, 1,41 en Pollán y 1,96 en el presente estudio podrían ser considerados como globalmente próximos como para aceptar la inclusión de este factor de riesgo dentro del modelo general multivariable. Estas desviaciones deben corroborarse y estudiarse en profundidad ampliando las cohortes con más mujeres y centros de referencia. Respecto a la densidad mamaria, los resultados basados en la cohorte representativa de la población española corroboran que este factor de riesgo es uno de los más significativos y fuertemente asociados a cáncer de mama esporádico tal como indica la literatura científica. Se puede observar como a medida que la densidad mamaria aumenta, el porcentaje de controles de la cohorte estudiada disminuye y aumenta la de los casos, siendo superior en número a partir de una densidad mamaria superior al 51%. Estos resultados coinciden con los descritos por Pollán anteriormente, aunque el OR obtenido en la cohorte representativa de la población española es ligeramente menor respecto a los OR obtenidos por Pollán. Esta concordancia entre lo obtenido y lo esperado

permite valorar positivamente la inclusión de este factor al modelo multivariable.

Los otros factores reproductivos incorporados en el presente estudio, la edad de la menarquia y la edad del primer nacimiento de un hijo vivo, presentaron valores de significación estadística menores, con pValue de 0,061 y 0,03 respectivamente. Estos resultados tienen una concordancia dispar respecto a los resultados obtenidos por Pollán. Mientras que en la edad de la menarquia no fue significativa en su estudio (pValue 0,35), la edad del nacimiento del primer hijo vivo sí que fue sustancialmente más significativo (pValue <0,001).

Respecto a la edad de la menarquia, la orientación del riesgo obtenido es concordante con lo esperable según la literatura científica donde a cada año que la mujer tiene su primera menstruación antes de la media, el riesgo a padecer cáncer de mama aumenta [111]. Estos resultados corroboran la creencia de que el tiempo de exposición a estrógenos es uno de los factores de riesgo más importantes, de ahí que cuando la mujer tiene la menstruación antes de la edad esperada está expuesta durante más años a esta hormona respecto a la media de la población general. El valor de significación estadística para este factor de riesgo fue relativamente bajo. Este valor indica la posible dispersión observada en la cohorte estudiada donde la tendencia no está clara, sin embargo, la orientación y magnitud de los OR obtenidos, concordantes con los descritos en la literatura científica, reafirman la incorporación de este factor de riesgo al modelo multivariable.

En relación a la edad de la madre al momento del nacimiento de su primer hijo vivo, el valor de significación estadística fue mayor al anteriormente descrito con pValue de 0,03. Según los datos observados, las mujeres que han dado a luz a una edad superior a 30 años o que incluso no han dado a luz presentan un mayor riesgo a padecer cáncer de mama. Esta tendencia se observa con los valores de prevalencia en la cohorte representativa de la población española, donde existe una mayor frecuencia de controles en las mujeres más jóvenes con edad comprendidas entre 20 y 30 años respecto a mujeres de mayor edad siendo la diferencia más elevada en mujeres que no han dado a luz. Esta tendencia coincide perfectamente con los resultados obtenidos por Pollán,

incluso el cambio de tendencia a los 30 años de edad. Estos resultados reafirman la incorporación de este factor de riesgo al modelo multivariable.

En este estudio se validaron positivamente los factores reproductivos estudiados en la cohorte representativa de la población española. Este hecho reafirma que el factor reproductivo, y, por tanto, la exposición a estrógenos es un factor de riesgo importante a padecer cáncer de mama esporádico en dicha población, permitiendo determinar que, a pesar de las desviaciones y sesgos puntuales identificados y descritos que este estudio haya podido tener en la toma de las muestras, es una prueba conceptual valida de esta metodología en dicha población.

En lo que respecta a los antecedentes familiares como factor de riesgo, en el presente estudio, no se obtuvo significación estadística (pValue 0,34). Este resultado no coincide con los obtenidos en el estudio de Pharoah [12]. En dicho estudio se observa una significación estadística importante además de OR elevados para algunas de las categorías. El resultado en la cohorte representativa de la población española puede ser por el bajo número de mujeres analizado, ya que en la mayoría no hay antecedentes (~73 % de los controles y 68 % de los casos), quedando muy bajos de cada categoría descrita para estimarse adecuadamente. Sin embargo, al ser un factor de riesgo respaldado por la comunidad científica, se decidió incluirlo en el modelo multivariable.

Una vez estudiados y corroborados todos los riesgos, tanto genéticos como fenotípicos, a nivel individual se evaluó su posible uso combinado en un modelo más complejo. En los últimos años se han descrito diferentes estrategias que combinan el factor genético con la densidad mamaria o con la edad de la menopausia, pero el uso de modelos con mayor número de variables es algo relativamente novedoso en este contexto. El primer paso es corroborar la no existencia de colinealidad entre las variables, es decir, que dos o más factores de riesgos dan la misma información y son correlacionadas linealmente. El estudio de correlación entre los riesgos no dio ninguna señal positiva significativa, este hecho, indica que todos los factores podrían estar explicando una parte diferente de la variabilidad total del sistema y que, por lo

tanto, un modelo acumulativo podría incrementar la capacidad de estratificación de la cohorte respecto a la obtenida utilizando cada factor de manera individual.

En la selección del método multivariable para el modelo se eligió una regresión logística simple por varios motivos. El primer motivo es que esta regresión es la más referenciada para este tipo de estudios por la comunidad científica el segundo motivo es debido al bajo número de mujeres de la cohorte, donde el uso de métodos más complejos como *"random Forest"* podría generar un valor alto de sobreestimación debido a la baja proporción que se podría tener entre el test de entrenamiento y de testeo. Adicionalmente, el método logístico permite pesar cada riesgo dentro del modelo, además de introducir interacciones entre términos que puede, desde el punto de vista matemático, captar algunas relaciones entre variables que son adecuadas desde el punto de vista clínico, como la relación entre la edad de las mujeres y la densidad mamaria. Estas variables podrían tener una relación natural debido al hecho de que mujeres más jóvenes tendrían tendencia a tener densidad mamaria mayores [112] y, por tanto, en un modelo sin considerar estas relaciones tendría tendencia a sobreestimar el riesgo a padecer cáncer de mama en mujeres jóvenes de manera artificial.

En la cohorte representativa de la población española se obtuvieron con valores estadísticamente significativos dos interacciones, la edad de las mujeres y la densidad mamaria, explicada antes y la edad de la mujer con la edad de la menopausia.

El modelo multivariable, utilizando las variables seleccionadas, fenotípicos y genéticos, presentó mejoras sustanciales y estadísticamente significativas respecto a la capacidad discriminante de cada variable de manera individual, obteniéndose un aumento de AUC medio de 0,74. Estos resultados corroboran la no colinealidad de los riesgos y la capacidad de explicar un porcentaje diferente de la variabilidad de la cohorte estudiada. Es decir, que el método tiene una capacidad discriminante mayor que la suma de los modelos utilizando la estrategia univariable.

Adicionalmente, se evaluó el modelo multivariable añadiendo las interacciones anteriormente mencionadas. Los resultados indican que el modelo es

estadísticamente significativo aumentando la capacidad discriminante a valores de AUC de 0,80 (CI 95%; 0,77-0,83). Esta diferencia fue significativa estadísticamente respecto a los resultados obtenidos utilizando el modelo sin las interacciones. Esa diferencia observada podría indicar que las variables edad de la mujer con la densidad mamaria y la edad de la mujer con la edad de la menopausia están relacionadas y podrían compartir parte de la varianza (covarianza). El uso de esta información dentro del modelo podría ajustarse mejor a las relaciones genéticas y fenotípicas del cáncer de mama esporádico, permitiendo al modelo adaptarse mejor y obtener mejores resultados en la estratificación de la cohorte estudiada.

Estratificando en diferentes segmentos la cohorte representativa de la población española en base al valor obtenido del modelo multivariable con interacciones y normalizado por la mediana del mismo valor en la población general (en este caso los controles), se observó una clara discriminación. Considerando los extremos, se observó cómo el decil <10% de la cohorte se obtenía una prevalencia de casi el 9,39% de los controles respecto a casos, donde se observó valores menores al 0,7 % de los casos, con un OR medio y significativo de 0,097. Esto indica que si una mujer, analizando los factores genéticos y fenotípicos, y aplicando el modelo multivariable con interacciones obtiene un valor normalizado dentro del primer decil, tiene 9 veces más probabilidad de no padecer cáncer de mama esporádico respecto a la población general. Estos resultados, alejándonos del concepto de diagnóstico clínico ya que este estudio está enfocado a una estratificación, son resultados muy aceptables. Los posibles positivos dentro de este decil puede ser debido a diferentes factores, en primer lugar, el cáncer de mama esporádico tiene un cierto carácter probabilístico lo que hace que a pesar de que la mujer tenga menos probabilidad de padecer el fenotipo que la población general es posible que lo desarrolle por componentes aleatorios. Por otro lado, estamos estratificando a la población general usando solo algunos factores de riesgo, pero no todos, algunos como la terapia hormonal podrían ser incorporados en ampliaciones del presente estudio, otros son poco conocidos y/o evaluados por la comunidad científica como tabaquismo o alcoholismo y otros son sencillamente desconocidos (SNPs adicionales, factores epigenéticos etc).

Por otro lado, el último decil, mayor al 90% de la cohorte estudiada, presenta una prevalencia antagónica con un 9% de los casos y 1 % de los controles aproximativamente. Estos resultados indican que las mujeres estratificadas dentro del último decil en función del valor del modelo multivariable usando las dos interacciones tienen 9 veces más probabilidad de padecer cáncer de mama esporádico respecto a la población general. Este resultado podría tener un interés importante para validar el posible uso de este tipo de modelos en rutina, ya que, si una mujer es identificada dentro de este decil, por ejemplo, podría ser considerada como de alto riesgo ya que tendría claramente una mayor probabilidad de padecer cáncer de mama respecto a la población general. La identificación de este tipo de mujeres y la posible aplicación de pautas de seguimiento específicas, como aumentar la frecuencia del estudio mamario a una vez al año, permitiría reducir la probabilidad de aparición de cánceres de intervalo, aumentar la probabilidad de identificación precoz aumentando la tasa de supervivencia de la mujer igual que se reduce la probabilidad de aplicación de posibles acciones física y psicológica invasivas, como la mastectomía o la quimioterapia. Por otro lado, es importante determinar que un 1% de las mujeres clasificadas dentro de este decil no desarrollaron finalmente cáncer de mama en los siguientes 5 años. Este hecho hace que se ponga foco en el control del sobrediagnóstico o la posible tasa de falsos positivos de este tipo de análisis, es decir, identificar que umbrales permiten tener un buen balance de verdaderos positivos frente a los falsos positivos, ya que el sobrediagnóstico puede afectar psicológicamente a las mujeres a lo largo de su vida, además del gasto que conlleva aumentar la frecuencia de los chequeos sin un control estricto. Por este motivo, aumentar el conocimiento de los riesgos y de los modelos para mejorar la precisión en la estratificación de la población general, además de completar las guías y protocolos que permita la gestión psicológica de las mujeres clasificadas como alto riesgo, son de alta importancia en los próximos años para la evaluación en rutina hospitalaria de este tipo de ensayos.

Por otro lado, la propuesta de medidas de seguimiento individuales en las mujeres clasificadas como bajo riesgos genera un mayor nivel de debate en el ámbito hospitalario y clínico, ya que a pesar de que la prevalencia de casos, es decir, de mujeres que desarrollan cáncer, es pequeño en este tipo de deciles

respecto a las mujeres que no lo desarrollan, las evidencias científicas no permiten aún la recomendación fiable del aumento de la frecuencia de los controles mamográficos de estas mujeres respecto lo que le corresponde por edad o por otro tipo de factores en los programas de cribado habituales.

En resumen, los resultados indican que el uso del modelo logístico multivariable y la combinación de variables genéticas, fenotípicas y de interacción es un enfoque eficaz para estratificar a las mujeres de la población española según el riesgo individual de padecer cáncer de mama en un periodo de 5 años, con una capacidad similar a la observada en otros estudios en poblaciones europeas y no europeas. Debido a la naturaleza del estudio, diferentes sesgos podrían haber afectado la precisión de los resultados. Además, el pequeño tamaño de nuestra cohorte podría haber generado un sobreajuste del modelo en términos de estimación del riesgo o la representación excesiva o insuficiente de un tipo de tumor específico. Pero, a pesar de estas limitaciones, este análisis da una prueba de concepto en una población que no ha sido muy estudiada como la española.

Es necesario añadir series más amplias para corroborar los resultados, definir con mayor precisión los umbrales de bajo y alto riesgo además de evaluar el sobrediagnóstico en caso de mujeres de alto riesgo, además de profundizar en el protocolo de las medidas de seguimientos individuales y el control psicológico de los mismos. Hay muchos factores que deben evaluarse y corroborarse para su inclusión rutinaria en los programas de cribado, pero el estudio permite abrir una vía de desarrollo al probar su utilidad en la estratificación de la población general española, siendo una posible base de nuevas investigaciones para los avances necesarios en esta dirección.

Capítulo 6. Conclusiones

1-El cáncer de mama esporádico es una enfermedad. La estrategia de seleccionar SNPs relativamente frecuentes en la población general pero asociados cáncer de mama y combinarlos en un modelo PRS permite una mejor discriminación de las mujeres presentes en la cohorte representativa de la población española en función de su probabilidad a tener cáncer de mama.

2-Además del componente genético, el cáncer de mama esporádico presenta diferentes marcadores de riesgos fenotípicos ampliamente descritos en la literatura científica. Entre los estudiados se encuentran los factores reproductivos, densidad mamaria e historia familiar. En este contexto, estos marcadores han sido corroborados en la cohorte representativa de la población española, tanto en orientación del riesgo como la significancia estadística.

3-La combinación de los factores fenotípicos con el valor genético PRS en un único modelo es capaz de explicar una mayor proporción de mujeres que desarrollaron cáncer, corroborando la naturaleza compleja del cáncer de mama esporádico.

4-Finalmente, el uso de diferentes interacciones entre variables permite obtener los mejores valores discriminantes en la cohorte representativa de la población española. Esas interacciones, podrían absorber mejor la naturaleza compleja entre variables en la descripción del cáncer de mama esporádico, como por ejemplo, la posible dependencia de la densidad mamaria y la edad de la mujer.

Referencias

[1] Offarm El origen genético del cáncer de mama, Vol, 22, Núm, 6, Páginas 108-112 (junio 2003).

[2] Ponti G, De Angelis C, Ponti R, Pongetti L, Losi L, Sticchi A, Tomasi A, Ozben T, Hereditary breast and ovarian cancer: from genes to molecular targeted therapies, Crit Rev Clin Lab Sci, 2023 Jul 16:1-11, doi: 10,1080/10408363,2023,2234488, Epub ahead of print, PMID: 37455374.

[3] Wang L, Early Diagnosis of Breast Cancer, Sensors (Basel), 2017 Jul 5;17(7):1572, doi: 10,3390/s17071572, PMID: 28678153; PMCID: PMC5539491,

[4] Görner M, Just M, Gerull S, Mammakarzinom im lokalisierten Stadium - Strategien für die systemisch-adjuvante Therapie [Early-stage breast cancer - strategies for adjuvant systemic therapy], Handchir Mikrochir Plast Chir, 2008 Aug;40(4):230-8, German, doi: 10,1055/s-2008-1038926, Epub 2008 Aug 20, PMID: 18716990

[5] U,S, Population Data 1969-2019 with Other Software: (downloaded from SEER Web site) Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2019) (www,seer,cancer,gov/popdata), National Cancer Institute, DCCPS, Surveillance Research Program, released February 2021.

[6] Independent UK Panel on Breast Cancer Screening, The benefits and harms of breast cancer screening: an independent review, Lancet, 2012 Nov 17;380(9855):1778-86, doi: 10,1016/S0140-6736(12)61611-0, Epub 2012 Oct 30, PMID: 23117178.

[7] Pashayan N, Morris S, Gilbert FJ, Pharoah PDP, Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model, JAMA Oncol, 2018 Nov 1;4(11):1504-1510, doi: 10,1001/jamaoncol,2018,1901, Erratum in: JAMA Oncol, 2022 Mar 1;8(3):484, PMID: 29978189; PMCID: PMC6230256.

[8] Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al, False-positive results in mammographic screening for breast cancer in Europe: a

literature review and survey of service screening programmes, J Med Screen, 2012;19 Suppl 1:57-66, doi: 10,1258/jms,2012,012083, PMID: 22972811.

[9] Tice JA, Miglioretti DL, Li CS, Vachon CM, Gard CC, Kerlikowske K, Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer, J Clin Oncol, 2015 Oct 1;33(28):3137-43, doi: 10,1200/JCO,2015,60,8869, Epub 2015 Aug 17, PMID: 26282663; PMCID: PMC4582144.

[10] Zheng T, Holford TR, Mayne ST, Owens PH, Zhang Y, Zhang B, Boyle P, Zahm SH, Lactation and breast cancer risk: a case-control study in Connecticut, Br J Cancer, 2001 Jun 1;84(11):1472-6, doi: 10,1054/bjoc,2001,1793, PMID: 11384096; PMCID: PMC2363665.

[11] Hsieh CC, Trichopoulos D, Katsouyanni K, Yuasa S, Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study, Int J Cancer, 1990 Nov 15;46(5):796-800, doi: 10,1002/ijc,2910460508, PMID: 2228308.

[12] Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA, Family history and the risk of breast cancer: a systematic review and meta-analysis, Int J Cancer, 1997 May 29;71(5):800-9, doi: 10,1002/(sici)1097-0215(19970529)71:5<800:aid-ijc18>3,0,co;2-b, PMID: 9180149.

[13] Hiatt RA, Brody JG, Environmental Determinants of Breast Cancer, Annu Rev Public Health, 2018 Apr 1;39:113-133, doi: 10,1146/annurev-publhealth-040617-014101, Epub 2018 Jan 12, PMID: 29328875.

[14] Voutsadakis IA. Vitamin D baseline levels at diagnosis of breast cancer: A systematic review and meta-analysis. Hematol Oncol Stem Cell Ther. 2021 Mar;14(1):16-26. doi: 10.1016/j.hemonc.2020.08.005. Epub 2020 Sep 26. PMID: 33002425.

[15] Videnros C, Selander J, Wiebert P, Albin M, Plato N, Borgquist S, Manjer J, Gustavsson P, Investigating the risk of breast cancer among women exposed to chemicals: a nested case-control study using improved exposure estimates, Int Arch Occup Environ Health, 2020 Feb;93(2):261-269, doi: 10,1007/s00420-019-01479-4, Epub 2019 Oct 24, PMID: 31650237; PMCID: PMC7007902.

[16] Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K, Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland, N Engl J Med, 2000 Jul 13;343(2):78-85, doi: 10,1056/NEJM200007133430201, PMID: 10891514.

[17] Petrucelli N, Daly MB, Pal T, BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer, 1998 Sep 4 [updated 2022 May 26], In: Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, Gripp KW, Amemiya A, editors, GeneReviews® [Internet], Seattle (WA): University of Washington, Seattle; 1993–2023, PMID: 20301425.

[18] Ghoussaini M, Pharoah PDP, Easton DF, Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? Am J Pathol, 2013 Oct;183(4):1038-1051, doi: 10,1016/j,ajpath,2013,07,003, Epub 2013 Aug 23, PMID: 23973388.

[19] Yoshida K, Miki Y; Miki (November 2004), "Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage", Cancer Science, 95 (11): 866–871, doi:10,1111/j,1349-7006,2004,tb02195, x, PMID 15546503, S2CID 24297965.

[20] Friedenson B, BRCA1 and BRCA2 pathways and the risk of cancers other than breast or ovarian, MedGenMed, 2005 Jun 29;7(2):60, PMID: 16369438; PMCID: PMC1681605.

[21] Wilcox N, Dumont M, González-Neira A, Carvalho S, Joly Beauparlant C, Crotti M, et al, Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk, Nat Genet, 2023 Sep;55(9):1435-1439, doi: 10,1038/s41588-023-01466-z, Epub 2023 Aug 17, Erratum in: Nat Genet, 2023 Sep 26;: PMID: 37592023; PMCID: PMC10484782.

[22] Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DG, Chenevix-Trench G, Rahman N, Robson M, Domchek SM, Foulkes WD. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med.

2015 Jun 4;372(23):2243-57. doi: 10.1056/NEJMsr1501341. Epub 2015 May 27. PMID: 26014596; PMCID: PMC4610139.

[23] Broeks A, Urbanus JH, Floore AN, Dahler EC, Klijn JG, Rutgers EJ, Devilee P, Russell NS, van Leeuwen FE, van 't Veer LJ. ATM-heterozygous germline mutations contribute to breast cancer-susceptibility. Am J Hum Genet. 2000 Feb;66(2):494-500. doi: 10.1086/302746. PMID: 10677309; PMCID: PMC1288102.

[24] Gallagher S, Hughes E, Kurian AW, Domchek SM, Garber J, Probst B, Morris B, Tshiaba P, Meek S, Rosenthal E, Roa B, Slavin TP, Wagner S, Weitzel J, Gutin A, Lanchbury JS, Robson M, Comprehensive Breast Cancer Risk Assessment for CHEK2 and ATM Pathogenic Variant Carriers Incorporating a Polygenic Risk Score and the Tyrer-Cuzick Model, JCO Precis Oncol, 2021 Jun 24;5:PO,20,00484, doi: 10,1200/PO,20,00484, PMID: 34322652; PMCID: PMC8238281.

[25] Gallagher S, Hughes E, Wagner S, Tshiaba P, Rosenthal E, Roa BB, Kurian AW, Domchek SM, Garber J, Lancaster J, Weitzel JN, Gutin A, Lanchbury JS, Robson M, Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes, JAMA Netw Open, 2020 Jul 1;3(7): e208501, doi: 10,1001/jamanetworkopen,2020,8501, PMID: 32609350; PMCID: PMC7330720.

[26] Antoniou AC, Easton DF. Models of genetic susceptibility to breast cancer. Oncogene. 2006 Sep 25;25(43):5898-905. doi: 10.1038/sj.onc.1209879. PMID: 16998504.

[27] Mars N, Widén E, Kerminen S, Meretoja T, Pirinen M, Della Briotta Parolo P, Palta P; FinnGen, Palotie A, Kaprio J, Joensuu H, Daly M, Ripatti S, The role of polygenic risk and susceptibility genes in breast cancer over the course of life, Nat Commun, 2020 Dec 14;11(1):6383, doi: 10,1038/s41467-020-19966-5, PMID: 33318493; PMCID: PMC7736877.

[28] Dudbridge F (March 2013), "Power and predictive accuracy of polygenic risk scores", PLoS Genetics, 9 (3): e1003348, doi:10,1371/journal,pgen,1003348, PMC 3605113, PMID 23555274.

[29] Torkamani A, Wineinger NE, Topol EJ, The personal and clinical utility of polygenic risk scores, Nat Rev Genet, 2018 Sep;19(9):581-590, doi: 10,1038/s41576-018-0018-x, PMID: 29789686.

[30] Lello L, Raben TG, Yong SY, Tellier LC, Hsu SD (October 2019), "Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer", Scientific Reports, 9 (1): 15286, Bibcode:2019NatSR,,915286L, doi:10,1038/s41598-019-51258-x, PMC 6814833, PMID 31653892.

[31] Witte JS, Hoffmann TJ, Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer, OMICS, 2011 Jun;15(6):393-8, doi: 10,1089/omi,2010,0090, Epub 2011 Feb 24, PMID: 21348634; PMCID: PMC3125548.

[32] Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, Li J, Ho PJ, et al, European polygenic risk score for prediction of breast cancer shows similar performance in Asian women, Nat Commun, 2020 Jul 31;11(1):3833, doi: 10,1038/s41467-020-17680-w, PMID: 32737321; PMCID: PMC7395776.

[33] Miñambres R, Pita G, Sanchez I, Serra D, Rosar R, Rubio-Solsona E, Palacios S, Llaneza A, Marron P, Hoyas S, Lluch A, Cano A, Gonzalez A, Triviño JC, Benítez J, Prediction of cancer risk based on study of genetic variants in healthy women in the Spanish population, Rev Senol Patol Mamar, 2019;32(3):94–9, https://doi,org/10,1016/j,senol, 2019,07,001.

[34] Ho WK, Tai MC, Dennis J, Shu X, Li J, et al , Polygenic risk scores for prediction of breast cancer risk in Asian populations, Genet Med, 2022 Mar;24(3):586-600, doi: 10,1016/j,gim,2021,11,008, Epub 2021 Dec 15, PMID: 34906514; PMCID: PMC7612481.

[35] Evans DG, van Veen EM, Byers H, Roberts E, Howell A, Howell SJ, Harkness EF, Brentnall A, Cuzick J, Newman WG, The importance of ethnicity: Are breast cancer polygenic risk scores ready for women who are not of White

European origin? Int J Cancer, 2022 Jan 1;150(1):73-79, doi: 10,1002/ijc,33782, Epub 2021 Sep 7, PMID: 34460111.

[36] Siegel RL, Miller KD, Fuchs HE, Jemal A, Cancer Statistics, 2021, CA Cancer J Clin, 2021 Jan;71(1):7-33, doi: 10,3322/caac,21654, Epub 2021 Jan 12, Erratum in: CA Cancer J Clin, 2021 Jul;71(4):359, PMID: 33433946.

[37] Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E, Breast cancer risk factors, Prz Menopauzalny, 2015 Sep;14(3):196-202, doi: 10,5114/pm,2015,54346, Epub 2015 Sep 30, PMID: 26528110; PMCID: PMC4612558.

[38] Berg WA, Tailored supplemental screening for breast cancer: what now and what next? AJR Am J Roentgenol, 2009 Feb;192(2):390-9, doi: 10,2214/AJR,08,1706, PMID: 19155400.

[39] Berg WA, Beyond standard mammographic screening: mammography at age extremes, ultrasound, and MR imaging, Radiol Clin North Am, 2007 Sep;45(5):895-906, vii, doi: 10,1016/j,rcl,2007,06,001, PMID: 17888776,.

[40] Al-Shami K, Awadi S, Khamees A, Alsheikh AM, Al-Sharif S, Ala' Bereshy R, Al-Eitan SF, Banikhaled SH, Al-Qudimat AR, Al-Zoubi RM, Al Zoubi MS, Estrogens and the risk of breast cancer: A narrative review of literature, Heliyon, 2023 Sep 17;9(9):e20224, doi: 10,1016/j,heliyon,2023,e20224, PMID: 37809638; PMCID: PMC10559995.

[41] Sickles, EA, D'Orsi CJ, Bassett LW, et al, ACR BI-RADS® Mammography, In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, Reston, VA, American College of Radiology; 2013.

[42] Bertrand KA, Tamimi RM, Scott CG, Jensen MR, Pankratz V, Visscher D, Norman A, Couch F, Shepherd J, Fan B, Chen YY, Ma L, Beck AH, Cummings SR, Kerlikowske K, Vachon CM, Mammographic density and risk of breast cancer by age and tumor characteristics, Breast Cancer Res, 2013 Nov 4;15(6):R104, doi: 10,1186/bcr3570, PMID: 24188089; PMCID: PMC3978749.

[43] Dupont WD, Page DL, Risk factors for breast cancer in women with proliferative breast disease, N Engl J Med, 1985 Jan 17;312(3):146-51, doi: 10,1056/NEJM198501173120303, PMID: 3965932.

[44] Neal L, Sandhu NP, Hieken TJ, Glazebrook KN, Mac Bride MB, Dilaveri CA, Wahner-Roedler DL, Ghosh K, Visscher DW, Diagnosis and management of benign, atypical, and indeterminate breast lesions detected on core needle biopsy, Mayo Clin Proc, 2014 Apr;89(4):536-47, doi: 10,1016/j,mayocp,2014,02,004, PMID: 24684875.

[45] Haagensen CD, Lane N, Lattes R, Bodian C, Lobular neoplasia (so-called lobular carcinoma in situ) of the breast, Cancer, 1978 Aug;42(2):737-69, doi: 10,1002/1097-0142(197808)42:2<737::aid-cncr2820420247>3,0,co;2-t, PMID: 209887.

[46] Wong SM, Stout NK, Punglia RS, Prakash I, Sagara Y, Golshan M, Breast cancer prevention strategies in lobular carcinoma in situ: A decision analysis, Cancer, 2017 Jul 15;123(14):2609-2617, doi: 10,1002/cncr,30644, Epub 2017 Feb 21, PMID: 28221673.

[47] Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS, Validation studies for models projecting the risk of invasive and total breast cancer incidence, J Natl Cancer Inst, 1999 Sep 15;91(18):1541-8, doi: 10,1093/jnci/91,18,1541, PMID: 10491430.

[48] General Assembly of the World Medical Association, World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects, J Am Coll Dent, 2014 Summer;81(3):14-8, PMID: 25951678.

[49] Gao J, Warren R, Warren-Forward H, Forbes JF, Reproducibility of visual assessment on mammographic density, Breast Cancer Res Treat, 2008 Mar;108(1):121-7, doi: 10,1007/s10549-007-9581-0, Epub 2007 Jul 7, PMID: 17616811.

[50] Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al, Large-scale genotyping identifies 41 new loci associated with breast cancer risk, Nat Genet, 2013 Apr;45(4):353-61, 361e1-2, doi: 10,1038/ng,2563, PMID: 23535729; PMCID: PMC3771688.

[51] Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al, The OncoArray Consortium: A Network for Understanding the Genetic

Architecture of Common Cancers, Cancer Epidemiol Biomarkers Prev, 2017 Jan;26(1):126-135, doi: 10,1158/1055-9965,EPI-16-0106, Epub 2016 Oct 3, PMID: 27697780; PMCID: PMC5224974.

[52] Jernström H, Polygenes, risk prediction, and targeted prevention of breast cancer, N Engl J Med, 2008 Sep 25;359(13):1407, PMID: 18822457.

[53] Pollán M, Ascunce N, Ederra M, Murillo A, Erdozáin N, Alés-Martínez J, Pastor-Barriuso R, Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: a Spanish population-based case-control study, Breast Cancer Res, 2013 Jan 29;15(1):R9, doi: 10,1186/bcr3380, PMID: 23360535; PMCID: PMC3672793.

[54] Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al, Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States, JAMA Oncol, 2016 Oct 1;2(10):1295-1302, doi: 10,1001/jamaoncol,2016,1025, Erratum in: JAMA Oncol, 2016 Oct 1;2(10):1374, PMID: 27228256; PMCID: PMC5719876.

[55] Hughes J, reghelper: helper functions for regression analysis, R package version 0,3,5; 2020.

[56] Draper NR, Smith H, Applied regression analysis, 3rd ed, New York: Wiley;1998.

[57] Hosmer DW, Lemeshow S, Sturdivant RX, Applied logistic regression: third edition; 2013.

[58] Rosner B, Glynn RJ. Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. Biometrics. 2009 Mar;65(1):188-97. doi: 10.1111/j.1541-0420.2008.01062.x. Epub 2008 May 28. PMID: 18510654.

[59] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12:77. doi: 10.1186/1471-2105-12-77. PMID: 21414208; PMCID: PMC3068975.

[60] Tworoger SS, Zhang X, Eliassen AH, Qian J, Colditz GA, Willett WC, Rosner BA, Kraft P, Hankinson SE. Inclusion of endogenous hormone levels in risk prediction models of postmenopausal breast cancer. J Clin Oncol. 2014 Oct 1;32(28):3111-7. doi: 10.1200/JCO.2014.56.1068. Epub 2014 Aug 18. PMID: 25135988; PMCID: PMC4171356.

[61] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F, The Ensembl Variant Effect Predictor, Genome Biol, 2016 Jun 6;17(1):122, doi: 10,1186/s13059-016-0974-4, PMID: 27268795; PMCID: PMC4893825.

[62] Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H, PANTHER: Making genome-scale phylogenetics accessible to all, Protein Sci, 2022 Jan;31(1):8-22, doi: 10,1002/pro,4218, Epub 2021 Nov 25, PMID: 34717010; PMCID: PMC8740835.

[63] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A, Pfam: The protein families database in 2021, Nucleic Acids Res, 2021 Jan 8;49(D1): D412-D419, doi: 10,1093/nar/gkaa913, PMID: 33125078; PMCID: PMC7779014.

[64] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2023 Dec 6. doi: 10.1038/s41586-023-06045-0. Epub ahead of print. PMID: 38057664.

[65] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062-D1067. doi: 10.1093/nar/gkx1153. PMID: 29165669; PMCID: PMC5753237.

[66] Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC, SIFT missense predictions for genomes, Nat Protoc, 2016 Jan;11(1):1-9, doi: 10,1038/nprot,2015,123, Epub 2015 Dec 3, PMID: 26633127,

[67] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76. PMID: 23315928; PMCID: PMC4480630.

[68] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019 Jan 8;47(D1):D419-D426. doi: 10.1093/nar/gky1038. PMID: 30407594; PMCID: PMC6323939.

[69] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M, KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Res, 2023 Jan 6;51(D1): D587-D592, doi: 10,1093/nar/gkac963, PMID: 36300620; PMCID: PMC9825424.

[70] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C, Deng C, Varusai T, Ragueneau E, Haider Y, May B, Shamovsky V, Weiser J, Brunson T, Sanati N, Beckman L, Shao X, Fabregat A, Sidiropoulos K, Murillo J, Viteri G, Cook J, Shorser S, Bader G, Demir E, Sander C, Haw R, Wu G, Stein L, Hermjakob H, D'Eustachio P, The reactome pathway knowledgebase 2022, Nucleic Acids Res, 2022 Jan 7;50(D1): D687-D692, doi: 10,1093/nar/gkab1028, PMID: 34788843; PMCID: PMC8689983.

[71] Chen J, Bardes EE, Aronow BJ, Jegga AG, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res, 2009 Jul;37(Web Server issue): W305-11, doi: 10,1093/nar/gkp427, Epub 2009 May 22, PMID: 19465376; PMCID: PMC2703978.

[72] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res, 2003 Nov;13(11):2498-504, doi: 10,1101/gr,1239303, PMID: 14597658; PMCID: PMC403769.

[73] Ma'ayan A, Introduction to network analysis in systems biology, Sci Signal, 2011 Sep 6;4(190):tr5, doi: 10,1126/scisignal,2001965, PMID: 21917719; PMCID: PMC3196357.

[74] Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JWT, Tice JA, Ziv E, Kerlikowske K, Cummings SR, Joint relative risks for estrogen receptor-positive breast cancer from a clinical model, polygenic risk score, and sex hormones, Breast Cancer Res Treat, 2017 Nov;166(2):603-612, doi: 10,1007/s10549-017-4430-2, Epub 2017 Aug 8, PMID: 28791495; PMCID: PMC5669824.

[75] Dite GS, MacInnis RJ, Bickerstaffe A, Dowty JG, Allman R, Apicella C, Milne RL, Tsimiklis H, Phillips KA, Giles GG, Terry MB, Southey MC, Hopper JL, Breast Cancer Risk Prediction Using Clinical Models and 77 Independent Risk-Associated SNPs for Women Aged Under 50 Years: Australian Breast Cancer Family Registry, Cancer Epidemiol Biomarkers Prev, 2016 Feb;25(2):359-65, doi: 10,1158/1055-9965,EPI-15-0838, Epub 2015 Dec 16, PMID: 26677205; PMCID: PMC4767544.

[76] Gao G, Zhao F, Ahearn TU, Lunetta KL, Troester MA, Du Z, Ogundiran TO et al, Polygenic risk scores for prediction of breast cancer risk in women of African ancestry: a cross-ancestry approach, Hum Mol Genet, 2022 Sep 10;31(18):3133-3143, doi: 10,1093/hmg/ddac102, PMID: 35554533; PMCID: PMC9476624.

[77] Yang Y, Tao R, Shu X, Cai Q, Wen W, Gu K, Gao YT, Zheng Y, Kweon SS, Shin MH, Choi JY, Lee ES, Kong SY, Park B, Park MH, Jia G, Li B, Kang D, Shu XO, Long J, Zheng W, Incorporating Polygenic Risk Scores and Nongenetic Risk Factors for Breast Cancer Risk Prediction Among Asian Women, JAMA Netw Open, 2022 Mar 1;5(3):e2149030, doi: 10,1001/jamanetworkopen,2021,49030, PMID: 35311964; PMCID: PMC8938714.

[78] Choi SW, Mak TS, O'Reilly PF, Tutorial: a guide to performing polygenic risk score analyses, Nat Protoc, 2020 Sep;15(9):2759-2772, doi: 10,1038/s41596-020-0353-1, Epub 2020 Jul 24, PMID: 32709988; PMCID: PMC7612115.

[79] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, Genet Med, 2015 May;17(5):405-24, doi: 10,1038/gim,2015,30, Epub 2015 Mar 5, PMID: 25741868; PMCID: PMC4544753.

[80] Mathelier A, Wasserman WW, The next generation of transcription factor binding site prediction, PLoS Comput Biol, 2013;9(9):e1003214, doi: 10,1371/journal,pcbi,1003214, Epub 2013 Sep 5, PMID: 24039567; PMCID: PMC3764009.

[81] Steinhaus R, Robinson PN, Seelow D, FABIAN-variant: predicting the effects of DNA variants on transcription factor binding, Nucleic Acids Res, 2022 Jul 5;50(W1):W322-W329, doi: 10,1093/nar/gkac393, PMID: 35639768; PMCID: PMC9252790.

[82] Beck T, Rowlands T, Shorter T, Brookes AJ, GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies, Nucleic Acids Res, 2023 Jan 6;51(D1): D986-D993, doi: 10,1093/nar/gkac1017, PMID: 36350644; PMCID: PMC9825503.

[83] Chakravarty D, Solit DB, Clinical cancer genomic profiling, Nat Rev Genet, 2021 Aug;22(8):483-501, doi: 10,1038/s41576-021-00338-8, Epub 2021 Mar 24, PMID: 33762738.

[84] Yanes T, Young MA, Meiser B, James PA, Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field, Breast Cancer Res, 2020 Feb 17;22(1):21, doi: 10,1186/s13058-020-01260-3, PMID: 32066492; PMCID: PMC7026946.

[85] Surakasula A, Nagarjunapu GC, Raghavaiah KV, A comparative study of pre- and post-menopausal breast cancer: Risk factors, presentation, characteristics and management, J Res Pharm Pract, 2014 Jan;3(1):12-8, doi: 10,4103/2279-042X,132704, PMID: 24991630; PMCID: PMC4078652.

[86] Preston-Martin S, Pike MC, Ross RK, Henderson BE, Epidemiologic evidence for the increased cell proliferation model of carcinogenesis, Environ Health Perspect, 1993 Dec;101 Suppl 5(Suppl 5):137-8, doi: 10,1289/ehp,93101s5137, PMID: 8013400; PMCID: PMC1519434.

[87] Preston-Martin S, Pike MC, Ross RK, Jones PA, Henderson BE, Increased cell division as a cause of human cancer, Cancer Res, 1990 Dec 1;50(23):7415-21, PMID: 2174724.

[88] Yue W, Wang JP, Li Y, Fan P, Liu G, Zhang N, Conaway M, Wang H, Korach KS, Bocchinfuso W, Santen R, Effects of estrogen on breast cancer development: Role of estrogen receptor independent mechanisms, Int J Cancer, 2010 Oct 15;127(8):1748-57, doi: 10,1002/ijc,25207, PMID: 20104523; PMCID: PMC4775086.

[89] MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S, Age at first birth and breast cancer risk, Bull World Health Organ, 1970;43(2):209-21, PMID: 5312521; PMCID: PMC2427645.

[90] Lowe CR, MacMahon B, Breast cancer and reproductive history of women in South Wales, Lancet, 1970 Jan 24;1(7639):153-6, doi: 10,1016/s0140-6736(70)90401-0, PMID: 4189235.

[91] Albrektsen G, Heuch I, Hansen S, Kvåle G, Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects, Br J Cancer, 2005 Jan 17;92(1):167-75, doi: 10,1038/sj,bjc,6602302, PMID: 15597097; PMCID: PMC2361726.

[92] Lambe M, Hsieh C, Trichopoulos D, Ekbom A, Pavia M, Adami HO, Transient increase in the risk of breast cancer after giving birth, N Engl J Med, 1994 Jul 7;331(1):5-9, doi: 10,1056/NEJM199407073310102, PMID: 8202106.

[93] Rosner B, Colditz GA, Willett WC, Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study, Am J Epidemiol, 1994 Apr 15;139(8):819-35, doi: 10,1093/oxfordjournals,aje,a117079, PMID: 8178795.

[94] Boyd NF, Mammographic density and risk of breast cancer, Am Soc Clin Oncol Educ Book, 2013, doi: 10,1200/EdBook_AM,2013,33,e57, PMID: 23714456.

[95] Nazari SS, Mukherjee P, An overview of mammographic density and its association with breast cancer, Breast Cancer, 2018 May;25(3):259-267, doi: 10,1007/s12282-018-0857-5, Epub 2018 Apr 12, PMID: 29651637; PMCID: PMC5906528.

[96] Sickles, EA, D'Orsi CJ, Bassett LW, et al, ACR BI-RADS® Mammography, In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, Reston, VA, American College of Radiology; 2013.

[97] Gemici AA, Bayram E, Hocaoglu E, Inci E, Comparison of breast density assessments according to BI-RADS 4th and 5th editions and experience level, Acta Radiol Open, 2020 Jul 20;9(7):2058460120937381, doi: 10,1177/2058460120937381, PMID: 32733694; PMCID: PMC7372628.

[98] Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S, Yaffe MJ, Mammographic density and the risk and detection of breast cancer, N Engl J Med, 2007 Jan 18;356(3):227-36, doi: 10,1056/NEJMoa062790, PMID: 17229950.

[99] Fletcher SW, Elmore JG, Clinical practice, Mammographic screening for breast cancer, N Engl J Med, 2003 Apr 24;348(17):1672-80, doi: 10,1056/NEJMcp021804, PMID: 12711743; PMCID: PMC3157308.

[100] Wolfe JN, Risk for breast cancer development determined by mammographic parenchymal pattern, Cancer, 1976 May;37(5):2486-92, doi: 10,1002/1097-0142(197605)37:5<2486::aid-cncr2820370542>3,0,co;2-8, PMID: 1260729, /1097-0142(197605) 37:5<2486::AID-CNCR2820370542>3,0,CO;2-8.

[101] Wolfe JN, Breast patterns as an index of risk for developing breast cancer, AJR Am J Roentgenol, 1976 Jun;126(6):1130-7, doi: 10,2214/ajr,126,6,1130, PMID: 179369.

[102] McCormack VA, dos Santos Silva I, Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis, Cancer Epidemiol

Biomarkers Prev, 2006 Jun;15(6):1159-69, doi: 10,1158/1055-9965,EPI-06-0034, PMID: 16775176.

[103] Ursin G, Lillie EO, Lee E, Cockburn M, Schork NJ, Cozen W, Parisky YR, Hamilton AS, Astrahan MA, Mack T, The relative importance of genetics and environment on mammographic density, Cancer Epidemiol Biomarkers Prev, 2009 Jan;18(1):102-12, doi: 10,1158/1055-9965,EPI-07-2857, PMID: 19124487.

[104] Li T, Sun L, Miller N, Nicklee T, Woo J, Hulse-Smith L, Tsao MS, Khokha R, Martin L, Boyd N, The association of measured breast tissue characteristics with mammographic density and other risk factors for breast cancer, Cancer Epidemiol Biomarkers Prev, 2005 Feb;14(2):343-9, doi: 10,1158/1055-9965,EPI-04-0490, PMID: 15734956.

[105] del Carmen MG, Halpern EF, Kopans DB, Moy B, Moore RH, Goss PE, Hughes KS, Mammographic breast density and race, AJR Am J Roentgenol, 2007 Apr;188(4):1147-50, doi: 10,2214/AJR,06,0619, PMID: 17377060.

[106] Boletín Oficial del Estado, Real Decreto 1030/2006, de 15 de septiembre, por el que se establece la cartera de servicios comunes del Sistema Nacional de Salud y el procedimiento para su actualización, BOE núm, 222, de 16-9-2006.

[107] Ascunce N, Ederra M, Barcos A, Zubizarreta R, Fernández AB, Casamitjana M, Situación del cribado de cáncer de mama en España: características y principales resultados de los programas existentes, En: Castells X, Sala M, Ascunce N, Salas D, Zubizarreta R, Casamitjana M, eds, Descripción del Cribado del Cáncer en España Proyecto DESCRIC, Madrid: Agència d´Avaluació de Tecnologia i Recerca Mèdiques; 2007, Informes de Evaluación de Tecnologias Sanitarias: AATRM 2006/01, p, 31-73.

[108] Zuleika Saz-Parkinson, Olga Monteagudo-Piqueras, Joaquin Granados Ortega, Encarnación Martínez Mondéjar y Mª Vicenta Labrador Cañadas (3) “European commission initiative on breast cancer”. Recomendaciones seleccionadas de cribado de cancer de mama de las guias europeas. Rev Esp Salud Pública, 2020; Vol, 94.

[109] Sanchez Gómez S, Torres Tabanera M, Vega Bolivar A, Sainz Miranda M, Baroja Mazo A, Ruiz Diaz M, Martinez Miravete P, Lag Asturiano E, Muñoz Cacho P, Delgado Macias T, Impact of a CAD system in a screen-film mammography screening program: a prospective study, Eur J Radiol, 2011 Dec;80(3):e317-21, doi: 10,1016/j,ejrad,2010,08,031, Epub 2010 Sep 22, PMID: 20863639.

[110] Blanch J, Sala M, Ibáñez J, Domingo L, Fernandez B, Otegi A, Barata T, Zubizarreta R, Ferrer J, Castells X, Rué M, Salas D; INCA Study Group, Impact of risk factors on different interval cancer subtypes in a population-based breast cancer screening programme, PLoS One, 2014 Oct 21;9(10):e110207, doi: 10,1371/journal,pone,0110207, PMID: 25333936; PMCID: PMC4204862.

[111] Collaborative Group on Hormonal Factors in Breast Cancer, Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies, Lancet Oncol, 2012 Nov;13(11):1141-51, doi: 10,1016/S1470-2045(12)70425-4, Epub 2012 Oct 17, PMID: 23084519; PMCID: PMC3488186.

[112] Kang YJ, Ahn SK, Kim SJ, Oh H, Han J, Ko E, Relationship between Mammographic Density and Age in the United Arab Emirates Population, J Oncol, 2019 Aug 5; 2019:7351350, doi: 10,1155/2019/7351350, PMID: 31467543; PMCID: PMC6701291.