UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

MÉTODOS Y HERRAMIENTAS BIOINFORMÁTICAS: CONTRIBUCIONES A LA MEJORA DEL DIAGNÓSTICO Y LA GESTIÓN DE PACIENTES

D. Alejandro Cisterna García

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

MÉTODOS Y HERRAMIENTAS BIOINFORMÁTICAS: CONTRIBUCIONES A LA MEJORA DEL DIAGNÓSTICO Y LA GESTIÓN DE PACIENTES

Autor: D. Alejandro Cisterna García

Director/es: D. Juan Antonio Botía Blaya y D. Paolo Maietta

UNIVERSIDAD DE MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR
Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Alejandro Cisterna García

doctorando del Programa de Doctorado en

Envejecimiento y Fragilidad

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Métodos y herramientas bioinformáticas: contribuciones a la mejora del diagnóstico y la gestión de pacientes

y dirigida por,

D./Dña. Juan Antonio Botía Blaya

D./Dña. Paolo Maietta

D./Dña.

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Si la tesis hubiera sido autorizada como tesis por compendio de publicaciones o incluyese 1 o 2 publicaciones (como prevé el artículo 29.8 del reglamento), declarar que cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 26 de julio de 2023

Firmado por CISTERNA GARCIA ALEJANDRO - ***6364** el día 26/07/2023 con un certificado emitido por AC FNMT Usuarios

Fdo.: Alejandro Cisterna García

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Información básica sobre protección de sus datos personales aportados	
Responsable:	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación:	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad:	Gestionar su declaración de autoría y originalidad
Destinatarios:	No se prevén comunicaciones de datos
Derechos:	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Dedicatoria y Reconocimientos

Un trabajo como una tesis no es posible sin el esfuerzo y dedicación de muchas personas a lo largo del tiempo. Por esta razón, quiero expresar mi gratitud a aquellos que han contribuido de alguna manera a su realización.

En primer lugar, quiero dar las gracias a mi director de tesis, Juan A. Botía, por su paciencia, tiempo y conocimiento. He tenido la suerte de contar contigo primero como profesor y luego como mentor. Tu guía, dedicación y enseñanzas tienen un gran valor para mí y tendrán un gran efecto lo largo de mi carrera.

Agradezco también a Laura Ibáñez, mi supervisora en la Washington University de St. Louis, por su apoyo, estímulo y enseñanzas durante mi estancia allí. Gracias por tu forma de transmitirme nuevos conocimientos y por tu manera de hacer las cosas. Con mujeres como tú la ciencia y la sociedad tienen mucho que ganar.

Quiero agradecer a la gente de NIMGenetics S.L., implicada en esta tesis como empresa cofinanciadora, por su compromiso con la ciencia, su conocimiento y su generosidad en compartirlo. En especial a Paolo Maietta, cotutor de esta tesis, pero sin olvidar la suerte de contar con Sara Álvarez, María Hoyos, Javier Suela e Irene Diez, entre otros.

Agradezco al Servicio Murciano de Salud, especialmente a Manuel Escudero Sánchez, Francisco de Francisco Verdú y a Ramón Rodríguez Iborra, por ayudarnos y proporcionarnos datos clínicos utilizados en esta tesis.

Agradezco a todas las fuentes de datos por sus contribuciones y, especialmente, a quienes nos han permitido utilizar sus logotipos e imágenes en esta publicación.

A mis coautores en los trabajos publicados, les agradezco por hacer ciencia juntos.

A mis compañeros de facultad y de estancia: Aurora, Alicia, Antonio, José Luis, Enrique, Fran, Jesús, Kiril, María, Horacio, Juan, Adelaida..., les agradezco por los incontables momentos de diversión, risas y complicidad.

A mis amigos, estoy inmesamente agradecido por tenerlos en mi vida: José Alberto, José Manuel, Alejandro, David, Antonio, Marcial, Manuel, Juanjo...

A mi pareja, Pilar, gracias por estar a mi lado en los momentos buenos y malos durante este proceso. Tu sonrisa siempre ha sido un rayo de luz en los días más difíciles.

A mi madre, agradezco su amor incondicional, apoyo y confianza. Gracias por darme la oportunidad de llegar hasta aquí.

A mi padre, aunque no esté aquí, le agradezco por enseñarme tanto.

A todos vosotros, no tengo forma de expresar con palabras la profunda gratitud que siento por toda vuestra ayuda.

TABLA DE CONTENIDOS

LISTA DE TABLAS

TABLA **Página**

LISTA DE FIGURAS

FIGURA **Página**

Acronyms

AD Alzheimer Disease

AI Artificial Intelligence

BAM Binary Alignment Map

BMI Body Mass Index

BWA Burrows-Wheeler Aligner

CDC Centers from Disease Control and Prevention

CDR Clinical Dementia Rating

cfRNA cell-free RNA

CGI The Cancer Genome Interpreter

CI Confidence Interval

ClinGen The Clinical Genome Resource

COPD Chronic Obstructive Pulmonary Disease

COVID-19 Coronavirus Disease 2019

CPU Central Processing Unit

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

CSF Cerebrospinal fluid

CSV Comma-Separated Values

CTD The Comparative Toxicogenomics Database

DECIPHER DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources

DNA Deoxyribonucleic Acid

EHR Electronic Health Record

EOD Early Onset Dystonia

FDA U.S. Food and Drug Administration

FDR False Discovery Rate

FI Feature Importance

FTD Frontotemporal Dementia

GATK Genome Analysis Toolkit

GDPR General Data Protection Regulation

GEO Gene Expression Omnibus

GO Gene Ontology

GPU Graphics Processing Unit

GTEx Genotype-Tissue Expression

HDD Hard Disk Drive

HGNC HUGO Gene Nomenclature Committee

HPO Human Phenotype Ontology

IGV Integrated Genomics Viewer

IPIP Identical Partitions for Imbalance Problems

IQR Interquartile Range

KLD Kullback-Leibler Divergence

LBD Lewy body dementia

MGD Mouse Genome Database

ML Machine Learning

NGS Next-Generation Sequencing

NIH National Institutes of Health

OMIM Online Mendelian Inheritance in Man

OR Odds Ratio

OWL Web Ontology Language

PCA Principal Component Analysis

PD Parkison Disease

PET Positron Emission Tomograph

POR Phenotypic Overlap Ratio

RAM Random Access Memory

RDF Resource Description Framework

RNA Ribonucleic Acid

RT-PCR Reverse-Transcriptase Polymerase-Chain-Reaction

SAM Sequence Alignment Map

SAMtools Sequence Alignment/Map tools

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2

SMS Servicio Murciano de Salud

SQL Structured Query Language

SSD Solid State Drive

UMLS Unified Medical Language System

VaD Vascular Dementia

VCF Variant Call Format

CAPÍTULO 1

INTRODUCCIÓN

Este capítulo se orienta a la exposición de los conceptos fundamentales abordados en la tesis, ofreciendo una visión general de la bioinformática y sus aplicaciones. A lo largo del mismo, se recopilan los diversos modos en que la bioinformática nos ayuda para la manipulación y análisis de varios tipos de datos, que abarcan desde la información clínica hasta la genómica y transcriptómica. En ese sentido, este capítulo cumple un papel crucial al sentar las bases y trazar el hilo conductor para la comprensión global de la tesis. Por lo tanto, se presentarán resúmenes de diversos conceptos y se definirán los objetivos a seguir. La tesis se centrará en el desarrollo de métodos y herramientas bioinformáticas que manejan datos de diversa naturaleza, con el objetivo de mejorar el proceso de diagnóstico clínico de enfermedades y la gestión de los pacientes.

La **bioinformática** es un campo interdisciplinar que combina biología, informática y matemáticas para el análisis de datos biológicos y médicos. Además, aplica el tratamiento automático de información a la biología y la medicina [1]. Su impacto en las ciencias de la salud ha sido significativo, permitiendo una mayor comprensión de las enfermedades y una mejor capacidad para prevenirlas, diagnosticarlas y tratarlas.

Debido a los constantes progresos técnicos y tecnológicos en el campo de la medicina y biología, especialmente en la genética, cada vez disponemos de una cantidad mayor de información relevante sobre un individuo o una enfermedad en particular. La bioinformática ha permitido el análisis a gran escala de datos clínicos y genéticos. Por tanto, ha posibilitado una mayor comprensión de los genes y las variantes, también

conocidas como **mutaciones**, que causan **enfermedades genéticas**.

Además, la bioinformática también ha contribuido a la identificación de nuevas dianas terapéuticas y al desarrollo de tratamientos personalizados basados en la genética. Por otro lado, también ha tenido un impacto importante en el diagnóstico clínico, mejorando la precisión y la eficacia. Esto ayuda tanto a la prevención de las enfermedades como a la búsqueda de una cura. Por ejemplo, los sistemas de apoyo al diagnóstico o manejo de pacientes basados en la bioinformática facilitan a los médicos tomar decisiones clínicas con mayor información y precisión [2, 3, 4, 5]. Una de las ventajas del uso de herramientas bioinformáticas es que han permitido integrar y analizar datos clínicos y genómicos ayudando a médicos e investigadores en diferentes tareas.

Es importante puntualizar que existen diferentes perfiles bioinformáticos. Podemos decir que hay bioinformáticos que se centran en la aplicación de herramientas de análisis de datos en las ciencias biológicas y medicina. Estos han utilizado recursos que se han ido creando y validando en los últimos años. Por otra parte, hay otra vertiente que acerca a los bioinformáticos a los ingenieros de software; esto pasa cuando se centran más en el desarrollo de herramientas, con todos los matices que esto supone. En esta tesis hemos trabajado en los dos ámbitos, analizando, creando métodos y desarrollando herramientas.

Al tratarse de un campo en constante desarrollo hay una serie de desafíos y problemas que necesitan de la integración de diferentes especialidades para poder resolverse. Actualmente contamos con un ingente cantidad de datos para analizar. La bioinformática debe hacer uso de la **estadística**, la programación, la **inteligencia artificial** , *Artificial Intelligence* en inglés (**AI**), y sus aplicaciones particulares, como el Aprendizaje Automático, *Machine Learning* en inglés (**ML**), junto con el conocimiento especializado de biólogos y médicos para obtener provecho de esos datos. Encontrar **patrones** en los datos y asociar esa información a eventos clínicos o enfermedades es una de las tareas propia de la bioinformática. En estas tareas la AI junto con el ML son fundamentales para lograr el objetivo. Los sistemas de ML utilizan algoritmos y modelos estadísticos para analizar y extraer patrones de datos, que luego usan para realizar predicciones (más información en la sección 5.2.4.1 y 5.2.4.4). En resumen, transformar los datos en información útil es lo que se conoce como el proceso de análisis inteligente de datos.

En definitiva, la bioinformática ha tenido un impacto transformador en las ciencias de la salud, permitiendo una mayor comprensión de las enfermedades, mejores

diagnósticos y tratamientos personalizados. Su aplicación continúa expandiéndose y evolucionando, lo que supone un futuro esperanzador para el manejo y tratamiento de las enfermedades.

En las siguientes secciones vamos a introducir los principales campos de los que se nutre la bioinformática y en los que puede ayudar. Durante este capítulo nos centraremos en los más relevantes tratados en los capítulos de resultados de la tesis. Profundizaremos en la **diferente naturaleza y fuentes de los datos clínicos, genómicos y transcriptómicos**. Además, presentaremos algunos recursos, técnicas, tecnologías y herramientas que se utilizan, a la par que definiremos algunos de los retos y problemas a los que se enfrenta la medicina y la biología para mejorar el proceso de diagnóstico clínico.

1.1 Genética, producción de datos y ciencias ómicas

La **genética** es un campo de estudio de la biología dedicado a comprender los cambios en el ácido desoxirribonucleico, *Deoxyribonucleic Acid* en inglés (DNA), y la herencia de rasgos o cualidades entre padres e hijos. El **DNA** contiene la información genética y se ordena de forma determinada en lo que se conoce como genes. El DNA tiene numerosas características, por ejemplo, es una molécula que se puede replicar y cuyas propiedades fisico-químicas le aportan estabilidad. Un **gen** es una secuencia específica de DNA que codifica la información que se transcribe en otro ácido nucleico llamado **RNA**, *Ribonucleic Acid* en inglés, y que se traduce en una proteína, biomolécula que tiene diferentes funciones esenciales para la vida. Este flujo de como se transmite la información genética es conocido como el dogma central de la biología, en la figura 1.1 encontramos un esquema del mismo[1].

Por tanto, la composición genómica de un individuo, conocida como genotipo, es la información hereditaria completa de la secuencia de sus genes y determina una buena parte de los rasgos que podemos observar en un ser vivo. Estos rasgos observables se conocen como el fenotipo del individuo, el cual se determina por la interacción del genotipo con factores ambientales. En el ámbito médico, las enfermedades son definidas y caracterizadas por un conjunto de rasgos observables, es decir, un conjunto de fenotipos distintos de los que se consideran normales. Por consiguiente, saber el genotipo de un individuo nos permite conocer una información muy valiosa para su salud. Los cambios en la secuencia del DNA, llamados mutaciones o variantes, pueden causar enfermedad,

[1] `https://www.genome.gov/es/genetics-glossary/Central-Dogma/`

los que causan enfermedad se denominan mutaciones patogénicas. La mayoría de estas mutaciones son neutras, no tienen ningún efecto en la salud, pero, en algunos casos, dan lugar a lo que se conocen como **enfermedades monogénicas**, en los que una mutación patogénica en un gen es suficiente para causar la enfermedad y que se pueden transmitir de padres a hijos o surgir por fallos en el mecanismo genético de un individuo. Por otro lado, tenemos lo que se conoce como **enfermedades poligénicas** o complejas, son causadas por una combinación de factores genéticos, normalmente varias mutaciones en distintos genes que predisponen a sufrir la enfermedad, junto con factores ambientales, estas no se transmiten de padres a hijos de la misma forma que las enfermedades monogénicas.

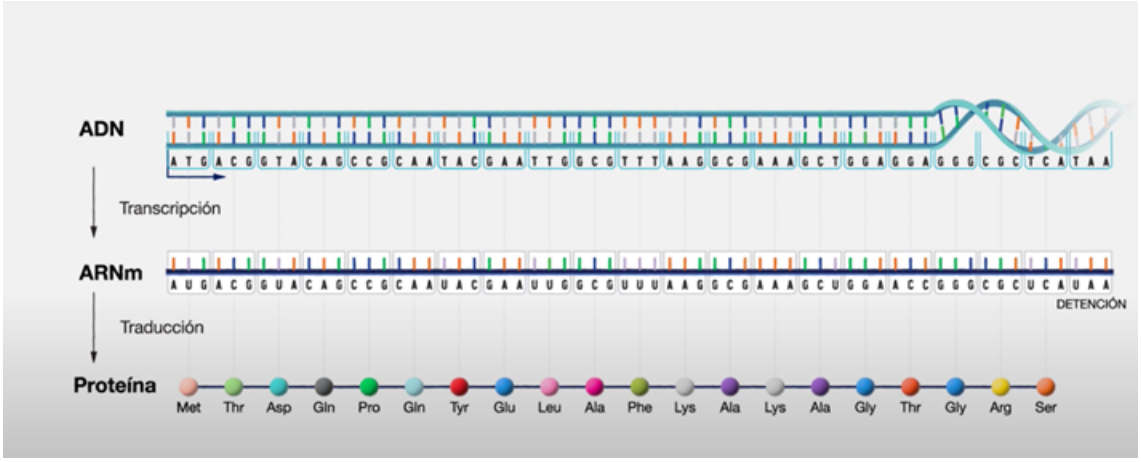


Figura 1.1: Esquema resumen de lo que se conoce como el dogma central de la biología.

El **diagnóstico genético** es un procedimiento realizado en la práctica clínica con la finalidad de determinar la presencia de anomalías genéticas que afectan al individuo y le hacen desarrollar o poder sufrir una enfermedad. Este tipo de pruebas puede realizarse antes del nacimiento (diagnóstico genético prenatal), durante la infancia o en la edad adulta, dependiendo de la enfermedad que se esté buscando en el individuo. Además, es una tarea que necesita de diferentes tecnologías y de la interacción de especialistas en diversas áreas. Aunque no hay una fecha determinada, se podría decir que el diagnóstico genético se lleva realizando desde principios de los **años 60**, lo cual permitió la asociación del **síndrome de Down** a defectos cromosómicos [6]. Posteriormente, a partir de los años 80 y 90 se avanzó mucho más en el campo con técnicas como la **secuenciación Sanger** y el estudio de multitud de enfermedades [7]. La **secuenciación del DNA** se utiliza para conocer el orden preciso de los nucleótidos (adenina [**A**], guanina [**G**], citosina [**C**] y timina [**T**]) en una molécula de DNA. El método Sanger de secuenciación del DNA se basa en el uso de dideoxinucleótidos, nucleótidos que carecen de un grupo 3'-hidroxilo

(-OH), que detienen la replicación del DNA cuando se incorporan a una cadena en crecimiento. Después de aislar y clonar el DNA de interés, se preparan cuatro muestras, cada una con un dideoxinucleótido diferente. Esto da lugar a fragmentos de DNA de diferentes longitudes que terminan en el punto de incorporación del dideoxinucleótido. Estos fragmentos se separan por electroforesis y de esta forma se puede determinar la secuencia del DNA original.

Desde la secuenciación del primer genoma humano, al inicio de los años 2000, el campo del diagnóstico genético ha experimentado un tremendo cambio [8]. Este desarrollo ha sido posibilitado por la implementación y avance de las tecnologías de **secuenciación masiva (NGS)**, *Next-Generation Sequencing* en inglés. Las NGS son un conjunto de técnicas que permiten secuenciar millones de fragmentos de DNA al mismo tiempo, superando a métodos anteriores como la secuenciación Sanger. Estas técnicas han conducido al abaratamiento de los costes de secuenciación del genoma y a la generalización del uso del diagnóstico genético en medicina [9]. El diagnóstico genético ha generado un aumento considerable de los datos a analizar. Por un lado, los datos en el ámbito de la investigación. Por otro, en el diagnóstico por parte de los facultativos que se encargan, en última estancia, de determinar la mutación causal en un paciente concreto. A continuación se exponen algunos datos de la cantidad de información que se puede extraer del genoma de una persona:

- Unos **3.2 billones americanos** de pares de bases, es decir, de fragmentos que contienen información genética y que se conocen como bases **(A, T, G y C)**.
- Unos **20000 genes** codificantes para proteínas que se estima producen unas **100000 proteínas** distintas [2].
- Si hablamos de un **genoma completo** en bioinformática para almacenarlo se utilizan los archivos BAM, *Binary Alignment Map* en inglés, o FASTQ que pueden pesar unos **180GB**.
- Teniendo en cuenta solamente las variantes diferentes de un individuo respecto a un genoma humano de referencia, normalmente entre 3 y 5 millones de variantes por individuo, obtenemos los archivos VCF, *Variant Call Format* en inglés, que son de aplicación clínica y que ocupan alrededor de 125MB. En la figura 1.2

[2] `https://nigms.nih.gov/education/Inside-Life-Science/Pages/genetics-by-the-numbers.aspx`

encontramos un esquema del proceso de secuenciación simplificado hasta llegar al fichero VCF[3].

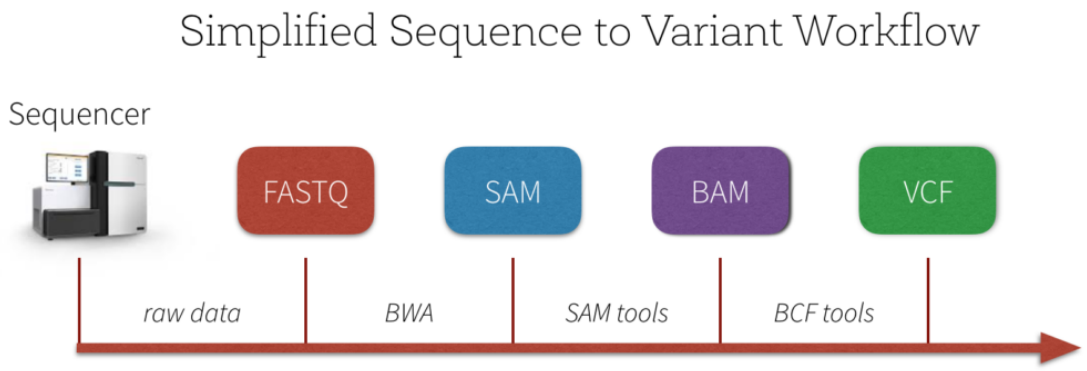


Figura 1.2: Simplificación del flujo de secuenciación hasta la obtención del archivo VCF.

Hasta ahora nos hemos centrado en la parte genómica de la genética y en la secuenciación, pero, podemos encontrar otras ramas dentro de lo que se conoce como las **ciencias ómicas** [10]. Se llaman ciencias ómicas a los campos de estudio de la biología que posibilitan estudiar diferentes tipos de moléculas, implicadas en el funcionamiento del organismo. Por su importancia se van a mencionar la siguientes ciencias ómicas:

- **Genómica**: La genómica estudia la función y estructura de los genes como hemos desarrollado en este capítulo.

- **Transcriptómica**: Es la rama que se refiere a todo lo que tiene que ver con la parte del proceso de **transcripción** del DNA al RNA que hemos comentado anteriormente [11]. Una de sus principales tareas es la **cuantificación** del RNA que nos da una idea de como se están expresando los genes utilizando técnicas como el **RNA-seq**, la secuenciación del RNA, y el uso de computación. Para comprender esta tesis es importante conocer el término transcrito. Un transcrito es una copia de la información genética que se encuentra en una región específica del DNA, que se transcribe a RNA para su posterior traducción en proteínas o para desempeñar otras funciones biológicas.

[3] `https://www.databricks.com/blog/2016/05/24/parallelizing-genome-variant-analysis.html`

- **Proteómica**: Esta rama se centra en conocer el conjunto de las proteínas de un organismo, así como, su función, estructura y cuantificación entre otros propósitos [12].

- Hay muchas otras ómicas que tienen menor relevancia en esta tesis como la **metabolómica** [13], **epigenómica** [14], **nutrigenómica** [15], **lipidómica** [16], etc.

En resumen, las ciencias ómicas tratan de caracterizar el funcionamiento y estado biológico de un individuo. Aportan una capa más detallada de información para caracterizar con datos concretos el estado de un individuo y, por tanto, generan datos de interés y relevancia relacionados con la presencia de enfermedades. La información que suministran puede funcionar como marcadores de enferme dad, asociarse a determinados fenotipos o ayudar al diagnóstico [17]. Los datos que proporcionan son utilizados para el tratamiento y manejo de enfermedades e incluso se pueden utilizar en conjunto para caracterizar más detalladamente un estado concreto [18].

En definitiva, cada vez tenemos más información y datos con los que trabajar para tratar de diagnosticar individuos como estamos comprobando a lo largo de la introducción. La bioinformática es crucial para analizar, manipular y asegurar la calidad de los datos provenientes de todos estos campos.

1.2 Mejora del diagnóstico clínico mediante la incorporación de datos clínicos, genética y genómica

Para entender el proceso de diagnóstico y tratamiento clínico es importante conocer los diversos factores que afectan al proceso en la actualidad. El diagnóstico clínico es un campo sujeto a numerosos cambios para mejorar los procesos tradicionales. En parte, estos cambios han sido originados gracias a los diferentes avances en biología, genética, medicina y tecnología. Tanto la investigación como la práctica clínica están generando ingentes cantidades de datos que deben ser analizados. Esto hace cada vez más necesario un enfoque multidisciplinario para mejorar estos procesos y aquí la bioinformática tiene un papel crucial.

En el transcurso del siglo XXI se han producido numerosos avances en el campo de la medicina y la ciencia, lo cual ha resultado en importantes logros en cuanto a la prevención, diagnóstico y tratamiento de diversas enfermedades. A modo de conocer algunos de los campos o avances más importantes creemos importante señalar y detenernos en ellos. A continuación se presentan algunos de los hechos más destacados en la prevención, diagnóstico, tratamiento de enfermedades y manejo de pacientes:

1. La **secuenciación del genoma humano** ha sido, quizás, uno de los logros más importantes a nivel científico. El cual permite un avance en en el conocimiento de la biología. Mejorando por tanto la medicina, posibilitando entender las causas genéticas de enfermedades y el desarrollo de tratamientos personalizados.

2. La llegada de la **terapia génica**, un tratamiento médico innovador que se está empezando a utilizar para tratar o prevenir enfermedades. Funciona reemplazando, inactivando o introduciendo en las células de un paciente mutaciones o partes del genoma para combatir o prevenir condiciones de salud. Este tipo de terapias han avanzado significativamente en el siglo XXI, permitiendo la modificación genética para tratar enfermedades, por ejemplo la administración estadounidense de alimentos y medicamentos, *U.S. Food and Drug Administration* en inglés (FDA), ha aprobado tratamientos para enfermedades raras o para el cancer[4].

3. La **medicina regenerativa** ha progresado mucho gracias al desarrollo de terapias basadas en células madre y la ingeniería de tejidos.

4. Los avances en la **inmunoterapia** han cambiado radicalmente el tratamiento de muchos tipos de cáncer, al utilizar el sistema inmunológico del paciente para atacar y destruir células cancerosas [19, 20, 21, 22, 23].

5. La atención a la salud mental ha avanzado significativamente ganando atención y recursos, con una mayor comprensión y tratamiento de enfermedades mentales como la depresión o la ansiedad.

6. Las **tecnologías de la información** y la comunicación se han utilizado para mejorar la toma de datos y comunicación entre los pacientes y los proveedores de atención médica, así como para proporcionar acceso a la atención médica en áreas de difícil acceso o a pacientes que lo requieren con la telemedicina.

[4] `https://www.fda.gov/consumers/consumer-updates/how-gene-therapy-can-cure-or-treat-diseases`

7. Se están utilizando **técnicas de AI y ML** en la medicina para la detección temprana de enfermedades y la **creación de modelos personalizados** de atención médica [24, 25, 26].

Como se ha mencionado anteriormente, en esta tesis vamos a desarrollar herramientas para la mejora del diagnóstico de enfermedades y para ayudar a la investigación de las mismas. Centrándonos en el diagnóstico de enfermedades es importante comprender que la utilización de tecnologías de la información, técnicas de AI y ML junto con el análisis de los nuevos datos disponibles, biológicos y clínicos, posibilita un gran avance en el campo, pero, también genera una serie de retos y necesidades. Ahora es crucial tener personal que pueda trabajar y entender esos datos y aquí es donde los profesionales formados en bioinformática pueden ayudar a mejorar los procesos clásicos de diagnóstico.

1.2.1 Proceso de diagnóstico clínico: problemas y posibles soluciones

En el párrafo anterior veíamos algunos de los campos y sectores que están posibilitando los avances en medicina. Ahora vamos a ver algunos de los problemas del proceso de diagnóstico y los elementos que creemos pueden ayudar a solucionarlos. En esta sección se va a resumir el proceso de diagnóstico clínico tradicional y las posibles mejoras que podemos obtener utilizando la bioinformática. Dentro del proceso de diagnóstico clínico tradicional encontramos una serie de pasos que los profesionales de la salud siguen para tratar de determinar la causa de los síntomas que presenta un paciente. Se puede resumir en los siguientes pasos:

1. **Consulta del historial médico**: El médico entrevista al paciente para obtener información sobre sus síntomas, su historial médico previo, su estilo de vida y otros factores relevantes. Tradicionalmente este historial médico ha sido registrado físicamente en papel pero ya hace un tiempo que gracias a las tecnologías de la información este se guarda en formato electrónico, *Electronic Health Record* en inglés (**EHR**). El EHR presenta una serie de ventajas como la accesibilidad, la facilidad de almacenamiento, la actualización y la facilidad de consulta para obtención de datos de cara a la investigación.

2. **Exploración física**: El médico examina al paciente para identificar signos físicos de enfermedad, como inflamación, dolor, presencia de infección, erupciones cutáneas

o tumores.

3. **Realización de pruebas de diagnóstico**: El médico puede necesitar realizar pruebas de diagnóstico para ayudarle a confirmar o descartar ciertas enfermedades. Este tipo de pruebas tradicionalmente han incluido análisis de sangre, imágenes médicas (como radiografías o tomografías) y pruebas de las funciones de órganos como el corazón o los pulmones, entre otras. En este paso encontramos que en ocasiones se solicitan pruebas de diagnóstico genético las cuales ayudan en la toma de decisiones. Cuando se solicitan pruebas para diagnóstico genético la bioinformática juega un papel clave en el análisis y generación de nuevas evidencias.

4. **Interpretación de las evidencias**: Ya con el conocimiento de los resultados de las pruebas, tradicionalmente el médico evalúa y compara con los síntomas y el historial del paciente para llegar al **diagnóstico**. Ahora, esa interpretación de las evidencias puede estar apoyada en los datos generados por la interpretación de un modelo de AI que aporte al médico una capa más de información y le ayude a tomar la decisión final. Debido a la cantidad de enfermedades, síntomas e información que hay que tener en cuenta se sabe que durante el proceso de diagnóstico médico ocurren **errores**. Estos errores son diversos y pueden ser humanos o errores técnicos producidos por la instrumentación, pruebas o máquinas utilizadas. Los errores de diagnóstico médico pueden presentarse entorno al **10-20%** de los casos[5] y ser producidos en diferentes pasos del proceso [27]. En ocasiones se producen porque varias enfermedades presentan fenotipos similares y es difícil tener un diagnóstico concreto. Se estima que la **tasa de errores** es **superior** cuando hablamos de **enfermedades raras**, se considera enfermedad rara a la que se presenta en muy pocos individuos 2 de cada 5.000, se conocen más de 6000 enfermedades de este tipo, la mayoría de ellas de origen genético, y si se toman en su conjunto se estima que afectan a entre el 3-7% de la población global [28].

5. **Tratamiento**: Una vez se llega al diagnóstico, el médico está en condiciones de proponer un tratamiento para el paciente. El plan de tratamiento puede incluir medicamentos, cambios en el estilo de vida y operación quirúrgica entre otros.

Por tanto, en base al proceso descrito y a los problemas mencionados se pueden proponer diferentes mejoras. Nosotros sostenemos que el proceso tradicional de diagnóstico clínico se puede mejorar incorporando información genética y tecnologías de la

[5] `https://www.improvediagnosis.org/what-is-diagnostic-error/`

información. La gran cantidad de datos que son tomados del paciente por el personal de salud durante las pruebas se guardan en distintos formatos que pueden ser explotables utilizando tecnologías de la información. La accesibilidad a gran cantidad de datos permite identificar patrones utilizando técnicas de AI y ML, desarrollar **herramientas que permitan mejorar el diagnóstico clínico** y dar apoyo a los profesionales de la salud en el proceso de toma de decisiones durante todas las etapas del diagnóstico. Esto puede aplicarse a diferentes campos de la medicina como la dermatología [29], enfermedad hepática [30], el diagnóstico del cáncer [31, 32], enfermedades coronarias [33] o a enfermedades neurodegenerativas entre otras [34].

Como hemos mencionado anteriormente, los avances en biología y genética han producido una nueva fuente de información. La utilización de datos biológicos para la práctica clínica ha sido una revolución. Si profundizamos un poco más y con el objetivo de introducir conceptos relevantes para la tesis es importante definir lo que conocemos como un biomarcador. Un **biomarcador** es una característica biológica que se puede medir objetivamente y que sirve como indicador de un proceso biológico normal o patológico, también, de una respuesta concreta a una exposición o tratamiento determinado en una enfermedad. Los biomarcadores son moléculas biológicas como proteínas, ácidos nucleicos, metabolitos, etc. En algunas enfermedades permiten la identificación temprana de las mismas, la evaluación de la progresión, la selección de tratamientos personalizados y la monitorización de la eficacia del tratamiento [35, 36].

Los biomarcadores se pueden clasificar según su función clínica. Los biomarcadores diagnósticos se utilizan para identificar una enfermedad, los biomarcadores pronósticos se utilizan para predecir la progresión de una enfermedad, y los biomarcadores predictivos se utilizan para predecir la respuesta a un tratamiento específico. Por tanto, los biomarcadores son un elemento clave para los procesos diagnóstico actuales. Además, son recursos que hacen que el proceso de diagnóstico sea incluso más importante ya que alguno de ellos se pueden utilizar antes de los síntomas y utilizarse de forma preventiva para el tratamiento.

Las **enfermedades de base genética** son otro de los campos mencionados anteriormente como problemáticos. Hemos mencionado que son **difíciles de diagnosticar** y que se producen fallos durante ese proceso. Cuando hablamos concretamente de diagnóstico genético de una enfermedad surge el concepto de **odisea diagnóstica** debido a la dificultad del proceso. Este concepto se define como el tiempo entre el momento en que se observa un síntoma o característica (fenotipo) que puede ser asociada a una enfermedad

genética o rara y el momento en que se realiza un diagnóstico final en el que se asocia el genotipo al fenotipo[6].

Hay diferentes factores que afectan durante el proceso de diagnóstico genético que dificultan el proceso: la falta de asociaciones gen-fenotipo, la cantidad de información que se tiene que estudiar, la falta de estandarización de los datos o el número de expertos implicados. Es por esto que en el proceso de diagnóstico genético también se está comenzando a utilizar AI para ayudar en los distintos problemas mencionados. Han surgido proyectos científicos y empresas, como Emedgene[7], Congenica[8] o Fabric genomics[9]. Por ejemplo, Emedgene ofrece un software que pretende ayudar al diagnóstico genético utilizando AI para aumentar por una parte la precisión en el diagnóstico y, por otra, reducir la carga de trabajo y el tiempo empleado por los facultativos. Emedgene hace uso de algoritmos de ML para la priorización de posibles variantes causales junto con grafos de conocimiento basados en datos de variantes, literatura, fenotipos y enfermedades, entre otros, para explicar y fundamentar las asociaciones sugeridas.

En definitiva, creemos que en el campo del **diagnóstico clínico** se puede **mejorar** utilizando la **bioinformática** con la finalidad de reducir la tasa de errores del diagnóstico y generar nuevos métodos diagnósticos. Además, se puede reducir el tiempo necesario para el diagnóstico, ayudar en el manejo de pacientes y ayudar en el futuro diagnóstico de enfermedades de las cuales no se conoce la causa previamente. Por esto, uno de los objetivos de esta tesis es la creación de herramientas bioinformáticas que ayuden a mejorar el diagnóstico de enfermedades y que manejen las evidencias disponibles. Para lograr esto objetivo es crucial conocer los diferentes tipos de datos que podemos utilizar en el proceso de diagnóstico clínico, vamos a verlos en la siguiente sección.

1.3 Tipos y fuentes de datos clínicos y biológicos

Anteriormente hemos visto algunos de los datos más relevantes que se pueden tomar para un paciente o caso en concreto. En esta sección vamos a tratar las fuentes más comunes de datos clínicos y biológicos. Se comentarán algunas de las bases de datos curadas donde se recoge información de interés para la comunidad científica. El término curada se refiere a que es un conjunto de datos en el que la información ha sido seleccionada,

[6] https://fdna.health/es/knowledge-base/what-is-a-diagnostic-odyssey/
[7] https://www.emedgene.com/
[8] https://www.congenica.com/
[9] https://fabricgenomics.com/

verificada, organizada y optimizada por expertos en el campo. Además, vamos a explicar con más detalle los posibles formatos que tienen esas bases de datos, junto con las diferentes posibilidades de obtener los datos.

En la sección de diagnóstico clínico definíamos que los médicos y el personal sanitario son los encargados de tomar los datos clínicos a los pacientes. La fuente más común en cuanto a datos clínicos es el **EHR**, aunque el EHR tiene mucha información que no acaba siendo objeto de análisis en investigación ya que con frecuencia se extrae, previamente, en otros formatos por parte de los informáticos encargados del manejo de los datos en los hospitales. Esto se traduce en que la mayoría de información que llega a los investigadores suele ser **tablas** en formato **CSV**, *comma-separated values* en inglés, o bases de datos relacionales de tipo SQL, *Structured Query Language* en inglés. Es necesario tener permisos concretos para contar con la información clínica de unos pacientes determinados, además esa información debe venir de forma anonimizada[10], es decir, que no se pueda identificar o relacionar el paciente en la vida real. Todo esto es debido al Reglamento General de Protección de Datos, *General Data Protection Regulation* en inglés (GDPR). Por lo tanto los equipos de investigación tienen que lidiar con estos procedimientos para recibir ese tipo de datos concretos de los pacientes. En esta tesis se han recibido datos de pacientes concretos para desarrollar la investigación expuesta en el capítulo 5 y en el 6 por lo que ha se han llevado a cabo el cumplimiento de estos procedimientos.

Cuando se investiga con datos clínicos se suelen obtener los datos de distintos hospitales y en ocasiones es difícil la estandarización de la toma de datos junto con la utilización de unos términos o vocabulario común para referirse al mismo síntoma o enfermedad [37, 38]. Para intentar mitigar este problema hay diversas iniciativas y consorcios que tratan de aportar soluciones a los diferentes aspectos que se deben tratar. La mayoría de estos recursos suelen ser ontologías.

Las **ontologías** definen los términos y las relaciones utilizadas para describir y representar los conceptos y objetos de un dominio, y establece una jerarquía de clases y subclases que reflejan la estructura de ese dominio. Las ontologías suelen ser comunes en el ámbito de la informática y se utilizan distintos formatos para representarlas, siendo los más frecuentes el formato OWL, *Web Ontology Language* en inglés, y el RDF, *Resource Description Framework* en inglés. Un ejemplo de una ontología que trata, entre otras cosas, de estandarizar los términos de fenotipos del paciente es la *Human Phenotype*

[10] `https://www.aepd.es/es`

Ontology (HPO). Otro ejemplo es el UMLS, *Unified Medical Language System* en inglés, un sistema que trata de integrar y unificar el vocabulario biomédico para mejorar la interoperabilidad [11]. Los datos biológicos se relacionan con los datos clínicos mediante bases de datos y suelen encontrarse representadas sus asociaciones en las mencionadas ontologías.

Por otro lado, los datos genómicos y transcriptómicos se obtienen después del proceso de secuenciación del genoma o del transcriptoma. Estos datos se guardan en distintos tipos ficheros y son tratados por bioinformáticos y especialistas de genética previo al manejo por parte del clínico. Son datos sensibles al igual que los datos clínicos y sujetos al GDPR. Para compartirlos se necesita permiso y se suelen depositar en repositorios como *Gene Expression Omnibus* (GEO) o guardar en servidores del centro de investigación.

En resumen, algunos de los problemas de los datos con los que trabajamos son los siguientes:

- Los datos son de distinta naturaleza y se presentan en distintos formatos. Tenemos datos clínicos, genómicos, transcriptómicos fenotipos, metadatos, etc.
- Son datos sensibles sujetos al GDPR.
- La interoperabilidad es difícil debido a la complejidad de la estandarización de los datos entre hospitales, regiones y países.
- La comunicación entre investigadores y clínicos es compleja. Se necesita personal formado en distintas áreas e incentivos para compartir los datos.

Por tanto, todos estos problemas hacen necesario mencionar, otra vez, que es crucial contar con la participación de diferentes campos para el uso efectivo y el tratamiento de los datos. Estos factores hay que tenerlos en cuenta de cara a la necesidad de crear recursos y herramientas para la comunidad científica. En resumen, son problemas a los que la bioinformática tiene que enfrentarse y tratar de solucionar.

[11] `https://www.nlm.nih.gov/research/umls/index.html`

1.3.1 Recursos disponibles de interés para la tesis

En esta sección se presentan con más detalle algunos de los recursos disponibles que son de interés en esta tesis y que nos aportan asociaciones entre datos biológicos, mayormente genómicos o transcriptómicos, y clínicos:

- *Online Mendelian Inheritance in Man* **(OMIM)**: Es un compendio exhaustivo y fidedigno de genes y fenotipos genéticos humanos que está disponible gratuitamente y se actualiza. El proyecto OMIM recoge un conjunto de reglas sobre la transmisión por herencia de los organismos padres a sus hijos, es una base de datos que cataloga todas las enfermedades humanas conocidas con un componente genético [39]. Además, trata de asociar los genes causantes de esas enfermedades. En su web podemos encontrar archivos disponibles y una API con información de asociaciones entre genes, enfermedades, tipos de herencia, fenotipos, etc. En la figura 1.3 podemos encontrar una vista de la sección resumen para un tipo de epilepsia concreto, aquí vemos el gen asociado, el tipo de herencia de la enfermedad y algunos de los fenotipos asociados. Este recurso es importante para el desarrollo de la herramienta presentada en el capítulo 4. Además, este recurso es una fuente de consulta y extracción de datos en el proyecto desarrollado para la empresa NIMGenetics con la finalidad de generar un priorizador de variantes.

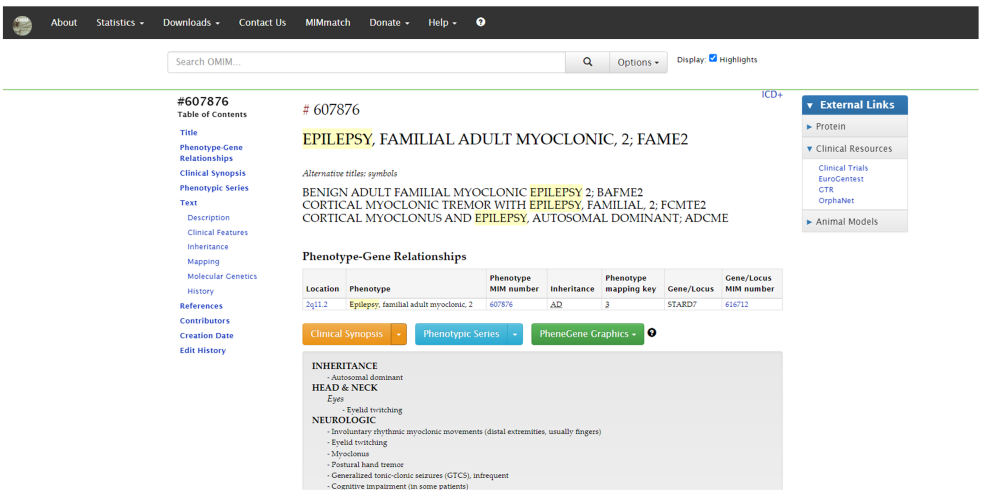


Figura 1.3: Captura de la web de OMIM para un tipo de epilepsia concreto

- **Orphanet**: Es una base de datos de origen europeo sobre enfermedades raras que fue fundada en 1997 [40]. En su web podemos encontrar recursos y términos concretos dirigidos a mejorar el entendimiento de las enfermedades raras. Este

recurso es importante para el desarrollo de la herramienta presentada en el capítulo 4.

- *DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources* **(DECIPHER)**: Es una base de datos y recurso web que recoge anomalías en la secuencia genómica de más de 45.000 pacientes que presentan diferentes enfermedades con sus respectivos fenotipos [41]. DECIPHER fue fundada en 2004 en el *Sanger Institute* de Reino Unido y continua tomando datos de pacientes y expandiendo el conocimiento. Este recurso es importante para el desarrollo de la herramienta presentada en el capítulo 4.

- **HPO**: Es una ontología en la que se estandariza el vocabulario específico para hablar de los fenotipos que se pueden encontrar en las enfermedades [42]. Utiliza literatura médica y otras fuentes de datos como Orphanet, DECIPHER y OMIM con la finalidad de ayudar al diagnóstico y a la investigación. Tiene una terminología concreta con unos códigos que se pueden utilizar para nombrar fenotipos concretos, por ejemplo, HP:0001250 hace referencia al término fenotípico de convulsiones. En su web se pueden encontrar numerosos ficheros de asociación entre fenotipos y genes. Este recurso es importante para el desarrollo de la herramienta presentada en el capítulo 4. Además, este recurso es una fuente de consulta y extracción de datos en el proyecto desarrollado para la empresa NIMGenetics con la finalidad de generar un priorizador de variantes.

- *Genotype-Tissue Expression* **(GTEx)**: Es un proyecto financiado por el Instituto Nacional de Salud de Estados Unidos, *National Institutes of Health* en inglés **(NIH)**, para construir una base de datos con información de la variación genómica y de la expresión de los genes, es decir, la información transcriptómica, en diferentes tejidos como el cerebro, el corazón, etc [43]. Este recurso es importante para poder desarrollar la investigación que se presenta en el capítulo 6.

- **GEO**: Es un repositorio público internacional donde se archivan y se distribuyen datos genómicos y transcriptómicos generados por la comunidad científica [44]. En el recurso web se pueden encontrar numerosos datos de distintos análisis realizados por la comunidad científica. Este recurso es importante para poder desarrollar la investigación que se presenta en el capítulo 6.

1.3.2 Proceso de secuenciación del genoma y generación de datos

Como hemos mencionado anteriormente, la secuenciación del DNA se utiliza para conocer el orden preciso de los nucleótidos (adenina [**A**], guanina [**G**], citosina [**C**] y timina [**T**]). Este es un proceso que se realiza con diferentes fines, ya sea investigación o en la práctica clínica. Mayoritariamente, cuando se hace secuenciación del DNA en la práctica clínica es para buscar variantes que confieren riesgo o certeza de desarrollar alguna enfermedad. Estos genes y variantes pueden estar recogidos en las bases de datos anteriormente mencionadas como OMIM o DECIPHER. Aunque, en muchos de los casos no están recogidas y se tienen que determinar su importancia. Normalmente son alteraciones de la secuencia de referencia con la que se compara. El genoma de referencia es una representación estandarizada de la secuencia de DNA de una especie. Normalmente se construye con varios individuos de la misma especie.

Desde el punto de vista de los datos, el proceso de secuenciación del genoma de un individuo genera una gran cantidad de datos. En esta sección vamos a tratar con más detalle el proceso y la generación de datos genómicos y transcriptómicos. Aunque el proceso de secuenciación es complejo y largo se puede resumir en los siguientes pasos que han sido esquematizados en la figura 1.4:

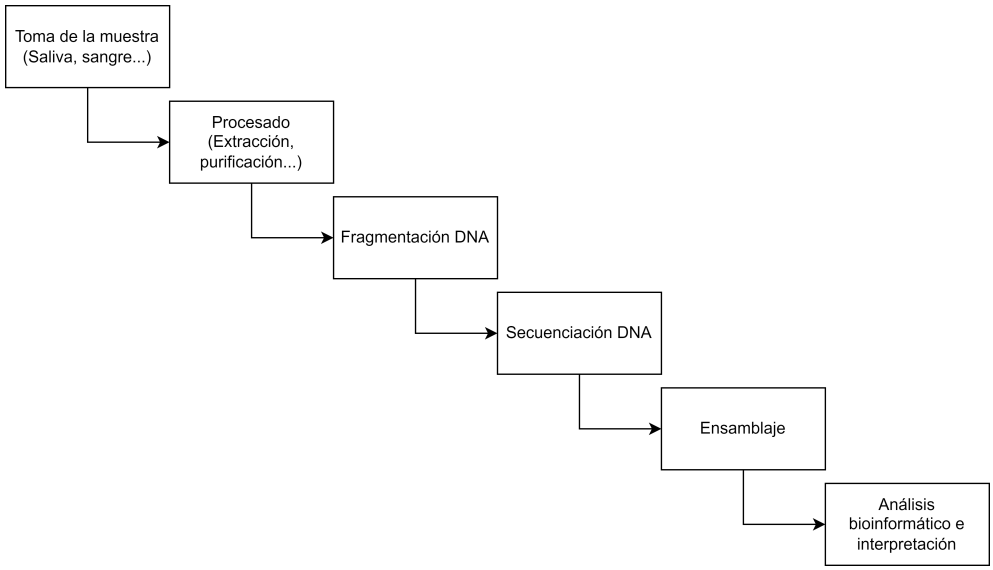


Figura 1.4: Esquema resumen del proceso de secuenciación del genoma

1. **Toma de la muestra**: El personal sanitario extrae una muestra de material biológico del individuo, normalmente suele ser sangre o saliva. La muestra se

guarda y se envía al centro de secuenciación en unas condiciones determinadas para asegurar su conservación.

2. **Procesado de la muestra**: Ya en el laboratorio los técnicos especializados extraen el DNA de la muestra, purificando y concentrándolo. Existen diversos protocolos para los pasos relativos a este punto dependiendo del objetivo y las necesidades finales.

3. **Fragmentación del DNA**: Se rompe el DNA en fragmentos relativamente pequeños de bases y se añaden adaptadores a los extremos de los fragmentos. La fragmentación se utiliza para asegurar que se pueda leer la secuencia con la suficiente calidad. Este proceso también se conoce como la preparación de librerías. Todo el proceso es más extenso y específico dependiendo del tipo de muestra y finalidad del análisis.

4. **Secuenciación del DNA**: Actualmente existen diferentes técnicas para secuenciar el DNA y el RNA. Tenemos técnicas de **NGS** como las desarrolladas por empresas como **Illumina** y **PacBio**. Cada una utiliza tecnologías diferentes para realizar la secuenciación. Illumina para secuenciar suele utilizar reactivos fluorescentes que reaccionan con cada una de las bases del DNA y emiten una longitud de onda determinada que se registra. En este proceso existen variaciones en la metodología basadas en distintas tecnologías que dan lugar a la variedad de secuenciadores que existen en el mercado. Estos secuenciadores tienen unas características diferentes y posibilitan utilizar la secuenciación para distintos propósitos [45].

5. **Ensamblaje**: En este paso se monta la secuencia de DNA para reconstruir los fragmentos y recuperar la secuencia completa. Existen diferentes algoritmos bioinformáticos para el alineamiento y ensamblaje de los fragmentos del DNA. Estos algoritmos han ido mejorando reduciendo el tiempo y los errores que se cometían en este proceso. Una vez termina este proceso tenemos diferentes tipos de archivos de datos para su análisis.

6. **Análisis bioinformático e interpretación**: La secuencia obtenida es objeto de distintos análisis dependiendo de los objetivos de la investigación. Por ejemplo, se puede comparar el **DNA** obtenido con el DNA que se tiene como referencia en la comunidad para **identificar** las **variantes** del sujeto. En el caso de trabajar con **RNA** podemos **cuantificar** las diferencias de expresión de los genes respecto a determinados fenotipos, por ejemplo, sano o enfermo. La interpretación de los

resultados es realizada por distintos especialistas y posibilita identificar la causa genética de la enfermedad, la reacción a ciertos fármacos, la predisposición o riesgo a sufrir la enfermedad y comprender la función de ciertos genes entre otras muchas cosas. Para conseguir cada uno de esos propósitos se utilizan ficheros de datos determinados y técnicas de análisis diferente o software específico.

En importante mencionar que en los diferentes pasos del proceso se comprueba la calidad de las muestras y la correcta realización de los mismos. En definitiva, se ha resumido el proceso para dar una visión global pero cada paso puede requerir ajustes específicos dependiendo de los objetivos con los que se analice el genoma. En nuestro caso, por un lado, utilizamos la información de la secuenciación del DNA para comparar y determinar las variantes que se asocian en un individuo a un fenotipo concreto. Esto lo hacemos para el proyecto del priorizador de variantes con la empresa NIMGenetics, en el que con esta información más la que se tiene en base de datos ayudamos con técnicas de ML a la priorización de las variantes candidatas en un caso determinado. Por otro lado, la cuantificación de RNA se utiliza para saber como se están expresando o comportando los genes en un tejido, condición y momento determinado. Este procedimiento es importante para el proyecto desarrollado en el capítulo 6.

1.4 Entornos de computación y software de análisis de datos en bioinformática

Hasta ahora hemos ido introduciendo conceptos y relacionando la bioinformática con diversos objetivos. En esta sección vamos a profundizar en las necesidades de computación generales y específicas para los grupos que trabajan en bioinformática. Además, vamos a resumir algunas de las herramientas, programas y técnicas de relevancia para manipular datos clínicos y genómicos. Todo esto es importante para lograr los diferentes objetivos propuestos en la tesis. ¿Qué es importante cuando se analiza el DNA o el RNA? ¿Qué debería tener instalado el servidor en el que trabajamos? ¿En qué nos van a ayudar estas herramientas?. Lidiar con estas cuestiones es importante en el trabajo de bioinformático y en el contexto de la tesis. Conocer las necesidades y el software específico para llevar a cabo la investigación.

Normalmente los grupos o los departamentos que se dedican a trabajar con datos genéticos tienen un servidor dedicado para ello. Suelen ser servidores a los que los

bioinformáticos acceden para procesar los datos. Normalmente estos servidores tienen instalado como sistema operativo alguna de las distribuciones de **Linux**, debido a que muchas de las herramientas con las que se trabaja son de software libre. Este es nuestro caso. Al generarse una gran cantidad de datos durante el proceso de secuenciación del genoma se necesita que el servidor donde se procesa y guarda la información tenga una gran capacidad de almacenaje. Esto se traduce en varios discos de estado sólido de gran capacidad donde normalmente se procesa, *Solid State Drive* en inglés **(SSD)**, discos duros mecánicos tradicionales para guardar la información, *Hard Disk Drive* en inglés **(HDD)**, y, en ocasiones, cinta magnética para conservar datos más antiguos debido a que es más económico. Hay que asegurarse de contar con la suficiente cantidad de almacenaje en el servidor, por ejemplo, en la investigación realizada en el capítulo 6 se han necesitado unos 4 TB de almacenamiento para realizar los análisis de RNA. Por el tipo de uso que se le da a estos servidores suelen contar con una gran cantidad de memoria **(RAM)**, *Random Access Memory* en inglés, y con procesadores, *Central Processing Unit* en inglés **(CPU)**, que consten de suficientes núcleos e hilos para ejecutar las tareas en paralelo. Es importante mencionar que si el grupo de investigación trabaja con AI normalmente se cuenta con tarjetas gráficas, *Graphics Processing Unit* en inglés **(GPU)**, ya que permiten paralelizar mucho más los problemas y obtener resultados más rápido.

Por otro lado, es importante conocer que en estos servidores se tiene instalado software específico para el procesado de datos genéticos. Hay multitud de programas, pero se suelen utilizar los aconsejados para seguir las buenas practicas definidas por el **Broad Institute**. El Broad Institute es una institución de investigación biomédica sin fines de lucro, situada en Cambridge, Massachusetts, Estados Unidos. El instituto tiene el objetivo de mejorar la comprensión de la genética humana, la biología y la medicina utilizando herramientas genómicas. Este instituto, el cual es una referencia en el campo, recomienda el software descrito en *Genome Analysis Toolkit* **(GATK)**[12]. GATK es un conjunto de herramientas de software de bioinformática para el análisis de datos genómicos. El software está diseñado para ser altamente flexible y escalable, lo que permite a los usuarios adaptarlo a una variedad de aplicaciones y tamaños de datos. También incluye características que manejan muchos de los problemas comunes en el análisis de datos genómicos, como errores de lectura de la secuenciación y sesgo en la representación de ciertas regiones del genoma. A continuación vamos a profundizar más en el software que se ha utilizado a lo largo de la tesis:

[12] `https://gatk.broadinstitute.org/`

1. *Burrows-Wheeler Aligner* **(BWA)**: Es un software que nos permite mapear o alinear la secuencia de DNA que tenemos con un genoma de referencia. Consta de tres algoritmos diseñados para alinear las lecturas al genoma de referencia [46].

2. *Sequence Alignment/Map tools* **(SAMtools)**: Es un software que nos aporta un conjunto de herramientas para interactuar con los distintos formatos de archivos obtenidos en el proceso de secuenciación como el SAM, Sequence Alignment Map en inglés, o el BAM [47]. Estas herramientas se han utilizado en el capítulo 6.

3. ***Picard***: También es un software para interactuar con los diferentes formatos de archivos obtenidos como el SAM, BAM o el VCF [13]. Estas herramientas se han utilizado en el capítulo 6.

4. **GATK**: Ofrece un conjunto de herramientas para manipular los ficheros y para el control de calidad del proceso. Estas herramientas se han utilizado en el capítulo 6.

5. *Integrated Genomics Viewer* **(IGV)**: Es una interfaz gráfica que permite visualizar los ficheros para una inspección visual del alineamiento o de la posición genómica concreta [48]. Estas herramientas se han utilizado en el capítulo 6.

6. ***R y RStudio***: Es un lenguaje y entorno de programación enfocado al tratamiento de datos estadísticos [49]. Al formar parte de un proyecto abierto y colaborativo hay una serie de usuarios y grupos que publican paquetes para añadir funcionalidades. Algunos de los recomendados por GATK en concreto es ggplot2 para la creación de gráficos y gsalib para cargar archivos con datos de secuenciado de GATK y manipularlos.

Junto con estas recomendaciones también merece la pena mencionar a PLINK. Este software nos permite utilizar una serie de comandos por linea para filtrar, convertir y analizar datos genéticos [50].

Si nos centramos en lenguajes de programación ampliamente utilizados en la bioinformática tenemos R y Python [14]. Estos lenguajes de programación nos permiten operar de forma precisa sobre los datos, crear software para tareas concretas y además tienen funcionalidades ya creadas y testadas por otros desarrolladores de las que podemos hacer uso. De R ya hemos comentado un poco en el punto anterior. Normalmente se

[13] `http://broadinstitute.github.io/picard/`
[14] `https://www.python.org/`

utiliza con RStudio a modo de entorno de desarrollo integrado. Algunos de los paquetes y utilidades más utilizados para bioinformática en R son:

- **ggplot2**: Visualización de datos y creación de gráficos complejos [51].
- **caret**: Es un paquete que incluye funciones para utilizar métodos de clasificación y regresión enfocado en crear modelos de predicción y evaluarlo [52]. Se utiliza también junto con keras para tareas de AI.
- **dplyr**: Se utiliza para facilitar la manipulación de datos en R [53].
- **Bioconductor**: Es un proyecto donde se suben gran cantidad de paquetes relacionados con el análisis de datos biológicos. Por ejemplo, paquetes como **biomaRt** [54], repositorio de datos biológicos para facilitar la integración, o **limma** [55], paquete para el análisis de RNAseq, están en bioconductor.

Es importante para el desarrollo de la tesis conocer las posibilidades que ofrece R. Nosotros como bioinformáticos podemos crear paquetes que funcionen dentro de R. En ese sentido uno de los objetivos de la tesis es crear herramientas que puedan ser utilizadas por otros, esto se desarrolla en profundidad en el capítulo 4. R nos permite integrar software específico y tener muchas funcionalidades interesantes. Además, más allá de desarrollar librerías para usuarios de R, el uso de **Shiny** en R nos permite generar webs que posibiliten a usuarios sin tantos conocimientos de bioinformática utilizar las **herramientas**. Esto se ha realizado para las investigaciones desarrolladas en el capítulo 4 y en el capítulo 5.

Por otro lado, **Python** es un lenguaje de programación ampliamente usado en diversos campos que también cuenta con utilidades específicas para el tratamiento de datos biológicos. Un ejemplo es biopython que tiene una serie de funciones para realizar diferentes operaciones sobre datos biológicos. Python suele utilizarse para la creación y producción de modelos y software que tienen que ver con la parte de AI. Hay diferentes bibliotecas como **Keras** [56], **NumPy** [57] o **TensorFlow** [58] que se pueden utilizar en Python para este propósito. En esta tesis se ha utilizado python en alguna funcionalidad concreta y en el desarrollo de proyectos en colaboración con la empresa NIMGenetics.

En definitiva, contamos con una serie de lenguajes de programación y software que permite que se puedan manipular datos biológicos. Además, estos lenguajes de

programación nos permiten crear nuevos paquetes y herramientas que añadan nuevas funcionalidades para la comunidad.

1.5 Motivación y Objetivos

En esta sección se profundiza en la motivación y objetivos que han sido mencionados a lo largo de la introducción.

El **propósito general** de esta tesis es ayudar **en las decisiones clínicas y en el diagnóstico** de enfermedades. Usaremos la **bioinformática** para analizar los datos y crear herramientas útiles para la comunidad. Es decir, intentaremos tanto **generar herramientas** de consulta y aplicación directa como otras más centradas en la investigación. Además, nos centraremos en las enfermedades con base genética, en las cuales a parte de los **datos clínicos** contamos con **datos genómicos**. Es necesario conocer que el programa de doctorado al que está asociado esta tesis es al de Envejecimiento y Fragilidad y que la rama del doctorado es **"Visión, Big Data y análisis inteligente de datos aplicados al envejecimiento"**. Por tanto, vamos a aplicar y desarrollar este tipo de técnicas de análisis de datos centrándonos en **enfermedades neurodegenerativas** propias del envejecimiento como el Alzheimer o el Parkinson.

Una de las motivaciones de esta tesis es **integrar datos** de distinta naturaleza. Esto es esencial para avanzar en el diagnóstico y tratamiento como hemos definido a lo largo de la introducción. Para tratar grandes cantidades de datos utilizaremos las técnicas de análisis inteligente. La finalidad es resolver las necesidades y problemas detectados, para ello se hace necesario un enfoque multidisciplinar. Por tanto, vamos a utilizar estadística y ML para extraer la información y los patrones de los datos. Además, una vez desarrollados los modelos o extraída la información útil vamos a generar y facilitar el uso de las herramientas desarrolladas.

Debido a que esta tesis cuenta con la cofinanciación de la empresa **NIMGenetics**, la cual principalmente se dedica al **diagnóstico genético**, una parte de la tesis se enfoca en herramientas que puedan ayudar al desarrollo de sus tareas.

Por último, es importante mencionar que, por las peculiaridades del periodo de realización de la tesis, hemos trabajado con enfermedades infecciosas tratando de ayudar en la clasificación de pacientes de riesgo.

A continuación se detallan los objetivos de forma más concreta y se puntualiza en

que capítulos se han tratado:

1. Revisar y sintetizar la literatura existente sobre los temas tratados. Esta tarea se ha llevado a cabo para cada una de las investigaciones realizadas.

2. Realizar un análisis exhaustivo de los actuales métodos bioinformáticos y herramientas utilizados en el diagnóstico clínico y en la investigación, en particular en el diagnóstico de enfermedades genéticas (capítulo 4), infecciosas (capítulo 5) y neurodegenerativas (capítulo 6). Centrando el análisis en los problemas detectados previamente.

3. Identificar las limitaciones y las necesidades no satisfechas o de particular interés en la comunidad. En particular, en el diagnóstico genético, de enfermedades infecciosas y de enfermedades neurodegenerativas que podrían abordarse mediante el desarrollo de nuevas herramientas bioinformáticas. Esta tarea se ha realizado previamente para cada uno de los proyectos expuestos.

4. Desarrollar nuevas herramientas bioinformáticas que permitan una mejor interpretación y utilización de los datos genómicos y biológicos en el diagnóstico clínico. Esto se tratará en el capítulo 4 y en el capítulo 6.

5. Aplicar técnicas de análisis inteligente de datos, estadística y ML para extraer información, ayudar al diagnóstico y a la clasificación de pacientes con enfermedades neurodegenerativas (capítulo 6). Por las fechas de desarrollo de la tesis, también a pacientes con enfermedades infecciosas (capítulo 5).

6. Colaborar con NIMGenetics para desarrollar y aplicar herramientas bioinformáticas que mejoren la eficacia y la eficiencia de sus procesos de diagnóstico genético utilizando técnicas de análisis inteligente de datos propias de bioinformática y estadística (capítulo 4). Además, se ha generado un proyecto de colaboración que surge de investigaciones desarrolladas a lo largo de la generación de la tesis (capítulo 8).

7. Validar la eficacia de las herramientas desarrolladas mediante la colaboración con la empresa de diagnóstico genético NIMGenetics o con datos clínicos reales. En el caso de validación con NIMGenetics se tratará en el capítulo 4 y para la herramienta que se está desarrollando en colaboración con NIMGenetics aplicando conocimientos obtenidos a lo largo de la tesis capítulo 8 en la sección 8.1.4. En el caso de la validación con datos clínicos reales se desarrollará en el capítulo 5 y con datos clínicos, genómicos y transcriptómicos en el capítulo 6.

8. Facilitar el uso de estas herramientas a toda la comunidad. Posibilitar el uso de estas herramientas a los médicos, biólogos y genetistas sin conocimientos bioinformáticos. Eso se ha abordado mediante la creación de webs para explotar la información de las herramientas desarrolladas en el capítulo 4 y en el capítulo 5.

9. Difundir los resultados y las herramientas desarrolladas a través de publicaciones científicas y conferencias para fomentar su adopción y uso por parte de la comunidad médica y científica. Esto se ha realizado para todas las investigaciones desarrolladas (capítulo 8).

1.6 Organización de la Tesis

Esta tesis se estructura en ocho capítulos bien diferenciados.

El primer capítulo, capítulo 1, constituye la introducción, en la que se brinda una motivación, se esbozan los conceptos necesarios para la comprensión de los temas abordados y se establecen los objetivos de la investigación.

A continuación, el capítulo 2 proporciona un conciso resumen en español de la tesis. De igual manera, el capítulo 3 presenta un resumen en inglés.

La sección de resultados se inicia con el capítulo 4. Aquí se presenta una herramienta para el análisis comparativo de fenotipos entre conjuntos de genes. Le sigue el capítulo 5, en el cual se desarrolla un modelo predictivo para clasificar el riesgo de hospitalización o fallecimiento en pacientes con coronavirus. Finalmente, en el capítulo 6, se utilizan datos genómicos y biológicos para construir modelos capaces de detectar Alzheimer en etapas tempranas, incluso en pacientes presintomáticos.

El capítulo 7 se dedica a discutir los hallazgos principales de las investigaciones, resumir las conclusiones de cada una, relacionarlas con los objetivos propuestos al principio y sugerir posibles direcciones para trabajos futuros.

En el capítulo 8 se mencionan las contribuciones principales de la tesis, la producción científica generada y los proyectos de colaboración que se han desprendido de la investigación realizada en este periodo.

Finalmente, el último apartado está dedicado a la bibliografía y reúne todas las referencias citadas a lo largo de la tesis.

CHAPTER 2

RESUMEN

En esta tesis hemos aplicado y desarrollado técnicas bioinformáticas para tratar de ayudar a los clínicos, a los biólogos y a otros bioinformáticos. Hemos integrado y usado datos médicos y genómicos con la finalidad de estudiar las relaciones gen-fenotipo (capítulo 4), ayudar en el manejo de pacientes con COVID-19 (capítulo 5) y encontrar biomarcadores para el diagnóstico de enfermedades neurodegenerativas como el Alzheimer (capítulo 6). En definitiva, el objetivo es ayudar en las decisiones clínicas y mejorar el diagnóstico de enfermedades.

Por un lado, se han analizado datos utilizando estadística y técnicas de Machine Learning (ML). Por otro, se han desarrollado métodos y herramientas bioinformáticas para intentar solucionar problemas y facilitar los análisis. Por tanto, tenemos aportes en las dos grandes ramas de la bioinformática, el análisis de datos y el desarrollo de herramientas.

Una de las grandes cuestiones de la biología y la medicina genética es la asociación entre los genes y el fenotipo. Una enfermedad puede asociarse a un determinado fenotipo y a unos genes concretos. Además, los investigadores y genetistas pueden definir enfermedades utilizando una lista de genes, esto se conoce como paneles de genes. Dadas dos enfermedades parecidas el diagnóstico diferencial y la correcta asociación gen-fenotipo puede ser compleja. Debido a la importancia de estas relaciones se han realizado esfuerzos para conocer las asociaciones gen-fenotipo. La información de estas asociaciones gen-fenotipo se ha ido recopilando generando multitud de bases de datos.

En el capítulo 4 desarrollamos una herramienta bioinformática (PhenoExam) para centralizar el uso de diferentes bases de datos posibilitando realizar pruebas estadísticas y estudiar las relaciones gen-fenotipo. Primero, detectamos una necesidad de comparar dos conjuntos de genes en base a sus fenotipos. Después, elaboramos métodos para realizar estas comparaciones determinando su similitud y diferencias. PhenoExam es capaz de concluir si esas similitudes entre fenotipos son estadísticamente relevantes. Además, es capaz de determinar en los paneles de genes de enfermedades muy similares sus diferencias. Hemos validado PhenoExam con enfermedades similares definidas por sus paneles de genes y utilizando genes derivados de investigaciones para intentar descubrir asociaciones gen-fenotipo. Esta herramienta se encuentra disponible en R y en Web.

Por otro lado, esta tesis ha tenido lugar durante un periodo de pandemia causado por el COVID-19. Gracias al proyecto de la Fundación Séneca y a los datos del Servicio Murciano de Salud hemos podido trabajar con datos médicos de unos 86.000 pacientes de COVID-19. Utilizando estos datos hemos realizando un estudio retrospectivo (capítulo 5). Hemos extraído información relevante utilizando estadística y ML para estudiar la relación entre sexo, edad y comorbilidades con los diversos tipos de pacientes. Además, hemos desarrollado un método para lidiar con el desbalanceo y lo hemos aplicado en la construcción de modelos predictivos. Estos modelos han determinado con una buena exactitud el estado final del paciente (fallece o sobrevive) o la necesidad de hospitalización (externo o ingreso) con los datos que conocíamos en el momento del diagnóstico (edad, sexo y comorbilidades).

Por último, utilizando datos clínicos, información genómica y transcriptómica hemos desarrollado un biomarcador de Alzheimer en fases tempranas de la enfermedad (capítulo 6). El Alzheimer es una enfermedad compleja y en la que es difícil obtener un diagnóstico temprano. Además, por sus síntomas, difícil de diferenciar de otras similares e incluso de confirmar hasta el fallecimiento del paciente. Utilizando la información transcriptómica procedente del RNA libre del plasma sanguíneo y técnicas de ML hemos desarrollado un biomarcador que detecta Alzheimer en fases tempranas con gran precisión. Hemos estudiado su comportamiento en diferentes estadios de la enfermedad obteniendo resultados prometedores. Finalmente, determinamos su especificidad comparando con el resultado obtenido en otras enfermedades neurodegenerativas.

CHAPTER 3

SUMMARY

In this thesis, we have applied and developed bioinformatics techniques in an attempt to help clinicians, biologists, and other bioinformaticians. We have integrated and utilized medical and genomic data with the purpose of studying gene-phenotype relationships (chapter 4), assisting in the management of COVID-19 patients (chapter 5), and finding biomarkers for the diagnosis of neurodegenerative diseases such as Alzheimer's (chapter 6). Ultimately, the goal is to assist in clinical decision-making and improve the diagnosis of diseases

On one hand, data has been analyzed using statistical and Machine Learning (ML) techniques. On the other hand, bioinformatics methods and tools have been developed to solve problems and facilitate analyses. Therefore, we have contributed to the two major branches of bioinformatics, data analysis and tool development.

One of the relevant questions in biology and genetic medicine is the association between genes and phenotype. A disease can be associated with a certain phenotype and specific genes. Additionally, researchers and geneticists can define diseases using a list of genes, known as gene panels. Given two similar diseases, differential diagnosis and the gene-phenotype association can be complex. Due to the importance of these relationships, efforts have been made to understand gene-phenotype associations. The information from gene-phenotype associations has been collected, generating numerous databases. In chapter 4, we develop a bioinformatics tool (PhenoExam) to integrate different databases, enabling the execution of statistical tests and the study of gene-phenotype relationships.

First, we identified a need to compare two sets of genes based on their phenotypes. Then, we developed methods to make these comparisons by determining their similarities and differences. Using randomization PhenoExam is capable of concluding whether these phenotype similarities are statistically significant. Additionally, it can identify differences in the gene panels of very similar diseases. We have validated PhenoExam with similar diseases defined by their gene panels and used genes derived from research to attempt to uncover gene-phenotype associations. This tool is available in R and on the web.

On the other hand, this thesis took place during a pandemic period caused by COVID-19. Thanks to the project of the Seneca Foundation and data from the Murcian Health Service, we were able to work with medical data from about 86,000 COVID-19 patients. Using these data, we conducted a retrospective study (chapter 5). We extracted relevant information using statistics and ML to study the relationship between sex, age, and comorbidities with various types of patients. Moreover, we developed a method to deal with imbalance, and we applied it in the construction of predictive models. These models have determined with good accuracy the final state of the patient (death or survival) or the need for hospitalization (outpatient or hospitalized) with the data we knew at the time of diagnosis (age, sex, and comorbidities).

Finally, using clinical data, genomic and transcriptomic information, we have developed an early-stage Alzheimer's biomarker (chapter 6). Alzheimer's is a complex disease and one in which it is challenging to obtain an early diagnosis. Moreover, due to its symptoms, it is hard to distinguish from other similar conditions and even confirm until the patient's death. Using transcriptomic information derived from free RNA in blood plasma and ML techniques, we have developed a biomarker that detects Alzheimer's in early stages with high accuracy. We have studied its performance in different stages of the disease, obtaining promising results. Finally, we determined its specificity by comparing it with the result obtained in other neurodegenerative diseases.

CAPÍTULO 4

PhenoExam: Un paquete R para el análisis de enriquecimiento de fenotipos

Las preguntas e hipótesis que nos hacemos en el capítulo 1 pretenden ser contestadas en los capítulos 4, 5 y 6. En estos capítulos se desarrollan diferentes investigaciones. En cada capítulo se introduce en mayor profundidad que en el capítulo 1 el tema y los problemas que pretendemos solucionar.

Particularmente, en este capítulo 4 presentamos la herramienta PhenoExam que hemos desarrollado para estudiar las relaciones entre genes y fenotipos. Como hemos tratado en la introducción, un fenotipo es una característica visible, por ejemplo, tenemos más generales como la altura de una persona o más específicos como la trombocitopenia. PhenoExam es un paquete desarrollado en R para el análisis de fenotipos en conjuntos de genes que tiene como objetivo detectar términos fenotípicos que son significativos en un conjuntos de genes. PhenoExam posibilita el análisis de enriquecimiento de fenotipos y enfermedades en un conjunto de genes; mide de forma estadística las similitudes de fenotipos entre diferentes conjuntos de genes y detecta fenotipos diferenciales entre los conjuntos de genes que se comparan. En este caso concreto hemos utilizado la bioinformática, la estadística y diferentes fuentes de datos gen-fenotipo para crear una nueva herramienta en el lenguaje R que pueda servir tanto a bioinformáticos, como a clínicos y genetistas. Se puede instalar como paquete en R a través de su GitHub [1] y

[1] https://github.com/alexcis95/PhenoExam

también se puede acceder a la utilidad web para usuarios [2].

PhenoExam ha sido publicada en la revista BMC Bioinformatics [3]. Con esta herramienta, pretendemos cubrir diferentes aspectos del análisis ómico. Por ejemplo, ¿Qué información podemos extraer a nivel fenotípico y de relación con enfermedades de un conjunto de genes? ¿En cuanto se parecen dos conjuntos de genes a nivel de sus consecuencias visibles o fenotipos? ¿Podemos distinguir en que se diferencian enfermedades similares utilizando sus conjuntos de genes relacionados? En este caso, nos basaremos en métodos estadísticos. Analizamos la sobrerepresentación de términos fenotípicos y la compararemos con la que muestran diferentes grupos de genes escogidos aleatoriamente. Resumiendo, con este planteamiento podemos determinar cuanto de sobrerepresentados están esos términos fenotípicos y si las similitudes que muestran dos conjuntos de genes son significativas. Todo esto nos sirve para comparar y obtener información valiosa de los conjuntos de genes.

4.1 Introducción

En la investigación en genética clínica uno de los principales objetivos es descubrir nuevas asociaciones gen-enfermedad [59, 60, 61, 62, 63, 64]. Una enfermedad se suele diagnosticar mediante la identificación de un conjunto de síntomas y signos asociados a un fenotipo clínico concreto y reconocido [65, 66, 67]. Mientras que algunos fenotipos se deben a la interacción de diferentes factores ambientales, si una enfermedad tiene base genética entonces la variación genética de ese individuo puede explicar parte o la totalidad del fenotipo que se observa [68]. Un ejemplo de enfermedad estrictamente hereditaria es la fibrosis quística. La fibrosis quística afecta principalmente los pulmones y el sistema digestivo, haciendo que los mocos producidos por el cuerpo sean muy espesos y pegajosos produciendo daño en los pulmones y mayor propensión a infecciones. Se sabe que la fibrosis quística está causada por mutaciones en el gen CFTR [69] y, por tanto, los factores ambientales son despreciables o no tienen importancia para su desarrollo. Por otro lado, la mayoría de enfermedades tienen un componente hereditario o genético, pero eso no significa que se clasifiquen como hereditarias en el sentido estricto de la palabra. De hecho, la mayoría de las enfermedades son multifactoriales, lo que significa que son el resultado de una interacción entre genes y factores ambientales, el genotipo y el ambiente [70]. Por ejemplo, enfermedades como la diabetes tipo 2, el cáncer o el alzheimer tiene

[2] `https://alejandrocisterna.shinyapps.io/phenoexamweb/`
[3] `https://doi.org/10.1186/s12859-022-05122-x`

un componente genético [71]. Las personas que tienen ciertos genotipos pueden ser más propensas a desarrollar estas enfermedades, pero también son fuertemente influenciadas por factores de estilo de vida y ambientales, como la dieta, la actividad física, el consumo de tabaco, el alcohol, y la exposición a ciertas sustancias químicas.

En este capítulo presentamos PhenoExam, una herramienta bioinformática que puede ayudar a identificar nuevas asociaciones gen-fenotipo y a mejorar el diagnóstico clínico de las mismas. Puede identificar nuevas posibles asociaciones gen-fenotipo por medio del estudio de conjuntos de genes que PhenoExam relaciona estadísticamente con un fenotipo. A partir de esos resultados los investigadores pueden tomar otros genes de se conjunto que todavía no está determinada esa relación para estudiarlos en profundidad. Esto puede ser un primer paso para centrarse en unos genes más probables de que pueda existir la relación gen-fenotipo concreta. PhenoExam se centra en cualquier enfermedad que tenga un componente genético, por tanto, puede servir tanto para enfermedades estrictamente hereditarias y para multifactoriales. PhenoExam aprovecha los recursos disponibles de anotación gen-fenotipo para proporcionar una serie de análisis sobre conjuntos de genes y fenotipos que el usuario puede utilizar para diversos propósitos.

Durante la última década, hemos visto intentos de estandarizar nuestro conocimiento de las enfermedades genéticas mediante la vinculación o asociación de genes a fenotipos concretos mediante la utilización de una terminología estandarizada, por ejemplo encontramos la previamente mencionada HPO [42] para humanos y la *Mouse Genome Database* (MGD) para ratón [72]. HPO es un conjunto estandarizado de términos fenotípicos humanos que se organizan jerárquicamente con un gráfico acíclico y se ha utilizado para anotar las entradas clínicas de la base de datos de OMIM. OMIM es un catálogo continuamente actualizado de genes humanos, enfermedades genéticas y rasgos, con un enfoque particular en la relación molecular entre la variación genética y fenotípica [39]. Por otro lado, MGD es la representación consensuada y curada manualmente de la información genotipo-fenotipo, incluyendo información detallada sobre genes y productos génicos. Es la fuente autorizada de conjuntos de datos biológicos de referencia relacionados con genes de ratón, funciones génicas, fenotipos y modelos de ratón de enfermedades humanas. MGD tiene más términos e información fenotípica detallada que HPO porque los científicos pueden realizar un conjunto más amplio de experimentos con ratones. Estas características aumentan nuestros conocimientos y pueden ayudar a priorizar o detectar posibles nuevas relaciones gen-fenotipo en humanos. Más allá de las bases de datos de fenotipos, PhenoExam también incluye bases de datos de asociación gen con

términos de enfermedad, concretamente UniProt [73], The Comparative Toxicogenomics Database (CTD) [74], Orphanet [28], The Clinical Genome Resource (ClinGen) [75], The Genomics England PanelApp [76], The Cancer Genome Interpreter (CGI) [77] y PsyGeNET [78]. A continuación las describimos brevemente:

- **UniProt**: Es una base de datos de información sobre proteínas, curada manualmente y con anotaciones precisas. La base de datos contiene una gran cantidad de información sobre la función de la proteína, las secuencias, las enfermedades asociadas, las interacciones, etc.

- **CTD**: Proporciona información detallada sobre la relación entre los productos químicos, las enfermedades y los genes. Proporciona herramientas para visualizar y analizar cómo las sustancias químicas, los genes y las enfermedades interactúan entre sí. Los datos son curados manualmente a partir de la literatura científica por expertos.

- **Orphanet**: Es una base de datos en línea dedicada a las enfermedades raras. Esta base de datos se estableció en Francia en 1997 con el objetivo de mejorar el diagnóstico, tratamiento y cuidado de los pacientes con enfermedades raras. Orphanet también proporciona información sobre las asociaciones entre enfermedades raras y genes específicos.

- **ClinGen**: Es un recurso orientado a profundizar nuestro entendimiento de cómo las variantes genéticas influyen en la salud humana. Esta meta se alcanza a través del desarrollo de una base de datos acreditada que recoge genes, variantes genéticas y su importancia para la medicina y relación con enfermedades, entre otras cosas.

- **The Genomics England PanelApp**: Es una herramienta online y una base de datos que permite a revisar y documentar paneles de genes para diagnósticos genómicos. Los paneles de genes se crean y actualizan a través de un proceso de revisión por pares.

- **CGI**: Es una herramienta que tiene como objetivo ayudar a los investigadores en la interpretación de las variantes genéticas relacionadas con tumores. Puede proporcionar información sobre si una variante específica se ha asociado previamente con el cáncer, si es posible que sea patogénica y qué implicaciones podría tener para el tratamiento del cáncer.

- **PsyGeNET**: Es una base de datos y una herramienta accesible que se dedica a la exploración de la genética de enfermedades psiquiátricas. Recopila y organiza información de la literatura científica sobre genes y variantes genéticas que se han asociado con trastornos psiquiátricos.

PhenoExam también incluye CRISPRbrain [79], el primer cribado de interferencia de repeticiones palindrómicas cortas agrupadas y regularmente espaciadas de DNA, *Clustered Regularly Interspaced Short Palindromic Repeats* en inglés (CRISPR) y activación CRISPR de todo el genoma en neuronas humanas para que podamos estudiar la posible asociación de términos fenotípicos a funciones específicas de estos genes en neuronas humanas. La tecnología CRISPR se ha convertido en una herramienta esencial en genética por su capacidad para editar genes de manera precisa. CRISPR también puede usarse para controlar la actividad de los genes sin cambiar su secuencia. A esto se le conoce como activación CRISPR. Se puede aumentar o disminuir la transcripción del gen y observar su comportamiento. CRIPRBrain recoge varios de estos experimentos con genes y neuronas. En definitiva, es un recurso muy interesante para usar en enfermedades con base genética relacionadas con neurología.

Además de ser una herramienta de uso general para la anotación de conjuntos de genes basada en el fenotipo, PhenoExam también puede ayudar en el diagnóstico de enfermedades genéticas. En la actualidad, menos de la mitad de los pacientes con sospecha de trastornos mendelianos (enfermedades genéticas debidas principalmente a alteraciones en un gen) reciben un diagnóstico molecular [80]. Las enfermedades con base genética se suelen diagnosticar buscando mutaciones causales en un panel de genes específicamente asociados a la enfermedad. Un panel de genes es una recopilación de genes que tienen relación con un determinado fenotipo o enfermedad. Estos paneles pueden ser muy útiles en el diagnóstico de enfermedades con base genética, particularmente aquellas que pueden ser causadas por mutaciones en alguno de los genes recogidos en el panel. Por ejemplo, hay paneles de genes para diferentes tipos de cáncer, trastornos del desarrollo neurológico, enfermedades cardíacas hereditarias, trastornos del espectro autista, y muchas otras condiciones. Estos paneles se diseñan seleccionando un conjunto de genes que están asociados por estudios con la enfermedad en cuestión. Las pruebas de panel de genes se utilizan a menudo cuando un individuo tiene síntomas de una enfermedad y el médico sospecha una causa genética, pero no está claro cuál de varios genes posibles podría ser la causa. La recopilación de todos los fenotipos asociados a los genes de un panel proporciona una descripción general a nivel de fenotipo más allá de la

Tabla 4.1: Comparativa de herramientas para el análisis de enriquecimiento de fenotipos.

Tool	*As web*	*As software tool*	*Open source*	*Model Organism*	*Phenotype sets*	*Gene sets*	*Multiple database at once*	*Phenotype Enrichment Analysis*	*Disease Enrichment Analysis*	*Differential phenotypes*	*Diagnosis based on phenotypes*	*Similarity scores*
PhenoExam	X	X	X	X	X	X	X	X	X	X	-	X
modPhEA	X	-	-	X	X	X	-	X	-	X	-	-
DisGeNET	X	X	X	-	X	X	X	-	X	-	-	-
Phenomizer	X	-	-	-	X	-	-	-	-	-	X	*
HPOSim	-	X	X	-	X	X	-	X	-	-	-	*
PhenoSim Web	X	-	-	-	X	X	-	-	-	-	-	*

enfermedad estudiada. Para mejorar la precisión del diagnóstico genético, necesitamos métodos que evalúen adecuadamente la similitud fenotípica a nivel de genes entre las enfermedades candidatas o sus paneles de genes. Además, la identificación de fenotipos diferenciales entre enfermedades también puede contribuir a un diagnóstico más preciso. La identificación de fenotipos exclusivos y/o compartidos entre paneles de genes puede demostrar una fisiopatología común o diferente [81], pero también puede ayudar a crear vínculos genéticos entre enfermedades a través de sus conjuntos de genes [82, 83].

4.2 Estado del arte

Ahora pasamos a revisar herramientas que realizan tareas similares a PhenoExam aunque no iguales. Podemos encontrar numerosos métodos basados en la medición de similitudes fenotípicas en enfermedades mediante la comparación de conjuntos de términos de fenotipo procedentes de HPO, por ejemplo, Phenomizer [84], HPOSim [85], y PhenoSimWeb [86], la tabla 4.1 ofrece una comparación detallada entre todas las herramientas. Todas estas utilizan términos de fenotipo y los comparan para ver cuanto de similares son, independientemente de los genes. También encontramos modPhEA [87], un recurso online para el análisis de enriquecimiento de fenotipos. modPheEA ayuda con el análisis de enriquecimiento de fenotipos basado en genes, pero sólo se centra en una base de datos de fenotipos a la vez y sin considerar análisis condicionales (dos conjuntos de genes que pueden ser obtenidos el uno del otro por diferentes técnicas). La herramienta modPheEA es quizás el recurso más parecido a PhenoExam ya que trabaja con listas de genes pero tiene algunas peculiaridades que no lo hacen tan completo (Tabla 4.1).

Phenomizer obtiene la similitud semántica entre conjuntos de fenotipos basándose en la ontología HPO, pero no se basa en el uso de los genes implicados en cada fenotipo. HPOSim es un paquete R que implementa medidas de similitud semántica basadas en

ontologías ampliamente utilizadas para cuantificar las similitudes entre fenotipos, y análisis de enriquecimiento a nivel de fenotipo utilizando un test hipergeométrico y el método NOA [88]. PhenoSimWeb es una herramienta en línea para medir y visualizar las similitudes de fenotipos utilizando HPO, utiliza una medición basada en el contenido de la información y explota el algoritmo PageRank [89]. Sin embargo, estas herramientas no tenían en cuenta algunos conceptos importantes. PhenoExam contribuye a este campo con nuevas características. Entre ellas se incluye la capacidad de detectar fenotipos diferenciales entre dos conjuntos de genes: fenotipos que son estadísticamente significativos sólo dentro de un conjunto de genes y no en otro. Esto es útil para detectar términos fenotípicos clave que hacen diferente a un conjunto de genes de otro, lo que puede ayudar a distinguir mejor entre enfermedades similares si tomamos las mismas como dos conjuntos de genes.Por ejemplo, esto es interesante en el caso de enfermedades parecidas en las que se comparte muchos genes como el caso del Parkinson y la distonía (Caso 1 en la sección 4.4.3), que son muy similares, pero que se logran detectar fenotipos diferenciales que pueden ayudar a distinguirlas. PhenoExam también combina términos de fenotipo y enfermedad. Esto es importante para vincular fenotipos a enfermedades específicas. Por último, intenta facilitar la interpretación de los resultados del análisis fenotípico utilizando métricas sencillas para clasificar los términos significativos, así como mensajes de resumen y gráficos interactivos. Además de las herramientas mencionadas, también encontramos una plataforma de gestión del conocimiento que integra y estandariza datos sobre genes asociados a enfermedades procedentes de múltiples fuentes llamada DisGeNET [90]. Aunque es similar a PhenoExam en la búsqueda de asociaciones gen-enfermedad, DisGeNET no ofrece, sin embargo, facilidades para el análisis de enriquecimiento del fenotipo basado en genes o para detectar similitudes fenotípicas en los análisis condicionales entre dos conjuntos de genes. PhenoExam utiliza como sustrato básico para las asociaciones gen-fenotipo y gen-enfermedad una serie de bases de datos configurables tanto en humano como en ratón que el usuario puede personalizar y adaptar en función del tipo de análisis a realizar. En PhenoExam, la similitud fenotípica entre dos grupos de genes se realiza evaluando la significación estadística del cociente de solapamiento fenotípico creado en PhenoExam, *Phenotypic Overlap Ratio* en inglés (POR), entre los fenotipos comunes que son significativos dentro de cada conjunto de genes (es decir, el número de fenotipos comunes enriquecidos entre los conjuntos de genes).

En definitiva, hemos desarrollado PhenoExam con la intención de ayudar a una gran variedad de usuarios, principalmente médicos, biólogos computacionales (profesio-

nal que combina el conocimiento en biología con técnicas de informática y estadística para entender y modelar sistemas biológicos complejos.) y genetistas (científico que estudia los genes, la herencia genética y la variación de organismos, aplicando su conocimiento para entender cómo los genes afectan las características de un individuo o una población.). PhenoExam acepta como entrada listas de genes. Asumimos que una enfermedad, un panel de genes o un conjuntos de genes interés derivado de otros estudios son definidos por el usuario en PhenoExam como una lista de genes. PhenoExam puede ayudar a los clínicos a encontrar fenotipos exclusivos de enfermedades entre un conjunto de posibles enfermedades genéticas cuyo diagnóstico se basa en paneles de secuenciación de genes (se presenta un ejemplo en el caso 1 en la sección 4.4.3). PhenoExam también es útil para los genetistas, ya que puede utilizarse para mejorar los paneles de genes, pero también para seleccionar con mayor precisión los genes implicados en estudios genéticos específicos (se presenta un ejemplo en el caso 2 en la sección de resultados). Por último, los biólogos computacionales pueden utilizar PhenoExam para descubrir nueva información sobre conjuntos de genes de interés gracias a la integración de múltiples bases de datos de fenotipos y para comparar fenotipos entre genes conocidos asociados a una enfermedad para la validación de genes asociados a enfermedades predichos con ML (se presenta un ejemplo en el caso 2 en la sección 4.4.4).

4.3 Métodos

Esta es la primera sección de métodos de la tesis. Para esta investigación se especifican los métodos empleados para desarrollar PhenoExam. El resto de capítulos en los que se explican las investigaciones desarrolladas tienen sus determinadas secciones de métodos específicas, concretamente la sección 5.2 para el capítulo 5 y la sección 6.2 para el capítulo 6.

4.3.1 Integración del acceso a las bases de datos

El conjunto de análisis realizados por PhenoExam se basa en lenguajes de fenotipos curados manualmente como HPO o MGD y las relaciones de genes-enfermedad emanadas de OMIM y recogidas en bases de datos como CTD. Por otro lado, también bases de datos basadas en cribado como CRISPRBrain, entre otras muchas que se han descrito en la introducción y que se resumen en la tabla 4.2. Teniendo en cuenta estos recursos PhenoExam puede realizar una gran variedad de análisis que son resumidos en la figura

4.1.

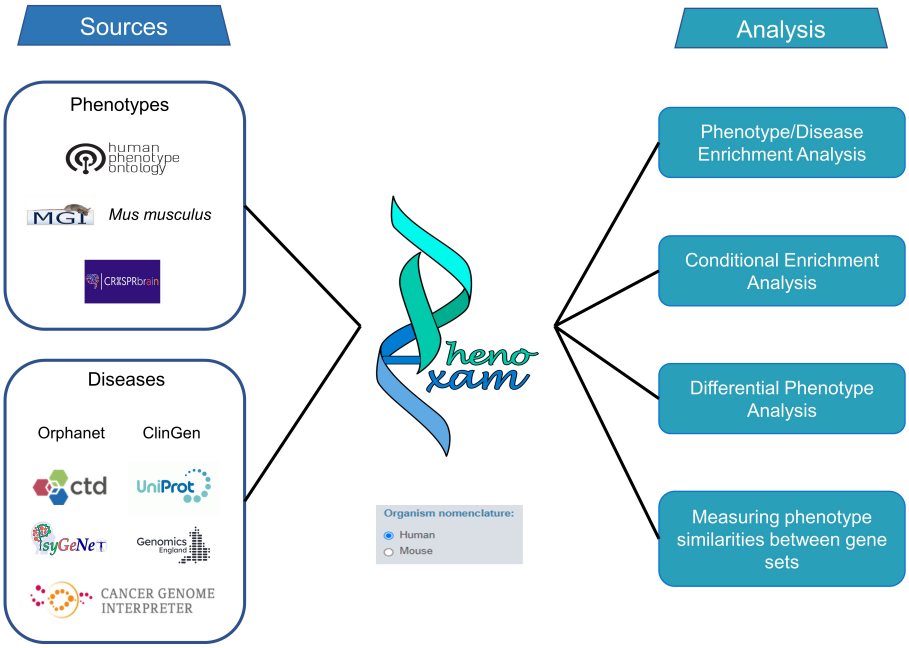


Figura 4.1: Presentación de las funcionalidades de PhenoExam

Hemos diseñado un proceso para integrar el acceso a las diferentes fuentes con la finalidad de conservar las particularidades de cada una. Por tanto, sin la finalidad de homogeneizar la semántica utilizada para las diferentes bases de datos y ontologías.

La integración del acceso a estas diferentes bases de datos es posible gracias a un proceso bien establecido de estandarización de genes y fenotipos. Utilizando el sistema de nomenclatura de genes del Comité de Nomenclatura de Genes HUGO [91], *HUGO Gene Nomenclature Committee* en inglés (HGNC) como forma común de identificar todos los genes humanos. Esto es necesario para homogeneizar el problema de búsqueda de genes. Un gen puede tener distintos símbolos, llamados sinónimos, y los mismos están contemplados como posibles opciones para acceder a ellos dentro de PhenoExam. Además, hemos definido un nuevo término de anotación dentro de cada base de datos de anotación para indicar los genes que no tienen ningún término de fenotipo o entrada asociada en la base de datos de interés. Esto es importante ya que creemos que los investigadores pueden consultar conjuntos de genes amplios y genes que todavía no hay una asociación. Hay unos 12000 genes codificantes de los que no se conoce asociación con términos en las diferentes fuentes integradas, sin embargo, es crucial tenerlos en cuenta para facilitar a

los investigadores la asociación de estos genes a términos de fenotipo significativos del conjunto de genes.

A continuación se recogen las direcciones y fuentes consultadas. La lista de genes HGNC se obtuvo de `https://www.genenames.org/download/statistics-and-files/`.

La lista de asociación gen-fenotipo HPO se obtuvo de `https://archive.monarchinitiative.org/latest/tsv/gene_associations/`. El nuevo término de asociación sin fenotipo (definido como HPO:XXX Sin fenotipo HPO) se añadió para todos los genes codificantes de proteínas sin asociación conocida con el fenotipo en HPO.

Para la base de datos de ratón es importante conocer a lo que nos referimos cuando hablamos de genes ortólogos, simplificando son genes que presentan una alta homología entre dos especies distintas. Para MGD, la base de fenotipos de ratón, los términos MP de genes ortólogos a humanos se obtuvieron de `http://www.informatics.jax.org/downloads/reports/index.html#go`, y la relación entre genes humanos y el fenotipo asociado a su ortologo en ratón se recopiló utilizando los archivos (MGI_PhenoGenoMP.rpt, HMD_HumanPhenotype.rpt, VOC_MammalianPhenotype.rpt). Se creó un nuevo término para la asociación sin fenotipo (MP:XXX Sin fenotipo) y todos los genes sin relación con el fenotipo se vincularon a este término.

Para CRISPRBrain, las relaciones gen-fenotipo se obtuvieron de `https://crisprbrain.org/simple-screen/`. Para la generación de esta base de datos, los fenotipos se codificaron en tres clases para cada análisis CRISPR: asociación al fenotipo (genes Positive-Hit y Negative-Hit en CRISPRBrain), asociación positiva (genes Positive-Hit en CRISPRBrain) y asociación negativa (genes Negative-Hit en CRISPRBrain). Esto se realizó de acuerdo con la etiqueta Hit-Class en CRISPRBrain (Positive-Hit, Negative-Hit). Se creó el fenotipo no relacionado (CRB:XXX No phenotype) y todos los genes que no estaban relacionados con ningún fenotipo se relacionaron con este término.

Integramos en PhenoExam la información de las bases de datos curadas (UniProt, CTD, Orphanet, ClinGen, The Genomics England PanelApp, CGI y PsyGeNET). A continuación, se creó el término de enfermedad sin relación (CXXX No diseases associated) y todos los genes que no estaban relacionados con ninguna enfermedad se relacionaron con este término.

Tras el proceso de integración y creación del fenotipo, la versión actual (v1.0) de

Tabla 4.2: Bases de datos utilizables en PhenoExam

Source	Genes	Phenotypes	Diseases	Assocs	Summary
HGCN	19.197	–	–	–	All protein coding genes
HPO	19.248	7.861	–	186.290	Human gene-phenotype associations
MGD	17.900	10.243	–	242.313	Mouse gene-phenotype associations
CRISPRBrain	19.275	55	–	43.481	Cell screen gene-phenotype associations
ClinGen	19.198	–	420	19.851	Human gene-disease associations
Genomics England	19.230	–	5.538	24.336	Human gene-disease associations
CTD	19.636	–	6.843	58.660	Human gene-disease associations
CGI	19.198	–	177	20.361	Human gene-disease (cancer) associations
UniProt	19.204	–	3.868	21.101	Human gene-disease associations
Orphanet	19.262	–	3.183	2.228	Human gene-disease (rare) associations
PsyGeNET	19.248	–	82	20.952	Human gene-disease associations
ALL	**20.209**	**18.159**	**9.348**	**544.022**	**PhenoExam tool**

PhenoExam utilizando todas las asociaciones de las bases de datos integradas contiene 659.634 asociaciones gen-fenotipo, que implican a 20.209 genes, 18.159 fenotipos diferentes y 9.348 enfermedades distintas (véanse los detalles en la tabla 4.2).

4.3.2 Métodos de los distintos análisis en PhenoExam

En las subsecciones siguientes vamos a desarrollar los métodos para los diferentes análisis que se pueden realizar con PhenoExam. Hemos querido utilizar una variedad de métodos y pruebas estadísticas para ofrecer al usuario la posibilidad de utilizar diferentes análisis que mejor se adapten a su propósito.

4.3.2.1 Análisis de enriquecimiento fenotípico sobre un conjunto de genes G

Supongamos que tenemos un conjunto de genes de los que queremos saber cuales son sus fenotipos prominentes o más representados, es decir, qué términos fenotípicos están

más enriquecidos. Esto es de utilidad, por ejemplo, cuando tenemos un panel de genes y queremos saber cuales son los términos de fenotipo o enfermedades relacionadas con ese panel. También cuando tenemos un conjunto de genes que se sabe están más expresado para una determinada condición en un experimento de RNAseq y se quiere consultar que fenotipos los caracteriza. Pues bien, vamos a explicar las diferentes formas que PhenoExam utiliza para calcular esto.

PhenoExam calcula los fenotipos estadísticamente significativos en un determinado conjunto de genes G dentro de una anotación de una base de datos de fenotipos/enfermedades de referencia D. Para calcular si un conjunto de genes G muestra enriquecimiento en un determinado término fenotípico p perteneciente a D, sea g el número de genes de G asociados a p. Sea también gdb el número de genes asociados a p y GDB el número total de genes de la base de datos, modelamos la probabilidad de enriquecimiento con una distribución hipergeométrica que se muestra en la ecuación 4.1. Este test es útil cuando queremos comparar el número de éxitos (en este caso, apariciones de un fenotipo específico) en una muestra (nuestro conjunto de genes) con lo que esperaríamos basándonos en una población más grande. Por ejemplo, puedo tener un conjunto de 200 genes que se están expresando fuertemente que he obtenido de un experimento de RNAseq y que será nuestro conjunto G. Saber que 20 genes g de ese conjunto están asociados a un determinado fenotipo p del que se sabe que hay 400 asociaciones gdb en la base de datos para un total de 19000 genes de los que hay registro en esa base de datos GDB.

$$P(X = g) = \frac{\begin{pmatrix} gdb \\ g \end{pmatrix}\begin{pmatrix} GDB - gdb \\ |G| - g \end{pmatrix}}{\begin{pmatrix} GDB \\ |G| \end{pmatrix}} \tag{4.1}$$

En este caso, el valor P obtenido de la prueba hipergeométrica es la probabilidad de obtener g, asociaciones gen-fenotipo específico, de nuestro conjunto de genes G, dado que la base de datos contiene gdb asociaciones fenotipo-gen en total en una población de GDB posibles genes. Cualquier fenotipo con P <0,05 estará enriquecido en el conjunto de genes G en comparación con lo que esperarías ver si los genes se seleccionaran al azar de la población general. Calculamos esta probabilidad para cada término fenotípico ph asociado con un gen o más en G y utilizamos estas probabilidades como valores P.

Es importante detenernos en los problemas generado por pruebas múltiples, lo que se conoce como *multiple testing* en inglés. Cuando hacemos este tipo de pruebas para un conjunto de genes y una base de datos de fenotipo hay que tener en cuenta que existen muchos fenotipos y posibles anotaciones que tenemos que consulta. Por lo tanto estamos realizar miles de pruebas de hipótesis al mismo tiempo. Uno de los problemas principales asociados con las pruebas múltiples es la del error de tipo I, que es la probabilidad de rechazar incorrectamente una hipótesis nula verdadera (falso positivo). La razón de esto es que si esta,ps probando miles de hipótesis y usando un umbral de significación del 5%, entonces por azar, esperaríamos que el 5% de las pruebas resultaran en un falso positivo. PhenoExam informa de los valores *P* brutos, ajustados por Bonferroni [92] y por la tasa de falsos descubrimientos, *False Discovery Rate en inglés* (FDR) [93]. Por un lado, la corrección de Bonferroni es un método muy conservador y controla la probabilidad de error de tipo I dividiendo el nivel de significancia por el número de pruebas Por otro, FDR controla la tasa de falsos descubrimientos, que es la proporción esperada de hipótesis incorrectamente rechazadas, este método es menos conservador.

4.3.2.2 Método para cálculo del POR

Supongamos que necesitamos saber si dos conjuntos de genes pueden producir fenotipos similares, esto nos puede servir para determinar si están relacionados o si producen patologías y enfermedades similares. Por tanto, tenemos dos conjuntos de genes y queremos comparar si los fenotipos importantes, es decir, los significativamente enriquecidos, son similares.

El enfoque de PhenoExam para medir la similitud entre dos conjuntos de genes G y G', dentro de una base de datos de anotaciones D, se basa en una puntuación denominada Phenotypic Overlap Ratio (POR). Sea Gp el número de términos significativamente enriquecidos en D para genes en G, y análogamente para G'p. PhenoExam ofrece que se calcule el POR de varías formas: (1) utilizando el ampliamente conocido índice de Jaccard o (2) el coeficiente de similitud de Forbes corregido por Alroy [94] sobre la concordancia entre los subconjuntos de fenotipos significativos. PhenoExam permite a los usuarios elegir entre estas dos opciones después de analizar los diversos resultados que pueden generar estas métricas publicados por *Salvatore et al.* [95].

Encontramos el índice de Jaccard en la ecuación 4.2:

$$POR(G,G') = \frac{Gp \cap G'p}{Gp \cup G'p} \tag{4.2}$$

Para el cálculo del POR utilizando el coeficiente de similitud de Forbes corregido por Alroy necesitamos las ecuaciones 4.3 y 4.4 para llegar a la ecuación 4.5:

$$N = Gp \cap G'p + Gp \setminus G'p + G'p \setminus Gp \tag{4.3}$$

(4.4)

$$A = (Gp \cap G'p + Gp \setminus G'p) \times (Gp \cap G'p + G'p \setminus Gp) + Gp \cap G'p \times \sqrt{N} + (Gp \setminus G'p \times G'p \setminus Gp) \div 2$$

$$POR(G,G') = \frac{Gp \cap G'p \times (N + \sqrt{N})}{A} \tag{4.5}$$

POR (G, G') toma valores entre [0,1], resultando 0 cuando no se comparte ningún fenotipo y 1 cuando los conjuntos comparten todos los fenotipos (índice de Jaccard) o al menos comparten todos los fenotipos de un conjunto (coeficiente de Forbes). Por lo tanto, el índice de Jaccard será más conservador y será más difícil obtener un valor de 1 que utilizando el coeficiente de Forbes, el cual puede ser más útil si los conjuntos de genes que necesitamos comparar son muy diferentes en cuanto a número de genes. Para más información se puede consultar al respecto la publicación de *Salvatore et al.* [95].

4.3.2.3 Determinar el valor de POR estadísticamente significativo

Para decir que un conjunto de genes es parecido a otro en términos fenotípicos hay que compararlos y determinar si el resultado obtenido es estadísticamente significativo. PhenoExam evalúa si el POR entre los conjuntos de genes G y G' es estadísticamente significativo mediante aleatorización. Es decir, se construyen conjuntos de genes aleatorios y se comparan con los del análisis para comprobar si lo obtenido entre los conjuntos se puede considerar estadísticamente significativo.

Tendremos dos modalidades del POR, dependiendo de si G y G' comparten genes o, por el contrario, son totalmente disjuntos por imposición (por ejemplo, G' se predijo

a partir de G). Cuando G y G' comparten genes, POR (G, G') se compara con POR (G, R) y con POR (G', R'), donde R tiene el mismo tamaño que G y R' el mismo que G'. Los genes de R y R' se eligen al azar dentro del conjunto de genes que codifican para proteínas. Repetimos este proceso para un número m de conjuntos de genes aleatorios (R1,R2,...,Rm) y (R'1,R'2,...,R'm) para obtener un valor P empírico con la proporción de conjuntos de genes aleatorios cuyo POR es mayor que el observado entre los conjuntos G y G'. Podemos por tanto definir el número total de conjuntos aleatorios que tienen un POR mayor al de los conjuntos de análisis como NRP. Utilizando una constante, en este caso un uno, para eliminar el resultado infinito cuando no hay ningún NRP y tomando m como el número de conjuntos aleatorios que se prueba en la aleatorizacion tenemos la ecuación 4.6:

$$p - valor = \frac{1 + NRP}{1 + m} \tag{4.6}$$

Por otro lado, cuando G' se obtiene utilizando G como entrada del proceso de generación, decimos que G' está condicionado a G. Por tanto, la prueba de significación del POR (G, G') se reduce ahora a obtener un valor P empírico basado en la proporción de veces que un POR aleatorio es superior pero teniendo en cuenta que el espacio de búsqueda fenotípico está reducido y que ningún gen del conjunto aleatorio puede pertenecer a G, es decir, evaluando con las mismas condiciones en las que se formó G'. Se calcula el valor de p de la misma forma utilizando aleatorización.

4.3.2.4 Cálculo del POR relajado

El POR sólo tiene en cuenta los fenotipos que se evaluaron como estadísticamente significativos. A veces, puede ser interesante relajar esta restricción para incorporar todos los términos de fenotipo/enfermedad asociados a G. Esto ocurre cuando tenemos conjuntos de pocos genes cuyo resultado de enriquecimiento muestra pocos o ningún término significativo pero queremos compararlos entre si.

En este caso, la puntuación se denomina POR relajado, *Relaxed Phenotypic Overlap Ratio* en inglés (RPOR). Se calcula de forma similar al POR pero con todos los fenotipos, estén éstos enriquecidos o no. Del mismo modo que con el POR, podemos determinar si el RPOR es estadísticamente significativo utilizando la aleatorización.

4.3.2.5 Análisis de asociación de relevancia fenotípica para conjuntos de genes

Con la finalidad de determinar si de dos conjuntos de genes de los que obtenemos fenotipos comunes esos fenotipos tienen el mismo peso, es decir, una distribución de genes parecida para cada fenotipo hemos planteado el siguiente análisis.

Una vez determinado que dos conjuntos de genes G y G' comparten cierto enriquecimiento de términos fenotípicos, y centrándonos sólo en los términos compartidos, podemos medir la correlación del número de genes de cada término fenotípico medido en G y G' mediante un modelo de regresión lineal e informar del R2 como la fuerza de esta correlación junto con el valor P de esa asociación. Los valores más altos de R2 sugerirían una asociación lineal entre la importancia de los términos fenotípicos en G y la importancia de los mismos genes en G'. Esto añade evidencia para saber si en los genes de ambos conjuntos la distribución de los fenotipos compartidos es similar.

4.3.3 Generación de la interfaz web

Hemos desarrollado PhenoExamWeb, una herramienta basada en web para realizar análisis fenotípicos utilizando R. Hemos creado una web para facilitar el uso de la herramienta para los genetistas o investigadores que no sean biólogos computacionales. PhenoExamWeb shiny app es accesible en `https://alejandrocisterna.shinyapps.io/phenoexamweb/`. R y el paquete shiny R [96] se utilizaron para la generación de scripts front-end de la interfaz web. Los scripts de R se utilizaron para la ejecución y el análisis back-end con el entorno de desarrollo de la versión 3.6.3 de R. El paquete R está disponible en `https://github.com/alexcis95/PhenoExam`. Es relevante que aunque ofrecemos PhenoExam a través de una aplicación web, podría ser una mejor opción considerar la instalación y el uso del paquete R localmente en aras de la flexibilidad o desplegar la aplicación shiny localmente en su estación de trabajo local para análisis computacionalmente más exigentes como, por ejemplo, un análisis de comparación fenotípica con más de 40 pruebas aleatorias. Basta con descargar el software de `https://github.com/alexcis95/PhenoExam/blob/master/PhenoExamWeb.zip` y ejecutar el archivo Rmd localmente.

4.3.4 Análisis utilizando la web de PhenoExam

La web de PhenoExam requiere símbolos de genes, humanos o de ratón, como archivo de entrada. A continuación, hay que seleccionar el tipo de análisis: Análisis de enriquecimiento de fenotipos (un conjunto de genes), llamado *Phenotype Enrichment Analysis* en inglés, o análisis de comparación fenotípica (necesita dos conjuntos de genes), *Phenotype Comparator* en inglés. También hay que especificar la base o bases de datos que queremos elegir. El flujo de trabajo de la web de PhenoExam se resume en la figura 4.2. Los usuarios pueden seguir el tutorial web en el sitio web `https://alejandrocisterna.shinyapps.io/phenoexamweb/#section-help` y el tutorial del paquete R en GitHub `https://raw.githack.com/alexcis95/PhenoExamWebTutorials/main/tutorial.html`.

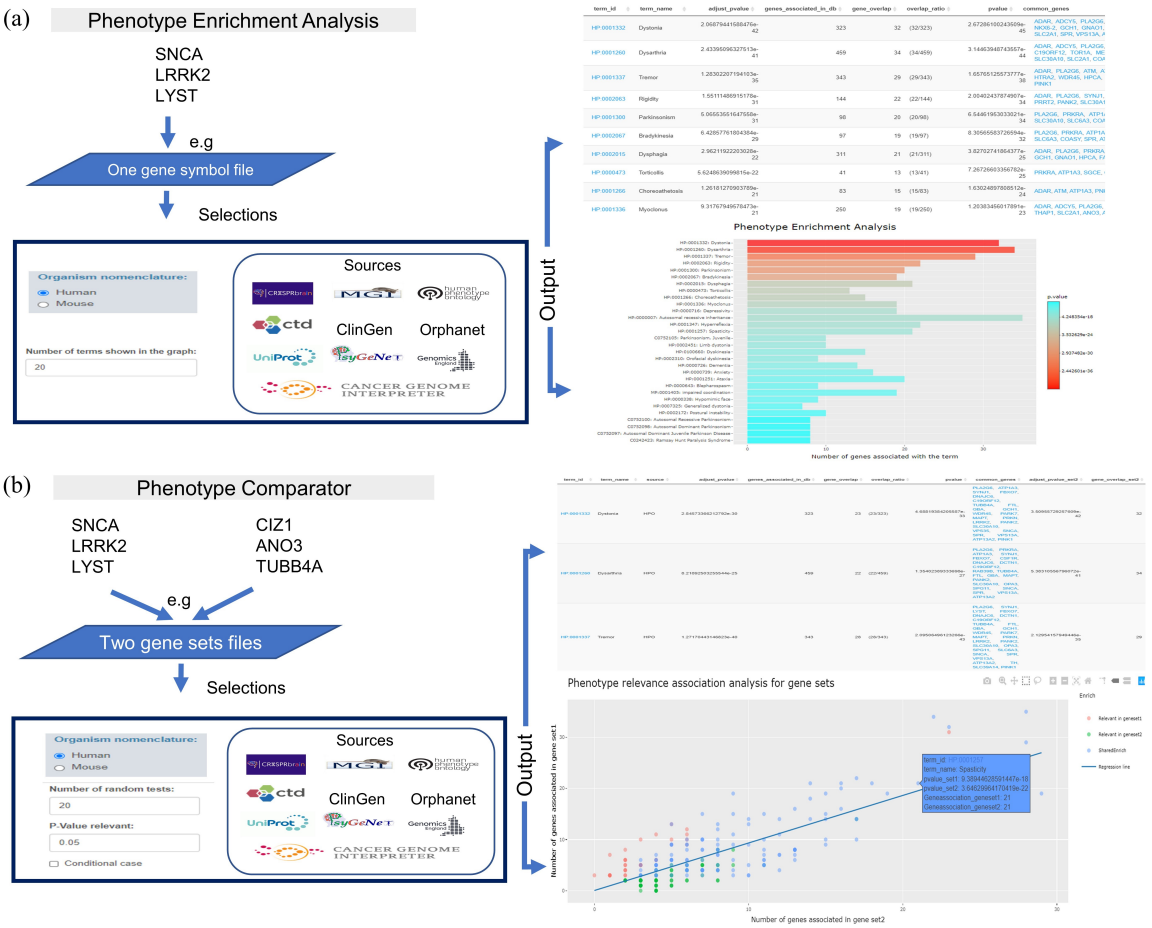


Figura 4.2: Flujo de trabajo en la web de PhenoExam

En la figura 4.2 se muestran los posibles flujos en la web PhenoExam. En la sección (a) Análisis de Enriquecimiento de Fenotipos: requiere un archivo de símbolos de genes

como archivo de entrada, la nomenclatura de símbolos de genes (nomenclatura de Organismo: Humano o Ratón), las bases de datos de anotación de fenotipos/enfermedades a considerar y el número máximo de términos mostrados en el gráfico. Los resultados generan una tabla y un gráfico interactivos que incluyen los fenotipos, los genes implicados con cada término y los valores P como salida. En la sección (b) el Comparador de Fenotipos requiere dos conjuntos de genes como entrada junto con la nomenclatura de símbolos de genes (Humano o Ratón) utilizada, las bases de datos de anotación de interés para el análisis y el número de pruebas aleatorias para obtener valores P empíricos, se recomienda que ese número de pruebas sea al menos de 1000, el umbral de valor P relevante y si nuestro análisis es un caso condicional (es decir, si un conjunto de genes se generó tras un análisis de predicción del otro y son conjuntos de genes totalmente diferentes). Por último, obtenemos el resumen del análisis con las puntuaciones de los fenotipos de similitud, los fenotipos diferenciales, tablas interactivas y gráficos con fenotipos, genes y valores P como salida para la inspección detallada y la presentación de resultados.

4.4 Resultados

Esta es la primera sección de resultados de la tesis. A continuación se desarrollan los resultados obtenidos de evaluar la fiabilidad de PhenoExam (subsecciones 4.4.1 y 4.4.2) y de su utilización con ejemplos de caso de uso reales (subsecciones 4.4.3 y 4.4.4).

El resto de capítulos en los que se explican las investigaciones desarrolladas tienen sus determinadas secciones de resultados específicas, concretamente la sección 5.3 para el capítulo 5 y la sección 6.3 para el capítulo 6.

4.4.1 PhenoExam controla el error de tipo I

Hemos considerado importante evaluar la fiabilidad de PhenoExam y el error esperado que pudiera cometer, error derivado de la utilización de bases de datos y pruebas estadísticas. En el contexto de pruebas de hipótesis estadísticas, un error de tipo I se define como el rechazo erróneo de una hipótesis nula verdadera. Este error se produce cuando se infiere erróneamente la existencia de un efecto o una diferencia que, en la realidad de los datos subyacentes, no está presente. El nivel de significancia, denotado como alfa, en una prueba de hipótesis estadística, representa la probabilidad de incurrir en un error de tipo I. Esto se interpreta como la probabilidad de rechazar erróneamente la hipótesis

nula a pesar de su veracidad. Por ejemplo, un nivel de significancia de 0.05 indica que existe un riesgo del 5% de rechazar la hipótesis nula cuando, de hecho, esta es verdadera. Cabe destacar que estos errores forman parte inherente de los análisis estadísticos y no necesariamente indican fallos en los métodos de análisis. No obstante, es de suma importancia minimizar la probabilidad de estos errores mediante la selección de un nivel de significancia adecuado, y asegurando la idoneidad, calidad y volumen de los datos utilizados.

Evaluamos PhenoExam para el error de tipo I dadas todas las bases de datos de fenotipos o enfermedades consideradas en la tarea de análisis de enriquecimiento fenotípico usando conjuntos de genes. En primer lugar, evaluamos la posibilidad de encontrar un término fenotípico erróneamente enriquecido, debido al azar, entre todos los términos de la base de datos, esto lo realizamos para conjuntos de genes de distintos tamaños. Para ello, realizamos simulaciones de análisis de enriquecimiento fenotípico para diferentes conjuntos de genes aleatorios con un número variable de tamaños de genes (5, 10, 20, 40, 80, 160, 320 y 640) con consultas a todas las bases de datos de anotación. Seleccionamos ese número de genes para estudiar el comportamiento en diferentes posibles tamaños dentro de un rango lógico para análisis biológicos. Además, nos centramos más en una cantidad pequeña de genes porque es aquí donde esperamos encontrar más errores. Para cada combinación de tamaño de conjunto de genes y base de datos se simularon 1.000 conjuntos diferentes, lo que arrojó un total de 80.000 simulaciones. En la figura 4.3 aparece una representación gráfica del resumen de los resultados.

PhenoExam mantiene el error tipo I bajo control para las bases de datos de fenotipo con un nivel de significación de 0.05 ya que el número de pruebas significativas está siempre muy por debajo de 0.05. Observamos una correlación negativa entre el tamaño del conjunto de genes y la proporción de pruebas positivas falsas, r = -0,453, P = 0,026. Esto es algo esperado, al tener menor muestra más posibilidad de errores. El error de tipo I es más difícil de controlar cuando se utilizan bases de datos con términos de enfermedad como Genomics England Panel App (GEL) y Orphanet. PhenoExam sólo controla el error de tipo I cuando el tamaño del conjunto de genes es superior a 80 para Orphanet y a 180 para Genomics England. Creemos que las dificultades para mantener bajo control el error de tipo I se deben al número medio de términos de enfermedad asociados a cada gen, es decir, 4,39 para GEL y 7 para Orphanet cuando para el resto de bases de datos de enfermedades es, de media, de 17,7. Además, existe

una correlación negativa entre el número de genes por conjunto de genes aleatorios y el error de tipo I, r = -0,381, P = 0,0038. Por lo tanto, tanto el número de términos asociados a cada gen como el tamaño de los conjuntos de genes utilizados como entrada son cruciales para obtener suficientes relaciones gen-fenotipo para mantener, de esta forma, el error de tipo I bajo control. Por estas razones, recomendamos utilizar CTD, HPO, MGD o CRB para los análisis que implican conjuntos de genes de tamaño 10. Estos son, aproximadamente, menos que el número de genes que podemos encontrar en muchas vías biológicas. Recomendamos utilizar PsyGeNET, ClinGen, UNIPROT o CGI con 40 genes o más. Estos suelen ser menos que el número de genes detectados en la mayoría de los estudios de asociación de genoma completo. Sólo se recomienda la inclusión de Orphanet y GEL cuando se dispone de al menos 80 y 180 genes respectivamente. Los usuarios pueden encontrar más información sobre qué base de datos deben utilizar en `https://alejandrocisterna.shinyapps.io/phenoexamweb/#section-help`.

En la figura 4.3 encontramos la tasa de falsos positivos del enriquecimiento de términos de fenotipo y enfermedad en función del tamaño del conjunto de genes (5, 10, 20, 40, 80, 160, 320, 640) por base de datos de fenotipo/enfermedad. Como señala la simulación, CRB, HPO y MGD son perfectamente utilizables para cualquier tamaño de conjunto de genes, CTD se recomienda para tamaños de conjunto de genes superiores a 10, PsyGeNET para 20, CGI, ClinGen y Uniprot para 40, Orphanet para 80 y GEL para tamaños de conjunto de genes superiores a 180.

4.4.2 PhenoExam distingue entre conjuntos de genes con fenotipos muy similares

Evaluamos la precisión de PhenoExam a la hora de calcular el POR (detectar similitudes fenotípicas) entre conjuntos de genes comparando formas genéticas de epilepsia (261 genes del panel de epilepsia de NIMGenetics) y conjuntos de genes artificiales construidos con POR variables respecto al conjunto de genes de epilepsia original. Escogimos epilepsia por dos motivos: 1) es una enfermedad que muestra una gran variedad de fenotipos; 2) es una enfermedad de interés para nuestro propósitos. A estos conjuntos artificiales les hemos introduciendo genes adicionales con conectividad fenotípica similar a la de estos pero no asociada a epilepsia. En estos genes adicionales inyectamos un 5% de ruido con genes asociados a términos fenotípicos de epilepsia. Realizamos 1000 simulaciones para los conjuntos de genes artificiales (261 genes) construidos con diferentes proporciones de genes de epilepsia entre (0-100%) y diferentes proporciones de otros genes (0-100%).

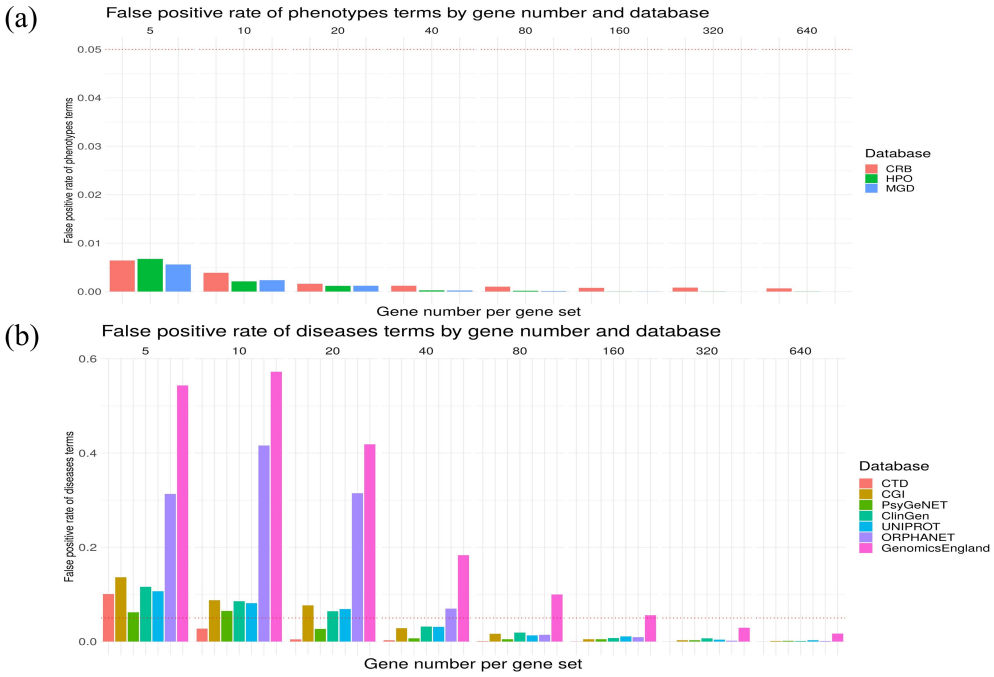


Figura 4.3: Simulaciones para detección error tipo I con PhenoExam

Calculamos la prueba de significación POR entre los conjuntos de genes reales y los artificiales que se muestra en la figura 4.4. PhenoExam es sensible en la detección de diferencias entre los cambios de composición de genes (aproximadamente el 1%) en diferentes conjuntos de genes, que en este caso ese 1% hace referencia a unos 3 genes en términos totales. Observamos una relación lineal positiva entre POR y las proporciones de genes de epilepsia en los conjuntos de genes artificiales, 0,9674 R2 ($P < 2{,}2 \times 10^{-16}$), esto se observa en la parte (a) de la figura 4.4. Evaluamos que PhenoExam puede distinguir bien entre los genes reales de epilepsia y los conjuntos de genes artificiales construidos con altas proporciones de genes de epilepsia (94-99% genes de epilepsia) los cuales constituyen conjuntos de genes con fenotipos muy similares a los del conjunto real de epilepsia pero PhenoExam muestra que puede distinguirlos con una prueba t para todos los casos mostrados en el recuadro (b) de la figura 4.4.

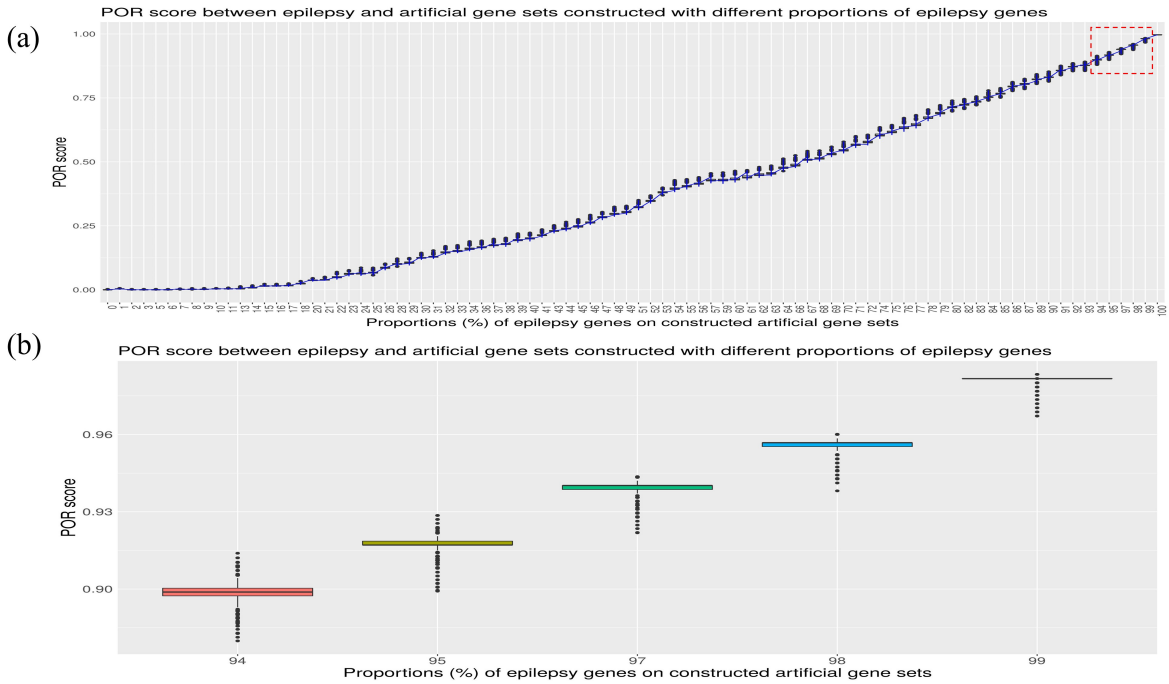


Figura 4.4: Diferencias de POR para detectar conjuntos similares a los del panel de epilepsia

4.4.3 Caso 1: El análisis entre la enfermedad de Parkinson y la distonía de inicio temprano revela que mantienen similitudes a nivel de fenotipo pero también fenotipos diferenciales potencialmente interesantes

Aplicamos PhenoExam a la detección de fenotipos diferenciales entre conjuntos de genes comparando dos enfermedades genéticas con síntomas o fenotipos similares: la enfermedad de Parkinson, *Parkison Disease* en inglés (PD), y la distonía de inicio precoz *Early Onset Dystonia* en inglés (EOD). Tanto la PD como la EOD son trastornos del movimiento, la PD está causada por una degeneración en los ganglios basales, y los síntomas predominantes consisten en temblor, rigidez, bradicinesia, inestabilidad postural y demencia progresiva [97]. La EOD es una enfermedad caracterizada por contracciones musculares involuntarias que conducen a posturas y movimientos anormales, y que ocurre con o sin otros síntomas neurológicos [98]. En nuestro caso comparamos 35 genes de PD y 50 genes de EOD del Genomics England PanelApp (los cuales se pueden consultar en la siguiente dirección `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM1_ESM.csv`), estos conjuntos comparten un total de 19 genes, es decir, en el conjunto de los genes de EOD

aparecen el 54,3 % de los genes del conjunto de genes de PD. Primero, ejecutamos un análisis de enriquecimiento de fenotipo por separado para PD y EOD, utilizando las bases de datos HPO, MGD, CTD y CRISPRBrain simultáneamente cuyos principales resultados podemos observar en la figura 4.5.

Obtuvimos una tabla con los fenotipos y el total de resultados para los genes de PD (se puede consultar en `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM2_ESM.xls`) y EOD (se pueden encontrar en `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM3_ESM.xls`). A continuación se nombran los dos fenotipos más enriquecidos para cada base de datos del análisis, para los genes de PD fueron Bradicinesia (HP: 0002067; P = 2.16×10^{-60}) y Parkinsonismo (HP: 0001300; P = 2.62×10^{-51}) para la base de datos HPO, Marcha anormal (MP: 0001406; P = $3.\ 78 \times 10^{-13}$) y Degeneración neuronal (MP: 0003224; P = 9,98 × 10-13) para MGD, Parkinsonismo juvenil (C0752105; P = $7{,}49 \times 10^{-28}$) y Síndrome de parálisis de Ramsay Hunt (C0242423; P = $7{,}49 \times 10^{-28}$) para CTD, y no se encontró enriquecimiento significativo para CRISPRBrain. Todos los términos de enriquecimiento encontrados están respaldados por la literatura [99, 100, 101, 102]. En el análisis para EOD, encontramos Distonía (HP: 0001332; P = $3{,}51 \times 10^{-42}$) y Disartria (HP: 0001260; P = $5{,}38 \times 10^{-41}$) para HPO, alteración de la coordinación (MP: 0001405; P = $7{,}4 \times 10^{-14}$) y Marcha anormal (MP: 0001406; P = $3.\ 17 \times 10^{-10}$) para MGD, Parkinsonismo juvenil (C0752105; P = $7{,}4 \times 10^{-13}$) y Síndrome de parálisis de Ramsay Hunt (C0242423; P = $7{,}4 \times 10^{-13}$) para CTD, y de nuevo ningún término enriquecido significativo para CRISPRBrain. Los términos fenotípicos mencionados anteriormente están asociados con la distonía según varios artículos [103, 104, 105, 106, 107].

Este caso es ideal para comparar los conjuntos de genes de PD y EOD, a través del análisis Comparador de Fenotipos en PhenoExam. En la figura 4.6 tenemos la configuración utilizada en la web de PhenoExam utilizando HPO, MGD, CTD y CRISPRBrain como las bases de datos seleccionadas, y para obtener los estadísticos se realizó una aleatorización basada en 1000 pruebas con conjuntos de genes de la aleatorización. Los resultados de esta comparación revelaros que los conjuntos de PD y de EOD tienen 139 términos fenotípicos significativos compartidos (de un total de 273 términos fenotípicos significativos únicos en ambos, POR = 0.509 (P <0.001). El análisis de asociación de relevancia fenotípica para PD y EOD (es decir, si los fenotipos compartidos son similares en relevancia, centrándonos en el número de genes asociados con ellos, dentro de cada conjunto de genes)

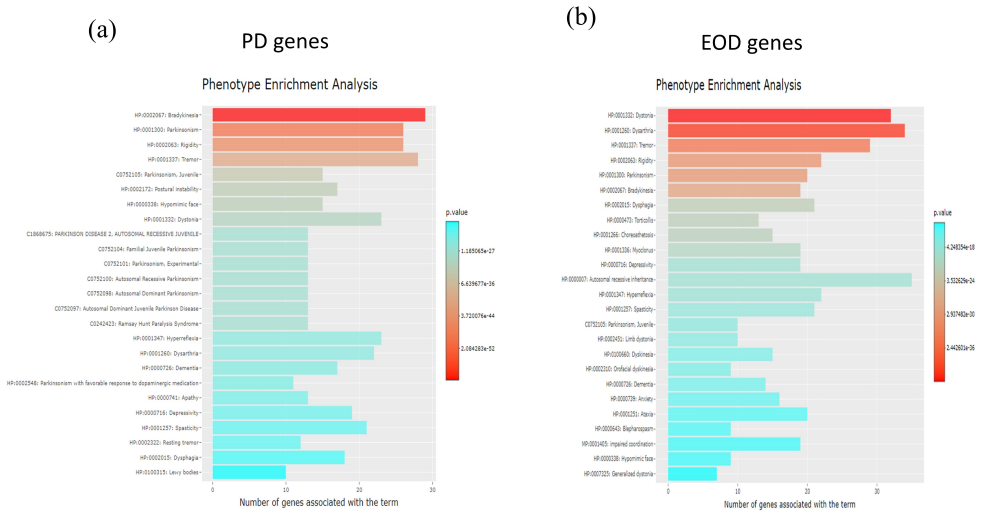


Figura 4.5: Análisis de enriquecimiento fenotípico para el conjunto de genes de Enfermedad de Parkinson (a) y para el de Distonía de Inicio Temprano (b).

resulta en un R cuadrado ajustado de 0.643 (P <9.23 × 10^{-63}) lo que sugiere que una porción importante de los fenotipos comunes son similares en relevancia entre los conjuntos. De hecho vemos que comparten términos fenotípicos como Temblor (HP: 0001337), Bradicinesia (HP: 0002067), Rigidez (HP: 0002063), Distonía (HP: 0001332), Marcha anormal (MP: 0001406) o Degeneración neuronal (MP: 0003224) (estos fenotipos compartidos se pueden consultar en `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM4_ESM.xls`). Pero también detectamos fenotipos diferenciales que pueden visualizarse mediante gráficos y tablas interactivos en la web. Por ejemplo, los términos significativos exclusivos de los fenotipos del conjunto de genes de PD incluyen Astrocitosis (MP: 0003354; P <5.17 × 10^{-12}), Gliosis de la Substantia nigra (HP: 0011960; P <4.15 × 10^{-11}), Pérdida neuronal en el sistema nervioso central (HP: 0002529; P <3. 74 × 10^{-6}), hipotensión ortostática debida a disfunción autonómica (HP: 0004926; P <9,96 × 10^{-6}) y enfermedad de los cuerpos de Lewy (C0752347; P <1,11 × 10-3) (puedes encontrar el archivo completo con todos los fenotipos significativos exclusivos de los genes de PD respecto a los de EOD EN `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM5_ESM.xls`). Los términos fenotípicos arriba mencionados están más o sólo asociados con PD de acuerdo a varios artículos

[108, 109, 110, 111, 112, 113]. El mismo análisis identificó al calambre de escritor (HP: 0002356; P <1.37 × 10^{-9}) como exclusivo de la EOD y esto se refiere a un tipo de distonía focal [114]. También encontramos Hipoplasia del cuerpo calloso (HP: 0002079; P <3,56 × 10^{-5}), un fenotipo controvertido y poco estudiado en la distonía [115, 116] y Acantocitosis (HP: 0001927; P <2,76 × 10^{-3}) un término normalmente asociado a la corea-acantocitosis, otra enfermedad con síntomas similares a la distonía [117]. La microcefalia (HP: 0000252; P <4,17 × 10^{-4}) está asociada a la distonía y a varios genes como KMT2B [118, 119]. También encontramos Discapacidad intelectual, leve (HP: 0001256; P <4,68 × 10^{-3}), Distonía, Primaria (C0752203; P <3,26 × 10^{-7}) y Reflejos tendinosos profundos hiperactivos (HP: 0006801; P <4,31 × 10^{-2}) que se asocia con Discinesia paroxística [120] (en el siguiente enlace se pueden encontrar los fenotipos significativos exclusivos de EOD respecto a PD `https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM6_ESM.xls`).

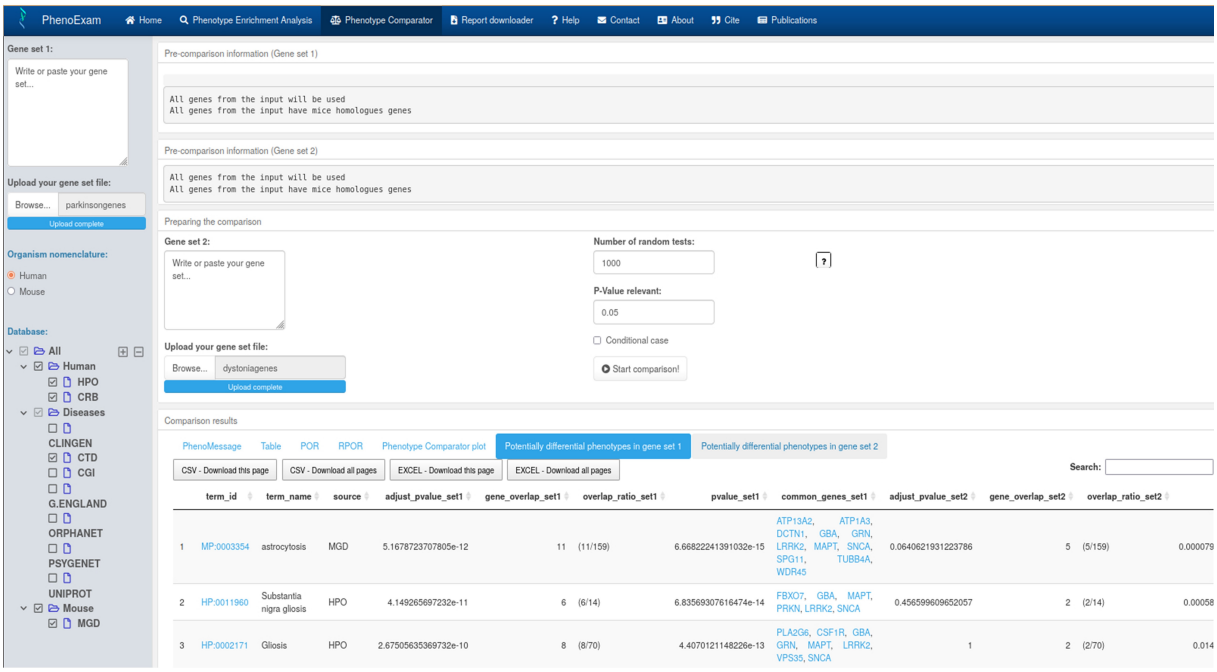


Figura 4.6: Vista del análisis comparador de fenotipos en la web de PhenoExam.

4.4.4 Caso 2: Demostrar que nuevos genes predichos utilizando el conjunto de epilepsia con G2PML recapitulan términos fenotípicos de epilepsia

Supongamos que es posible descubrir nuevos genes asociados a una enfermedad concreta (epilepsia en este caso) encontrando patrones no lineales de los genes de ese panel a

partir de su descripción mediante propiedades basadas en la genómica, transcriptómica y genética de cada gen con técnicas de aprendizaje automático. Por lo tanto, para descubrir nuevos genes, nuestro objetivo es encontrar genes muy similares en términos de esas propiedades, para ello hemos utilizado G2PML [121]). G2PML utilizada ML para descubrir genes que comparten características parecidas a las de un conjunto de genes determinado que el usuario especifica. Tiene como finalidad ayudar al descubrimiento de asociaciones entre enfermedades y genes todavía no caracterizadas y/o descritas en la literatura. Esta tarea es fundamental en el ámbito del diagnóstico genético clínico. Pero, el problema es que es una predicción y no sabemos si el conjunto de genes que propone puede tener efecto en un fenotipo o enfermedad determinado. PhenoExam pretende ayudar en esa tarea.

La pregunta a la que nos enfrentamos es: ¿recapitulan los genes que se predice que están relacionados con formas genéticas de epilepsia fenotipos similares a los genes del panel de epilepsia de origen? Cuanto más apoye la respuesta a una recapitulación del fenotipo, mejores serán las predicciones realizadas por G2PML. Este es un ejemplo de lo que llamamos un caso condicional, en el que se comparan los fenotipos de los conjuntos de genes G y G' cuando son disjuntos y G' se generó utilizando G como semilla. Más concretamente, G se refiere a genes de epilepsia de un panel de epilepsia mantenido internamente (261 genes) en la empresa cofinanciadora de la tesis NIMGenetics. Por otra parte, G' es un conjunto de 209 nuevos genes que han sido predichos utilizando el panel de NIMGenetics como semilla y G2PML, este conjunto de genes es totalmente disjunto por naturaleza ya que son nuevos genes diferentes a los utilizado como semilla.

Realizamos el análisis de comparación fenotípica en PhenoExam con la opción de caso condicional marcada, el conjunto de genes 1 fueron los genes de epilepsia, el conjunto de genes 2 fueron los nuevos genes probables de epilepsia predichos por G2PML, las bases de datos HPO, MGD, CRISPRBrain y CTD seleccionadas al mismo tiempo y elegimos 1000 pruebas aleatorias. Obtuvimos como mensaje informativo de la herramienta que estos dos conjuntos de genes compartían 106 términos fenotípicos significativos (de 734 términos fenotípicos significativos únicos en ambos), lo que arroja un POR de 0,144 (P <0,001). El análisis de asociación de relevancia fenotípica para los genes asociados a la epilepsia y los genes predichos para la epilepsia (es decir, si los fenotipos compartidos son similares en relevancia, centrándonos en el número de genes asociados a ellos, dentro de cada conjunto de genes) da como resultado un R cuadrado ajustado de 0,331 (P $<4{,}35 \times 10^{-66}$) que sugiere que una parte importante de los fenotipos comunes son

similares en relevancia. Los valores P se obtuvieron mediante la aleatorización de 1000 conjuntos de genes seleccionados al azar. También obtuvimos una tabla con todos los fenotipos compartidos entre los dos conjuntos de genes (esta tabla se puede consultar en https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM7_ESM.xls). Entre los nuevos genes probables de epilepsia predichos por G2PML encontramos que DDX3X, KCNH1, TBL1XR1, DLG4 o PDE2A, recapitulan términos fenotípicos de genes de epilepsia conocidos, comprobamos que comparten términos fenotípicos significativos de epilepsia como Convulsiones (HP: 0001250), Retraso global del desarrollo (HP: 0001263), Microcefalia (HP: 0000252), Morfología cerebral anormal (MP: 0002152), Hiperactividad (MP: 0001399) y términos de enfermedades como Epilepsia (C0014544) y Trastorno autista (C0004352). También encontramos que recapitulan términos CRISPRBrain interesantes como Asociación con Hierro Lábil (Intensidad FeRhoNox) en Neurona Glutamatérgica (CRB: 0000004) y Asociación Positiva con Lípidos Peroxidados (Intensidad Liperfluo) en Neurona Glutamatérgica (CRB: 0000008). Los términos fenotípicos mencionados están asociados con la epilepsia según varios artículos [122, 123, 124, 125, 126, 127, 128, 129, 130]. También hemos estudiado cual es el número de variantes genéticas del estudio Epi25 de secuenciación del exoma completo de casos y controles de cada gen de epilepsia predicho, obtuvimos 665 variantes genéticas en los casos y 446 en los controles, por lo tanto podemos decir que en los genes predichos se encuentran más variantes raras asociadas a casos de epilepsia con una razón de probabilidades, *odds ratio* en inglés (OR = 1,49) (se puede consultar el archivo con los genes y el número de variantes asociadas al fenotipo en el siguiente link https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-022-05122-x/MediaObjects/12859_2022_5122_MOESM8_ESM.xls) [131].

4.5 Conclusiones

Hemos desarrollado PhenoExam, un paquete R y una aplicación web, que realiza análisis de enriquecimiento de fenotipos y enfermedades en el conjunto de genes G, mide las similitudes de fenotipos estadísticamente significativas entre pares de conjuntos de genes G y G' y detecta fenotipos diferenciales estadísticamente significativos entre conjuntos de genes, a través de diferentes bases de datos. PhenoExam sólo requiere los nombres de los genes en los conjuntos de genes como entrada y marcar las bases de datos para realizar la búsqueda de fenotipos. Permite pasar del espacio génico al espacio fenotípico. PhenoExam integra datos de fenotipos de diferentes bases de datos, cada base de datos

se centra en enfermedades concretas y/o organismos específicos. Por lo tanto, la elección de una base de datos para los análisis requiere de un conocimiento básico por parte del usuario sobre las enfermedades que allí se utilizan para comprender adecuadamente el resultado del análisis. PhenoExam puede identificar los fenotipos estadísticamente significativos y diferenciales de un conjunto de genes como demostramos con los conjuntos de genes de PD y EOD, al igual que con los conjuntos de genes de epilepsia y los predichos como probablemente asociados a la epilepsia. Demostramos con simulaciones que es útil para distinguir entre conjuntos de genes o enfermedades con fenotipos muy similares mediante la proyección de genes en sus espacios fenotípicos basados en anotaciones. Con el ejemplo anterior de PD y EOD, vemos claramente que tienen similitudes a nivel de fenotipo, pero también fenotipos diferenciales potencialmente interesantes. El caso condicional estudiado entre genes asociados a epilepsia y genes predichos que pueden estar asociados a la epilepsia muestra que tienen en común términos de fenotipo asociados a los del conjunto de referencia de epilepsia, lo que es útil para la validación de genes de enfermedad predichos computacionalmente. Por lo tanto, PhenoExam descubre eficazmente los vínculos entre los términos fenotípicos a través de bases de datos de anotación mediante la integración de diferentes bases de datos de anotación. Todos estos hallazgos se apoyan con gráficos interactivos (ver tutoriales en el proyecto GitHub) para fomentar la visualización e interpretación de los resultados.

Se puede encontrar el repositorio GitHub con el paquete R y los tutoriales en `https://github.com/alexcis95/PhenoExam` y la web con la aplicación shiny disponible en `https://alejandrocisterna.shinyapps.io/phenoexamweb/`.

CAPÍTULO 5

MODELOS PARA LA PREDICCIÓN DEL RIESGO DE HOSPITALIZACIÓN O MUERTE EN EL MOMENTO DEL DIAGNÓSTICO DE COVID-19

En la introducción hemos hablado de la necesidad de integrar diferentes fuentes de datos clínicos. Hemos planteado que es crucial obtener información de datos médicos y desarrollar herramientas que faciliten la labor de los clínicos y el manejo de los pacientes. Además, hemos mencionado que la bioinformática debía utilizar técnicas de análisis de datos inteligentes, es decir, utilizar tanto la estadística como el ML para extraer información de los datos. Un periodo concreto en el que se tenía una gran cantidad de datos clínicos fue la pandemia causada por el coronavirus. Gracias al proyecto de la Fundación Séneca y a contar con los datos por parte del Servicio Murciano de Salud (SMS) hemos podido trabajar en esta tesis con esos datos en un problema tan importante y actual. Un escenario ideal para plantear técnicas de ML y enfoques novedosos que puedan ayudar ante situaciones similares.

Este capítulo presenta la investigación realizada con los datos de pacientes de la enfermedad producida por el coronavirus, *Coronavirus disease 2019* en inglés COVID-19, obtenidos gracias al Servicio Murciano de Salud (SMS). En este caso se han utilizado datos clínicos para realizar modelos que predigan las posibilidades de sobrevivir (fallece o sobrevive) o de hospitalización (necesidad de ingreso) del paciente en el mismo momento del diagnóstico basado en datos clínicos previos. El desarrollo de herramientas

que proporcionen un triaje precoz de los pacientes con COVID-19 con un uso mínimo de pruebas diagnósticas, basándose en datos fácilmente accesibles, puede ser de vital importancia para reducir las tasas de mortalidad por COVID-19 durante escenarios de alta incidencia, por ejemplo en las olas de la enfermedad. Esta investigación propone un modelo de aprendizaje automático para predecir la mortalidad y el riesgo de hospitalización utilizando características demográficas simples y comorbilidades obtenidas de 86.867 historias clínicas electrónicas de pacientes con COVID-19. Además se ha diseñado un nuevo método para tratar los problemas de desequilibrio de datos.

En esta investigación ha tenido gran relevancia la interpretación de datos clínicos, el uso de la estadística, procesado y del ML. Se han generado modelos de predicción que están disponibles en un repositorio GitHub[1] y también una shiny app[2] para facilitar el uso. Los resultados de la investigación se han publicado en Scientific Reports del grupo Nature[3].

5.1 Introducción

Hemos divido la introducción para facilitar al lector un seguimiento correcto.

5.1.1 Introducción al fenotipo

El virus responsable de la COVID-19, el coronavirus del síndrome respiratorio agudo severo 2, *severe acute respiratory syndrome coronavirus 2* en inglés (SARS-CoV-2), es un betacoronavirus altamente transmisible y patógeno que apareció a finales de 2019 en Wuhan, China [132]. Hasta febrero de 2022, ha tenido un efecto trágico en la salud de la población mundial, con más de 6,9 millones de muertes y 767 millones de casos (datos consultados el 20/06/23) en todo el mundo, convirtiéndose en la crisis sanitaria mundial más importante desde la época de la pandemia de gripe de 1918 [133, 134]. Los síntomas de la COVID-19 son amplios y pueden incluir fiebre, tos, fatiga, problemas gastrointestinales, dolor de garganta, anosmia, hiposmia y síntomas neurológicos [135, 136, 137, 138]. Algunos de estos síntomas pueden persistir tras la recuperación del paciente, en particular la fatiga y la disnea [139]. Según la Organización Mundial de la Salud, *World Health Organization* en inglés (WHO), la tasa de mortalidad mundial por

[1] https://github.com/antoniogt/ipip
[2] https://alejandrocisterna.shinyapps.io/PROVIA/
[3] https://www.nature.com/articles/s41598-022-22547-9

COVID-19 se sitúa en torno al 1,5% desde el inicio de la pandemia hasta febrero de 2022. Aunque, somos conscientes de la existencia de personas más propensas a desarrollar una enfermedad crítica y finalmente fallecer [140]. Además, se ha demostrado que las vacunas reducen las tasas de mortalidad por COVID-19 [141, 142]. Por ejemplo, el último comunicado de los Centros de Control y Prevención de Enfermedades, *Centers from Disease Control and Prevention* en inglés (CDC) de Estados Unidos concluyó que las personas no vacunadas tienen más riesgo de muerte asociada a COVID-19 (la tasa de mortalidad es de alrededor del 1,39%) que las personas totalmente vacunadas (la tasa de mortalidad es de alrededor del 0,78%) con o sin dosis de refuerzo [143]. Debido a la alta contagiosidad y rápida propagación del SRAS-CoV-2, muchos países tienen que gestionar periodos intensos de la enfermedad, que se conocen como olas [144]. En estos periodos, los recursos hospitalarios, la capacidad de las unidades de cuidados intensivos, *Intensive Care Unit* en inglés (ICU), y la saturación del sistema sanitario pueden contribuir al aumento de la letalidad [145]. Por ello, el manejo clínico de los pacientes, las estrategias rápidas de estratificación del riesgo y la optimización del uso de los recursos son importantes para reducir la tasa de letalidad [146, 147].

5.1.2 Trabajos previos

Es importante mencionar que los datos clínicos de la pandemia causada por el coronavirus tienen un horizonte temporal limitado. Se empezó a tener datos de la enfermedad a finales de 2019 en China. No hay trabajos relacionados con este problema concreto anteriores a 2020. Por tanto, no hay una tecnología nueva o algoritmos específicos para tratar de abordar el problema.

Como ya hemos comentado en el capítulo 1 de introducción, las historias clínicas electrónicas (EHR) son uno de los principales recursos para mejorar la forma de abordar la gestión de los pacientes y avanzar hacia un triaje más eficiente de los pacientes, en este caso concreto pacientes con COVID-19. Así, los datos demográficos y de salud de los pacientes disponibles a través de los sistemas sanitarios se han utilizado para el pronóstico y la evolución de los pacientes con COVID-19 mediante el uso de sistemas semiautomatizados de AI.

Para realizar esta investigación se han estudiado otras publicaciones similares. Por ejemplo, se ha desarrollado un modelo XGBoost basado en aprendizaje automático para predecir las tasas de mortalidad de los pacientes con más de 10 días de antelación con una exactitud de alrededor del 90%, utilizando tres biomarcadores como indicadores

principales para predecir el pronóstico de COVID-19, la lactato deshidrogenasa (LDH), la proteína C reactiva de alta sensibilidad (hs-CRP) y el recuento de linfocitos que se puede encontrar en la siguiente publicación [148]. Por otro lado, se obtuvo un ROC-AUC alto (0,96) con los datos clínicos de los pacientes utilizando cuatro métodos de aprendizaje automático: regresión logística, máquina de vectores soporte, árbol de decisión con gradient boosting y red neuronal [149]. Utilizando LASSO y una ecuación predictiva con regresión logística binaria basada en comorbilidades preexistentes y datos demográficos se concluyó que estas variables demostraban una buena capacidad para discriminar los resultados graves de los no graves utilizando únicamente esta información histórica con un AUC de 0,76 [150]. En otro estudio se desarrollaron modelos basados en aprendizaje automático con diferentes técnicas, LASSO, univariante novedosa y por pares, pero se concluyó que ningún modelo era capaz de superar a un modelo basado únicamente en la edad, en el que ésta tenía un AUC de 0,85 y una exactitud equilibrada de 0,7 [151]. Otro modelo fue capaz de predecir el riesgo de ingreso en hospital/ICU y de muerte ya en el momento del diagnóstico con un ROC-AUC de 0,902 centrándose únicamente en un número limitado de comorbilidades y variables demográficas, como la edad, el sexo y el indice de masa corporal, *Body Mass Index* en inglés (BMI) [152]. En todos los casos anteriores, el conjunto de datos sanitarios está desequilibrado, ya que los individuos con los episodios más graves de la enfermedad son minoría, por lo que un enfoque de aprendizaje automático supervisado dirigido a modelarlos sufrirá desequilibrios. Además, los predictores considerados para algunos de estos estudios son difíciles de obtener. Por ejemplo, la LDH, la albúmina (ALB), el nitrógeno ureico en sangre (BUN), la hs-CRP y los linfocitos requieren el uso de análisis de sangre o de urea; o bien la realización de mediciones con dispositivos físicos específicos, como es el caso del BMI, la temperatura y la saturación de oxígeno. Por lo tanto, estos modelos distan mucho de ser utilizables de forma realista para el triaje precoz de pacientes en momentos de sobresaturación de urgencias.

5.1.3 Plantemiento del trabajo

A diferencia de otros estudios, el nuestro presenta una técnica especialmente diseñada para abordar problemas de desequilibrio aplicada a la racionalización y mejora del triaje de pacientes con COVID-19 en función de su edad, sexo y comorbilidades, basada en datos fácilmente disponibles obtenidos del SMS. Esto nos permite realizar un estudio regional en el que los datos están organizados en cinco fuentes diferentes, incluyendo información sobre historia clínica, servicios de hospitalización, síntomas, constantes

vitales, tratamientos realizados en más de 100.000 pacientes con COVID-19 con fechas de diagnóstico que van desde el 4 de enero de 2020 al 4 de febrero de 2021. Cabe destacar que este estudio no puede centrarse en los efectos de la vacunación. Nótese que sólo el 1,4% de la población total de España estaba vacunada en la fecha en la que se inscribió el último paciente en nuestro estudio[4]. Esto nos deja con sólo aproximadamente 402 sujetos vacunados. La técnica presentada para tratar el desequilibrio consiste en dividir el problema original en p subproblemas, donde cada uno tendrá asociado un conjunto de datos perfectamente equilibrado, formado por muestras del conjunto original. Utilizando este razonamiento, es posible construir un modelo de regresión logística ensemble, una combinación de modelos, que permite obtener ROC-AUC de 0,94 para predecir el estado final del paciente (sobrevive o fallece) con resultados similares o superiores a los modelos complejos que combinan varios métodos de aprendizaje automático, y a los que combinan datos mucho más complejos basados en técnicas de laboratorio o asistenciales.

Como hemos mencionado anteriormente, los datos con los que hemos trabajado son un conjunto de COVID-19 único en el mundo por tamaño y localidad (en la región de Murcia). Estos datos se obtuvieron en el contexto de una convocatoria extraordinaria de proyectos que se desarrolló por la Fundación Séneca en pleno confinamiento y en la que he participado como miembro del equipo investigador a la par que redactando parte de la propuesta de investigación.

5.2 Métodos

5.2.1 Diseño del estudio

Nuestra cohorte se compuso por pacientes diagnosticados de COVID-19 de forma consecutiva entre el 4 de enero de 2020 y el 4 de febrero de 2021. Un caso confirmado de COVID-19 se define como un resultado positivo de la prueba de antígeno o del ensayo de reacción en cadena de la polimerasa con transcriptasa inversa en tiempo real, *reverse-transcriptase polymerase-chain-reaction* en inglés (RT-PCR), para muestras de hisopos nasales y faríngeos. Los pacientes que se excluyeron del análisis son pacientes con al menos una característica no disponible (6192 pacientes), o con la enfermedad COVID-19 todavía activa en la fecha de cierre de registros de pacientes (9513 pacientes), o pacientes con valores erróneos que no pudieron ser posibles (1 paciente). Los datos de nuestro análisis y modelos incluyeron 86.867 pacientes COVID-19 confirmados, se

[4] `https://ourworldindata.org/covid-vaccinations?country=ESP`

puede encontrar un diagrama en la figura 5.1. Las características incluidas en nuestro estudio procedían de la base de datos denominada estratificación de pacientes, que incluye información de cada paciente sobre edad, sexo y comorbilidades como: diabetes mellitus, demencia, obesidad, insuficiencia cardíaca, enfermedad pulmonar obstructiva crónica, *chronic obstructive pulmonary disease* en inglés (COPD), asma, hipertensión arterial, depresión, miocardiopatía isquémica, accidente cerebrovascular, insuficiencia renal, cirrosis, osteoporosis, artrosis, artritis, síndrome de inmunodeficiencia adquirida (VIH) y dolor crónico.

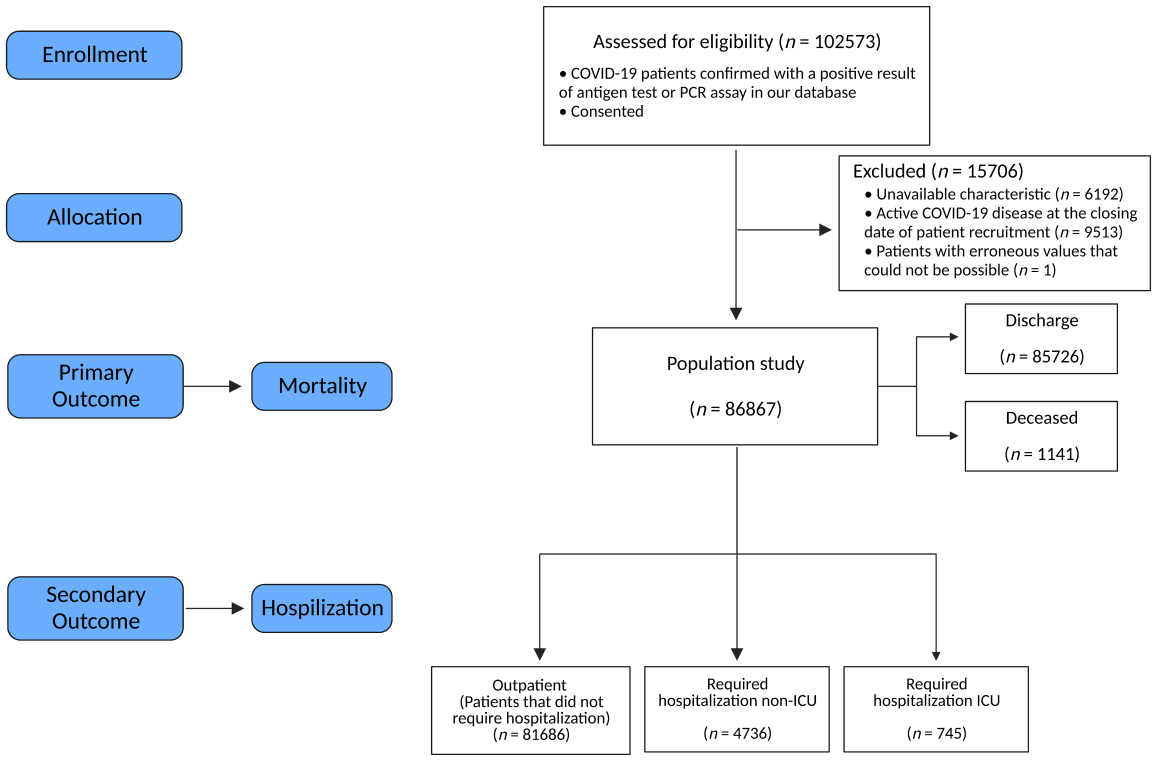


Figura 5.1: Diagrama CONSORT. Diagrama de flujo de los sujetos y cómo se analizan en el estudio.

5.2.2 Acceso a los datos personales

La información de interés para el estudio se ha obtenido de los archivos de historias clínicas del SMS, sin el consentimiento de los titulares de los datos. La necesidad de consentimiento informado fue dispensada por el Comité de Bioética de la Universidad de Murcia. Ello se hizo de acuerdo con los siguientes criterios

- El artículo 157 del Reglamento General de Protección de Datos (RGPD) reconoce el beneficio que el acceso a los registros puede aportar a la investigación sobre enfermedades, de forma que los resultados de estos estudios podrían ser más sólidos, basados en una población más amplia.

- El artículo 89.1 del RGPD confirma este principio siempre que se adopten las medidas adecuadas, en particular el respeto del principio de minimización de los datos personales. Estas medidas también pueden incluir el uso de datos seudonimizados o anonimizados.

- Por su parte, el artículo 14.5 b) del RGPD permite al responsable del tratamiento no informar a los interesados cuando *dicha información resulte imposible o suponga un esfuerzo desproporcionado en particular para el tratamiento con fines de interés público, fines de investigación científica o fines históricos o estadísticos sujetos a las condiciones y garantías indicadas en el artículo 89, apartado 1, o en la medida en que la obligación mencionada en el apartado 1 de este artículo pueda imposibilitar o dificultar gravemente la consecución de los objetivos de dicho tratamiento*.

- La disposición adicional decimoséptima de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales permite la utilización de datos de salud con fines de investigación científica siempre que estos datos hayan sido sometidos a un tratamiento previo de seudonimización o anonimización.

De acuerdo con los preceptos expuestos, el consentimiento no es necesario por los siguientes motivos:

1. El acceso a las historias clínicas con fines de investigación científica es lícito, siempre que se adopten determinadas garantías, entre ellas el respeto al principio de minimización de datos y que la información obtenida sea adecuadamente seudonimizada o anonimizada.

2. La obtención del consentimiento no sería un requisito legal, ya que no sólo supondría un esfuerzo desproporcionado, sino que podría entorpecer el desarrollo del estudio debido a la gran cantidad de información analizada.

3. El SMS, de acuerdo con lo establecido en la citada normativa, ha sometido esta información clínica a un proceso de anonimización, excluyendo cualquier dato identificativo de los pacientes, imposibilitando su identificación.

5.2.3 Descripción y preprocesamiento de datos

Los datos epidemiológicos y clínicos de COVID-19 se recogieron de los registros médicos electrónicos del SMS y su uso para este trabajo y dentro de todos los experimentos en él incluidos fue aprobado por el Comité de Ética de la Universidad de Murcia. Por lo tanto, todos los experimentos se realizaron de acuerdo con las directrices y normativas pertinentes. La base de datos que estratifica a los pacientes diagnosticados de COVID-19 (un total de 102.573 pacientes) incluye edad, sexo, hospital y unidad de atención primaria asignada al paciente, información sobre el ingreso y estado final (es decir, el paciente está curado o ha fallecido), información sobre comorbilidades, número de patologías crónicas y número de sistemas afectados, así como estrato de riesgo. Otros datos que limitaban el número de pacientes, como la medicación dispensada por el hospital (9.165 pacientes), la duración de la estancia en cada departamento (8.356 pacientes), la información sobre las constantes vitales (7.524 pacientes) o la información sobre los diversos síntomas que presentaban 89.769 pacientes (dolor de cabeza, fiebre, mareos, vómitos, entre otros) no se utilizaron para simplificar los datos necesarios para el triaje de pacientes en situaciones de colapso del sistema hospitalario durante un brote pandémico.

5.2.4 Modelos de predicción planteados

Utilizamos el aprendizaje automático para desarrollar modelos sobre las siguientes cuestiones de interés

- ¿Cuál será el estado final del sujeto? Es decir, el paciente sobrevivirá a la infección por COVID-19 o fallecerá.
- ¿Será necesario ingresar al paciente en el hospital?

Nuestro interés en ellas es ayudar a la toma de decisiones clínicas en situaciones de congestión del sistema. Estas situación han sido frecuentes durante diversos periodos y se denominan olas. Además, nuestro enfoque es responder a las mismas cuando el paciente acaba de ser diagnosticado, por lo que la única información que podemos utilizar

es la obtenida en una revisión médica o la información previa disponible en la EHR del paciente. En la siguiente sección introduciremos algunos conceptos básicos del ML.

5.2.4.1 Breve introducción al ML empleado

El aprendizaje automático (ML) es una disciplina de la inteligencia artificial que tiene como objetivo la construcción de sistemas que sean capaces de aprender a partir de los datos. Esto implica diseñar y utilizar algoritmos que faciliten a las máquinas la adquisición de conocimientos y la toma de decisiones o predicciones. Además, depende de para que problemas y como utilicemos los datos podemos hablar de:

- **Aprendizaje supervisado**: Este tipo de aprendizaje ocurre cuando los algoritmos aprenden a partir de datos previamente etiquetados. En el proceso de entrenamiento, al algoritmo se le proporciona un conjunto de entradas con las correspondientes salidas (etiquetas). El propósito es aprender una función que sea capaz de predecir la salida correcta para una entrada nueva. Cuando el modelo ha sido entrenado, puede ser utilizado para predecir las salidas de nuevas instancias. Los problemas de clasificación y regresión son ejemplos típicos del aprendizaje supervisado. Este es el tipo de ML aplicado a nuestro problema ya que es un estudio retrospectivo y conocemos el desenlace de cada paciente.
- **Aprendizaje no supervisado**: En el aprendizaje no supervisado, los algoritmos aprenden a partir de datos que no han sido etiquetados. A diferencia de predecir una salida, el objetivo aquí es descubrir la estructura inherente de los datos. Algunas técnicas frecuentes de aprendizaje no supervisado son el agrupamiento (clustering), la reducción de la dimensionalidad y la detección de anomalías.

También es necesario comprender el concepto de clasificador para entender lo que estamos proponiendo. Hay dos tipos de clasificadores:

- **Clasificador binario**: Es un tipo de clasificador que diferencia entre dos clases. Por ejemplo, un clasificador binario podría ser utilizado para predecir si un paciente fallece o sobrevive.
- **Clasificador multiclase**: Es un tipo de clasificador capaz de distinguir entre más de dos clases. Por ejemplo, un clasificador multiclase podría ser utilizado para

predecir el tipo de enfermedad que presenta un paciente (Parkinson, cancer de pulmón, diabetes, etc.) a partir de fenotipos o síntomas recogidos.

En definitiva, en este caso el tipo de aprendizaje es supervisado y el clasificador utilizado es binario. Binario en base a resultados de varias regresiones (sección 5.2.4.5).

5.2.4.2 El problema de ML es desbalanceado

En este estudio, lo que tenemos a nuestra disposición es una tabla que contiene información sobre la edad, el sexo, las comorbilidades, el estado de hospitalización y el resultado final de los pacientes. Los conjuntos de datos disponibles para ambas cuestiones están muy desequilibrados, ya que hay muchos menos individuos fallecidos (1141) que pacientes dados de alta (85.726), y hay más pacientes ambulatorios (81.386) que hospitalizados (no ICU 4736 y ICU 745).

5.2.4.3 El algortimo IPIP

Para resolver estos problemas, proponemos un nuevo método de aprendizaje automático basado en conjuntos y que tiene en cuenta los desequilibrios o desbalanceo. Se denomina IPIP, *Identical Partitions for Imbalance Problems* en inglés. En primer lugar, extraemos el 20% de los datos de la clase minoritaria y el mismo número de muestras de la clase mayoritaria para crear un conjunto de prueba. El resto va a un conjunto de entrenamiento. A continuación, dividimos el conjunto de datos de entrenamiento en p conjuntos de datos equilibrados. Para ello es necesario configurar un hiperparámetro del algoritmo, denominado variabilidad de la proporción. Por defecto, IPIP crea p conjuntos de datos perfectamente equilibrados (50-50%) para los problemas de clasificación binaria. Este hiperparámetro permite especificar un intervalo aleatorio de variabilidad de esa proporción. Por ejemplo, un valor de 0,05 incrementa aleatoriamente, desde el 50% hasta un 5% más, la proporción de la clase mayoritaria para cada una de las p remuestreos. En este problema concreto, el hiperparámetro de variabilidad de la proporción se fijó en 0,05 para predecir la condición final. Y para el problema de hospitalización utilizamos conjuntos de datos perfectamente equilibrados. Estos valores finales se obtuvieron probando empíricamente el rango de valores (figura 5.2) cómo se comportan los modelos IPIP en un rango de valores de variabilidad de proporción. Una vez creadas las muestras p, creamos para cada una de ellas un modelo ML básico. Todos los modelos entran en un conjunto cuya agregación de respuesta es una mayoría simple. IPIP selecciona p en función de n (número de muestras). Valores más altos de n conducen a valores más bajos de p. Para

este conjunto de datos en particular, el 75% de las muestras de clases minoritarias de los datos del tren conduce a p = 7 subconjuntos. En cada iteración p divide aleatoriamente los conjuntos de datos de entrenamiento y prueba y genera un modelo con los datos de entrenamiento y otro con los datos de prueba. Si el nuevo modelo mejora la calidad global del conjunto, se añade. Si no, se muestrea aleatoriamente y se vuelve a probar hasta un número máximo de intentos. Utilizamos el conjunto de prueba para evaluar el conjunto candidato a la mejora. Al realizar la inferencia, el clasificador predice una observación como miembro de la clase mayoritaria cuando al menos el 75% de los modelos clasificados como negativos (clase mayoritaria) serán clasificados como negativos.

La inferencia del ensemble final en modo producción se genera evaluando cada uno de los modelos que componen el ensemble final, si el 50% de ellos clasifican una muestra como negativa (clase mayoritaria), el modelo final la clasifica como negativa. Los modelos predictivos encontrados en la literatura se basan en una gran variedad de modelos: conjuntos de árboles de decisión [148], clasificadores de máquinas de vectores soporte, redes neuronales [149] o simplemente modelos de regresión logística lasso [150]. El modelo IPIP es un ensemble compuesto por un conjunto de modelos básicos de regresión logística (LR) creados a partir de diferentes submuestras de los datos originales. El clasificador global predice una observación como miembro de la clase mayoritaria cuando al menos el 75% de los modelos clasificados como negativos (clase mayoritaria) se clasifiquen como negativos. Por lo tanto, los modelos IPIP son extremadamente eficientes tanto en términos de requisitos informáticos como de almacenamiento.

5.2.4.4 Desbalanceo y métricas de evaluación en ML

Antes de estudiar el efecto del desbalanceo hemos mencionado que IPIP está desarrollado para intentar mitigar este problema. El desbalanceo en los datos en ML se refiere a la diferencia significativa en el número de observaciones en cada clase de un problema de clasificación. Un ejemplo, como en nuestro caso, es cuando estás tratando de predecir un evento muy poco frecuente (la muerte) en un conjunto donde la mayoría de los pacientes de COVID-19 sobreviven (digamos, el 98% de las veces sobreviven), por lo que tu conjunto de datos está "desbalanceado". Es decir, tienes una clase mayoritaria (sobrevive) que se registra el 98% de las veces y una minoritaria (fallecen) que se registra el 2% de las veces. Esto puede causar problemas, ya que la mayoría de los algoritmos de ML tienden a estar sesgados hacia la clase mayoritaria, ignorando a la minoritaria, que en la mayoría de las veces es la de interés. Además, Cuando se tienen conjuntos de datos

Tabla 5.1: Matriz de confusión predicciones y realidad

		Predicción	
		+	-
Real	+	TP	FN
	-	FP	TN

desbalanceados, las métricas tradicionales, como la exactitud, pueden no ser las más adecuadas para evaluar el rendimiento de un modelo. Algunas métricas que se suelen utilizar en casos de desbalanceo incluyen la sensibilidad (sensitivity), la especificidad (specificity), la exactitud equilibrada o balanceada (balanced accuracy), el área bajo la curva ROC (AUC-ROC) o el coeficiente Kappa de Cohen (Cohen's Kappa).

Para entender las métricas propuestas tenemos que conocer y comprender los resultados de una matriz de confusión en la que comparamos las predicciones de un modelo con la realidad (tabla 5.1). Habitualmente, esto es lo que se calcula cuando se evalúa un experimento de ML para un modelo de clasificación. En esta matiz de confusión tenemos los verdaderos positivos, (TP) *True Positive* en inglés, los verdaderos negativos, (TN) *True Negative* en inglés, los falsos positivos, (FP) *False Positive* en inglés, y los falsos negativos, (FN) *False Negative* en inglés. Los TP se calculan con las veces que el modelo indica que una condición existe cuando en la realidad existe o está presente, los TN son las veces que el modelo indica que una condición no existe cuando en la realidad la condición no existe, los FP son las veces que el modelo indica que una condición existe cuando en realidad no existe y los FN es cuando el modelo indica que una condición no existe cuando realmente si existe.

Con estos conceptos hemos utilizado las siguientes métricas para evaluar nuestro problema:

- **Sensitivity**: La sensibilidad se define en la ecuación 5.1. Representa la proporción de casos positivos que fueron correctamente identificadas por el modelo.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.1}$$

- **Specificity**: La especificidad se define en la ecuación 5.2. Representa la proporción de casos negativos que fueron correctamente identificadas por el modelo.

$$Sensitivity = \frac{TN}{TN + FP} \tag{5.2}$$

- **Positive predictive value** (PPV): El valor predictivo positivo es la probabilidad de que, si ha obtenido un resultado positivo en la predicción, realmente tenga la condición estudiada. Se define en la ecuación 5.3.

$$PPV = \frac{TP}{TP + FP} \tag{5.3}$$

- **Negative predictive value** (NPV): El valor predictivo negativo es la probabilidad de que, si ha obtenido un resultado negativo en la predicción, realmente no tenga la condición estudiada. Se define en la ecuación 5.4.

$$NPV = \frac{TN}{TN + FN} \tag{5.4}$$

- **Balanced accuracy**: La exactitud balanceada se define en la ecuación 5.5. La exactitud tradicional simplemente calcula la proporción de predicciones correctas, sin tener en cuenta si las predicciones eran de la clase mayoritaria o de la clase minoritaria. Esto puede ser problemático en conjuntos de datos desbalanceados, ya que un modelo podría obtener una exactitud muy alta simplemente prediciendo siempre la clase mayoritaria. La exactitud balanceada aborda este problema tomando en cuenta tanto la sensibilidad (la proporción de verdaderos positivos sobre todos los casos positivos reales) como la especificidad (la proporción de verdaderos negativos sobre todos los casos negativos reales). De esta manera, se garantiza que el modelo se comporta correctamente en ambas clases, no solo en la clase mayoritaria. En resumen, la exactitud balanceada es más justa porque evita que los modelos que simplemente predicen siempre la clase mayoritaria obtengan una alta exactitud.

$$Balanced \quad accuracy = \frac{Sensitivity + Specificity}{2} \tag{5.5}$$

- **Cohen's Kappa**: El coeficiente Kappa de Cohen mide la posibilidad de acierto real y la compara con la posibilidad de acierto al azar. Se expresa con la ecuación 5.6:

$$k = \frac{2 * (TP * TN - FP * FN)}{(TP + FP) * (FP + TN) * (TP + FN) * (FN + TN)} \tag{5.6}$$

En conjuntos de datos desbalanceados si un modelo simplemente está prediciendo siempre la clase mayoritaria, su exactitud observada puede ser alta, pero su exactitud esperada también sería alta, por lo que el coeficiente de Kappa podría ser bajo. El coeficiente de Kappa varía de -1 a 1. Un valor de 1 implica un acuerdo perfecto entre el modelo y la realidad. Un valor de 0 implica que el nivel de acuerdo es exactamente lo que se esperaría por azar. Los valores negativos sugieren que hay menos acuerdo del que se esperaría por azar. En la práctica, un coeficiente de Kappa alto (cercano a 1) sugiere que el modelo está funcionando bien, mientras que un coeficiente bajo (0, o peor aún, negativo) sugiere que el modelo no está funcionando mejor de lo que se esperaría por azar.

- **AUC-ROC**: La curva ROC es una representación gráfica que ilustra el rendimiento de un sistema de clasificación binario a medida que varía el umbral de discriminación. El AUC-ROC (Área Bajo la Curva - Característica Operativa del Receptor) es una métrica de rendimiento que calcula el área total debajo de esta curva ROC. Un área de 1 representa un modelo que hizo todas las predicciones correctamente, mientras que un área de 0.5 representa un modelo que no es mejor que una predicción aleatoria. Esta métrica es útil cuando se trata con conjuntos de datos desbalanceados, ya que mide cómo se clasifican las instancias de ambas clases en todas las configuraciones de umbral posibles.

5.2.4.5 Experimentos de ML y selección del modelo

Creamos dos modelos IPIP para abordar cada pregunta de modelado. Uno con un algoritmo de referencia, la regresión logística, y el otro con bosques aleatorios [153] como modelos básicos basados en el paquete Caret [154]. Todos los modelos se evaluaron mediante validación cruzada quíntuple. Se probaron diferentes valores del número de árboles de decisión para los modelos de bosque aleatorio, y decidimos que cada modelo de bosque aleatorio utilizara 200 árboles de decisión, donde la impureza es el modo de importancia de la variable, que es el índice de Gini para la clasificación. Se creó una cuadrícula de ajuste para elegir el mejor tamaño mínimo de los nodos de los árboles (1, 11 o 21) y el número de variables que posiblemente se dividirían en cada nodo (1, 4, 7, 10, 13, 16 o 19). Para decidir si un modelo básico mejora o no el conjunto, nos basamos en

la métrica Kappa de Cohen [155]. Es decir, si al añadir un modelo básico al conjunto de modelos básicos entrenados para un subconjunto específico perfectamente equilibrado mejoraba el Kappa en la evaluación sobre el conjunto de prueba disponible del nuevo conjunto de modelos básicos con respecto a los valores Kappa anteriores del mismo conjunto de evaluación, añadíamos ese modelo básico a ese conjunto de modelos básicos; en caso contrario, se descartaba. También obtuvimos las siguientes métricas: exactitud equilibrada, valor predictivo negativo (NPV), valor predictivo positivo (PPV), sensibilidad y especificidad. Además, calculamos el área bajo la curva de las características operativas del receptor (ROC-AUC) para el modelo conjunto final.

5.2.4.6 Efecto de la proporción de desbalanceo en los resultados de IPIP

En la figura 5.2 podemos ver el efecto de diferentes valores de variabilidad de la proporción en la exactitud balanceada y Kappa de Cohen. En la parte a) observamos la evolución de la exactitud equilibrada con respecto a la proporción. El eje X indica los diferentes valores de variabilidad de la proporción y el eje Y indica la exactitud obtenida. Las curvas roja y azul indican los resultados obtenidos por el modelo IPIP con regresión logística y bosque aleatorio, respectivamente, para el problema de condición final. Las curvas verde y morada indican los resultados obtenidos por el modelo IPIP con regresión logística y bosque aleatorio para el problema de hospitalización respectivamente. En la parte b) observamos la evolución del Kappa de Cohen con respecto a la variabilidad de la proporción. Mismo gráfico pero el eje y indica el Kappa de Cohen obtenido. Obsérvese que las curvas de ambos gráficos, que corresponden al problema de la hospitalización terminan en 0,25 de variabilidad porque, debido a la forma en que IPIP genera las remuestreos, no es posible ni necesario generar p remuestreos con las características que requiere IPIP y un desequilibrio superior al 25

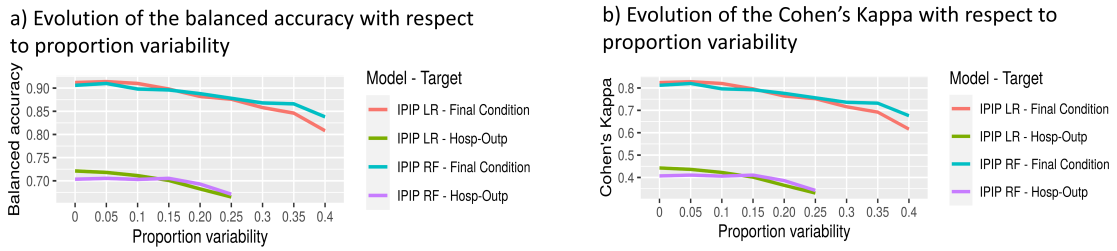


Figura 5.2: Efecto de los valores de variabilidad de la proporción en la exactitud y Kappa de Cohen.

5.2.5 Análisis estadístico

Los datos continuos se presentan como mediana con intervalo intercuartílico, *Interquartile Range* en inglés (IQR), y los datos categóricos se expresan como porcentajes (%). También se utilizaron OR junto con el intervalo de confianza, *Confidence Interval* (CI) del 95 %. Los OR se ajustaron por sexo, edad y otras comorbilidades. Las diferencias entre grupos se comprobaron mediante la U de Mann-Whitney para las variables numéricas, y se utilizó la prueba de la X^2 o la prueba exacta de Fisher para comprobar la significación de los datos categóricos. El análisis estadístico se realizó utilizando R (versión 3.6.3) con un umbral de significación de los valores P de 0,05.

5.3 Resultados

5.3.1 Descripción y diferencias de los distintos tipos de pacientes con COVID-19 en nuestro conjunto de datos

El análisis exploratorio de los datos de 86.867 pacientes permitió estratificar la base de datos obtenida por edad, sexo y comorbilidades específicas (Tabla 5.2), siguiendo el diagrama de flujo de la cohorte que se muestra en la figura 5.1. Entre los casos estudiados, el 93,7 % eran pacientes ambulatorios (N = 81.386), el 5,4 % eran pacientes hospitalizados fuera de ICU (N = 4736) y menos del 0,85 % eran pacientes ingresados en ICU (N = 745). Los síntomas más frecuentes entre los pacientes fueron tos (49,9 %), seguida de cefalea (38,3 %) y mialgia (36 %) (Se pueden encontrar datos de los sintomas en la tabla 5.3). Utilizando los datos de la tabla 5.2 podemos identificar los diferentes prototipos de pacientes con COVID-19.

A partir del análisis exploratorio se identificaron tres tipos de pacientes, el tipo de paciente COVID-19 más frecuente que no requirió hospitalización fue una mujer de 38 años (IQR: 22-52) (el 53,06 % de nuestros pacientes ambulatorios son mujeres), con 2 patologías crónicas y 2 sistemas afectados, cuyas patologías o comorbilidades más frecuentes fueron la hipertensión arterial (15,00 %), seguida de la obesidad (9,24 %), la depresión (9,02 %) y el asma (8,64 %). Por el contrario, el paciente COVID-19 prototípico hospitalizado fuera de ICU era un varón de 62 años (IQR: 47-79) (51,73 % de estos pacientes eran hombres), con 5 patologías crónicas, con 4 sistemas afectados, y con patologías o comorbilidades más frecuentes como hipertensión arterial (46,79 %), diabetes mellitus (25,27 %), obesidad (21,28 %), artrosis (18,03 %) y depresión (16,89 %). Finalmente, el

Tabla 5.2: Características demográficas, comorbilidades y resultado final de diferentes tipos de pacientes de COVID-19.

Characteristics	**Outpatient**	**Hospitalized (non-ICU)**	**ICU**
Number of individuals (N)	81,386	4736	745
Age median (IQR)	38.00 (22.00, 52.00)	62.00 (47.00, 79.00)	62 (52.00, 71.00)
Gender			
Male (%)	38,200 (46.94%)	2450 (51.73%)	524 (70.34%)
Female (%)	43,186 (53.06%)	2286 (48.27%)	221 (29.66%)
Comorbidities			
Number of Chronic diseases median (IQR)	2.00 (1.00, 4.00)	5.00 (2.00, 8.25)	5 (2.00, 8.00)
Number of systems affected median (IQR)	2.00 (1.00, 3.00)	4.00 (2.00, 5.00)	3 (2.00, 5.00)
Asthma (%)	7032 (8.64%)	398 (8.40%)	70 (9.40%)
Obesity (%)	7516 (9.24%)	1008 (21.28%)	220 (29.53%)
Diabetes mellitus (%)	5212 (6.40%)	1197 (25.27%)	214 (28.72%)
Heart failure (%)	646 (0.79%)	343 (7.24%)	33 (4.43%)
COPD (%)	1059 (1.30%)	350 (7.39%)	46 (6.17%)
Arterial hypertension (%)	12,210 (15.00%)	2216 (46.79%)	356 (47.79%)
Depression (%)	7345 (9.02%)	800 (16.89%)	103 (13.83%)
HIV (%)	121 (0.15%)	10 (0.21%)	3 (0.40%)
Ischemic cardiomyopathy (%)	1349 (1.66%)	420 (8.87%)	75 (10.07%)
Stroke (%)	961 (1.18%)	353 (7.45%)	37 (4.97%)
Renal insufficiency (%)	1267 (1.56%)	515 (10.87%)	66 (8.86%)
Cirrhosis (%)	1672 (2.05%)	256 (5.41%)	57 (7.65%)
Osteoporosis (%)	2453 (3.01%)	431 (9.10%)	51 (6.85%)
Osteoarthritis (%)	5212 (6.40%)	854 (18.03%)	121 (16.24%)
Arthritis (%)	1111 (1.37%)	132 (2.79%)	21 (2.82%)
Dementia (%)	859 (1.06%)	296 (6.25%)	9 (1.21%)
Chronic pain (%)	39 (0.05%)	9 (0.19%)	4 (0.54%)
Outcome			
Discharge (%)	81,132 (99.69%)	4082 (86.19%)	512 (68.72%)
Deceased (%)	254 (0.31%)	654 (13.81%)	233 (31.28%)

perfil del paciente ingresado en la ICU fue un varón de 62 años (IQR: 52-71) (70,34% de los pacientes ingresados en ICU eran varones), con 5 patologías crónicas, 3 sistemas afectados, y cuyas patologías o comorbilidades más frecuentes fueron la hipertensión arterial (47,79%), seguida de la obesidad 29,53%, la diabetes mellitus 28,72%, y la artrosis (18,03%). COVID-19 Los pacientes ingresados en la ICU tenían más del doble de posibilidades de morir que los hospitalizados no ingresados en la ICU (31,28% frente a 13,81%), muy lejos de la que presentaban los pacientes ambulatorios (0,31%) con apenas probabilidades de fallecer.

Para seguir estudiando las diferencias entre los pacientes COVID-19 dados de alta (supervivientes) y los que fallecieron (no supervivientes), los datos de la tabla 5.2 se reorganizaron en la tabla 5.4. El prototipo de paciente superviviente fue una mujer

Tabla 5.3: Síntomas y su frecuencia en pacientes con COVID-19.

Symptoms (Number of patients and percentages%)	
Number of individuals	89768
Cough (%)	44875 (49.99%)
Headache (%)	34419 (38.34%)
Myalgia (%)	32347 (36.03%)
Hyposmia (%)	26934 (30.00%)
Rhinorrhea (%)	26634 (29.67%)
Hypogeusia (%)	24412 (27.19%)
Nasal congestión (%)	23677 (26.38%)
Sore throat (%)	22186 (24.71%)
Expectoration (%)	11372 (12.67%)
Shivering (%)	10674 (11.89%)
Fever (%)	9810 (10.93%)
Chest pain (%)	7121 (7.93%)
Abdominal pain (%)	6680 (7.44%)
Dizziness (%)	6115 (6.81%)
Vomits (%)	3540 (3.94%)
Arterial hypertension (%)	2442 (2.72%)
Skin problems (%)	1556 (1.73%)
Eye problems (%)	1358 (1.51%)

de 39 años (IQR: 23-53) (52,72% son mujeres), con 2 patologías crónicas, 2 sistemas afectados, y con patologías o comorbilidades más frecuentes similares a las descritas previamente para los pacientes no hospitalizados. Por el contrario, el perfil del paciente no superviviente era claramente diferente y estaba representado por un hombre de 83 años (IQR: 75-88) (56,00% eran hombres), con 8 patologías crónicas, 5 sistemas afectados, y cuya patología o comorbilidad prevalente era la hipertensión arterial (75,64%), lejos de las posteriores como la diabetes mellitus (42,33%), la obesidad (29,36%), la artrosis (27,93%) y la depresión (23,31%). Además, tres variables fueron muy relevantes para el estado final del paciente, se muestra su distribución en la figura 5.3. Así, a mayor edad del paciente $t(1237) = 116{,}9$, $p < 2{,}2 \times 10^{-16}$, mayor número de patologías crónicas $t(1151) = 42{,}15$, $p < 2{,}2 \times 10^{-16}$ y mayor número de sistemas afectados $t(1163) = 47{,}2$, $p < 2{,}2 \times 10^{-16}$, mayor probabilidad de fallecer. Se observó una distribución similar asociada a más riesgo según la hospitalización de los pacientes para las variables anteriores cuando se dividió la población en los tres grupos iniciales (pacientes ambulatorios, hospitalizados no ICU y pacientes ICU) (figura 5.4). Se puede observar que los pacientes ambulatorios son más jóvenes y que tienen menos comorbilidades. Aunque se observa una cierta tendencia

Tabla 5.4: Características demográficas y comorbilidades en pacientes COVID-19 supervivientes y fallecidos.

Characteristics	**Discharge (survival)**	**Deceased (non-survival)**	**OR (95% CIs)**
Number of individuals (N)	85,726	1141	–
Age median (IQR)	39.00 (23.00, 53.00)	83 (75.00, 88.00)	–
Gender			
Male (%)	40,535 (47.28%)	639 (56.00%)	2.41 (2.11, 2.75)
Female (%)	45,191 (52.72%)	502 (44.00%)	–
Comorbidities			
Number of Chronic diseases median (IQR)	2.00 (1.00, 4.00)	8 (5.00, 12.00)	–
Number of systems affected median (IQR)	2.00 (1.00, 3.00)	5 (4.00, 7.00)	–
Asthma (%)	7403 (8.64%)	97 (8.50%)	1.13 (0.90, 1.42)
Obesity (%)	8409 (9.81%)	335 (29.36%)	1.57 (1.36, 1.81)
Diabetes mellitus (%)	6140 (7.16%)	483 (42.33%)	1.55 (1.36, 1.77)
Heart failure (%)	823 (0.96%)	199 (17.44%)	1.85 (1.53, 2.23)
COPD (%)	1290 (1.50%)	165 (14.46%)	1.49 (1.23, 1.81)
Arterial hypertension (%)	13,919 (16.24%)	863 (75.64%)	1.31 (1.12, 1.53)
Depression (%)	7982 (9.31%)	266 (23.31%)	1.24 (1.06, 1.44)
HIV (%)	134 (0.16%)	0 (0.00%)	–
Ischemic cardiomyopathy (%)	1631 (1.90%)	213 (18.67%)	1.58 (1.33, 1.88)
Stroke (%)	1145 (1.34%)	206 (18.05%)	1.84 (1.54, 2.20)
Renal insufficiency (%)	1591 (1.86%)	257 (22.52%)	1.90 (1.61, 2.24)
Cirrhosis (%)	1923 (2.24%)	62 (5.43%)	1.41 (1.07, 1.86)
Osteoporosis (%)	2758 (3.22%)	177 (15.51%)	1.06 (0.88, 1.29)
Osteoarthritis (%)	5868 (6.85%)	319 (27.96%)	1.03 (0.89, 1.20)
Arthritis (%)	1214 (1.42%)	50 (4.38%)	1.44 (1.05, 1.98)
Dementia (%)	911 (1.06%)	253 (22.17%)	1.78 (1.50, 2.12)
Chronic pain (%)	50 (0.06%)	2 (0.18%)	–

típica del triaje de ingresos en ICU a personas más jóvenes respectos a las hospitalizadas. La relación entre el sexo y el estado del paciente fue significativa (X^2(1, N = 86867) = 34,33, p = 4,64 × 10^{-9}). Así, los hombres tenían más probabilidades de morir que las mujeres (tabla 5.4), se puede observar un gráficos de odds ratios ajustados en la figura 5.5, también se identificó que las comorbilidades o patologías de mayor riesgo eran la insuficiencia renal (OR = 1,90; CI 95%: 1,61; 2. 24), insuficiencia cardíaca (OR = 1,85; CI 95%: 1,53, 2,23), ictus (OR = 1,84; CI 95%: 1,54, 2,20), demencia (OR = 1,78; CI 95%: 1,50, 1,81) y miocardiopatía isquémica (OR = 1,58; CI 95%: 1,33, 1,88). Sin embargo, no hubo una relación significativa entre el asma, la osteoartritis y la osteoporosis con la muerte relacionada con COVID-19.

5.3.2 Modelos predictivos generados con técnicas de ML

Se desarrollaron varios modelos de aprendizaje automático: (1) para predecir el estado final del paciente y (2) para predecir qué paciente necesitará ser hospitalizado. El conjunto

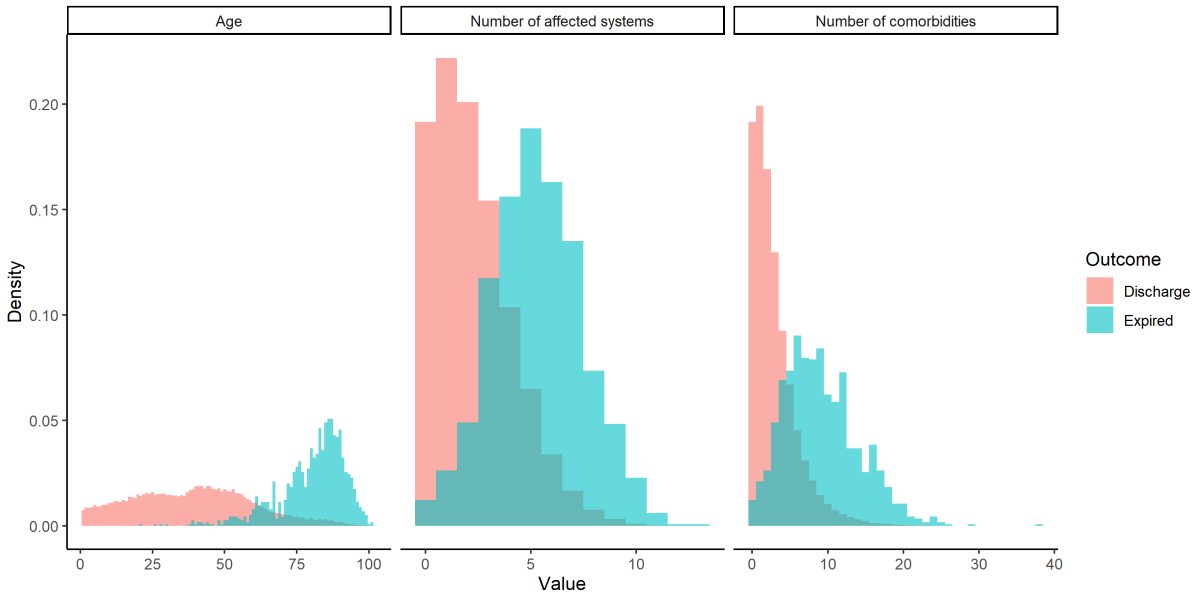


Figura 5.3: Distribución de edad, numero de comorbilidades y sistemas afectados según el estado final del paciente.

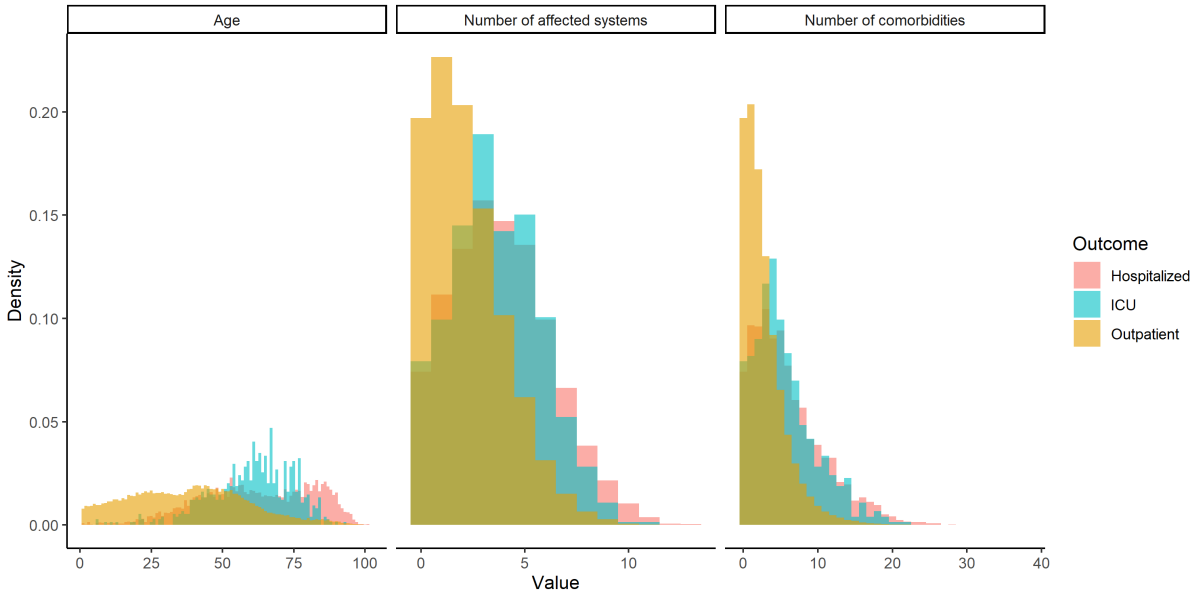


Figura 5.4: Distribución de edad, numero de comorbilidades y sistemas afectados según la hospitalización del paciente.

de datos de entrenamiento (85.476 pacientes supervivientes y 891 no supervivientes) se utilizó para entrenar el modelo de predicción del estado final del paciente y el conjunto de datos de prueba (500 pacientes; 250 pacientes de cada clase) se utilizó para evaluar este modelo. Para evaluar el modelo de forma realista, se crearon 101 conjuntos de prueba con la misma proporción que el conjunto inicial (1/75), es decir, por cada paciente COVID-19

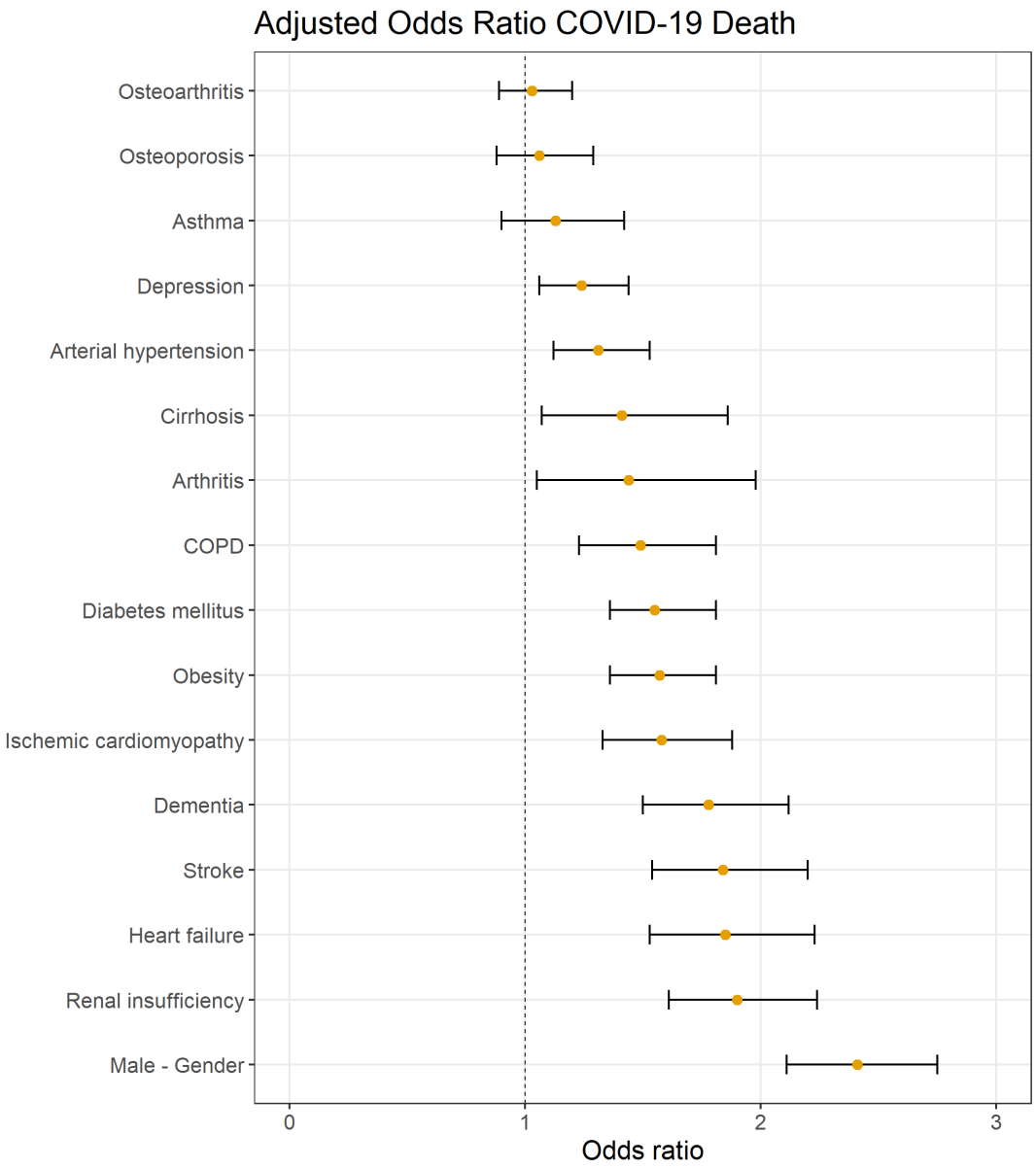


Figura 5.5: Odds ratio de riesgo de fallecer por COVID-19 ajustados de distintas características y comorbilidades.

fallecido se tomaron 75 pacientes supervivientes.

Se evaluaron dos algoritmos de aprendizaje automático Random Forest (RF) y Regresión logística (LR) con o sin IPIP, un método para tratar datos desequilibrados (véase la sección 5.2.4). La precisión y el Kappa de Cohen obtenidos en el conjunto de datos de prueba para los modelos ensemble se muestran en la figura 5.6 y mostraron que el modelo IPIP con regresión logística (LR-IPIP) obtuvo los mejores resultados en cuanto al estado final del paciente. Este modelo LR-IPIP combinó el resultado del conjunto

de 21 modelos de regresión logística. El estado final del paciente puede predecirse con el modelo LR-IPIP con una precisión equilibrada entre 0,90 y 0,93 (tabla 5.5) para los conjuntos de datos desequilibrados frente a 0,91 para los conjuntos de datos equilibrados, los conjuntos de datos equilibrados son los que se suelen utilizar en la literatura y que dan un coeficiente Kappa de Cohen más alto (0,83 frente a 0,20). Además, el ROC-AUC de este modelo para los conjuntos de datos desequilibrados fue de 0,94 (se puede observar en la figura 5.7). Los factores más importantes que determinan el estado final del paciente, *Feature Importance* en inglés (FI), obtenidos por este modelo LR-IPIP fueron en primer lugar la edad (FI: 1,0), seguido del sexo (FI: 0,366), la artrosis (FI: 0,194), la insuficiencia renal (FI: 0,144), la obesidad (FI: 0,132) y el número de sistemas afectados (FI: 0,117) (figura 5.8 parte a).

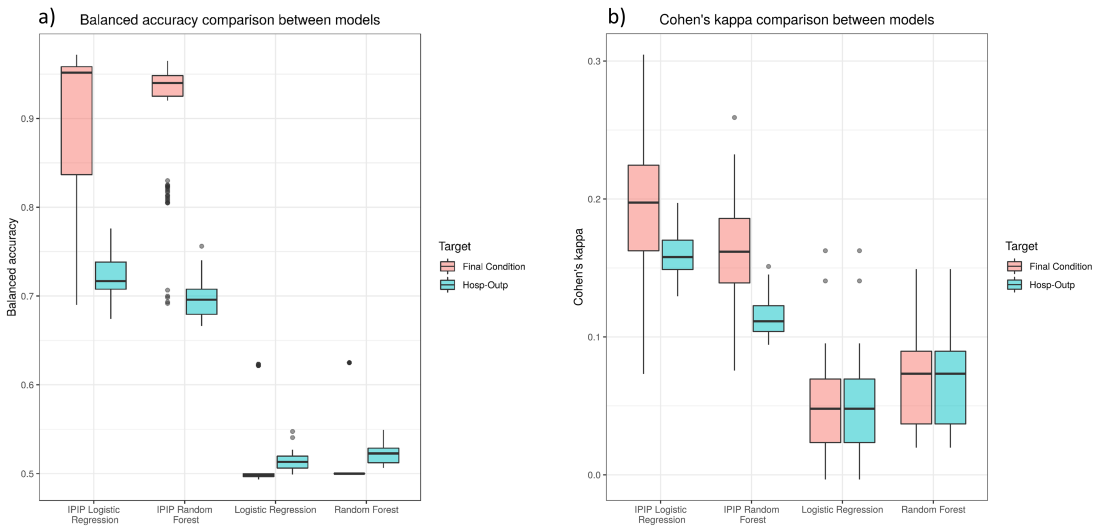


Figura 5.6: Comparación de los distintos modelos generados en cuanto a precisión y Kappa de Cohen.

Por otro lado, el modelo para predecir qué paciente necesitará ser hospitalizado se desarrolló utilizando un conjunto de datos de entrenamiento (4.385 pacientes hospitalizados y 80.290 pacientes ambulatorios), y para evaluar este modelo se utilizó un conjunto de datos de prueba (2.192 pacientes; 1.096 pacientes de cada clase). A su vez, estos datos se distribuyeron en 25 conjuntos de prueba con la misma proporción de pacientes hospitalizados/ambulatorios que en el conjunto inicial (1/15). De nuevo, el modelo con mejores resultados fue el LR-IPIP compuesto por 13 modelos de regresión

Tabla 5.5: Metricas obtenidas en los conjuntos de datos de test utilizando el mejor modelo preditivo generado en el entrenamiento.

Metrics	Deceased/discharge		Hospitalized/outpatient	
	Imbalanced tests	**Balanced test**	**Imbalanced tests**	**Balanced test**
Balanced accuracy	0.92 (0.90, 0.93)	0.91	0.72 (0.71, 0.73)	0.72
Cohen's Kappa	0.20 (0.18, 0.21)	0.83	0.16 (0.15, 0.17)	0.44
Sensitivity	0.93 (0.88, 0.96)	0.92	0.72 (0.70, 0.74)	0.71
Specificity	0.91 (0.90, 0.91)	0.91	0.73 (0.72, 0.73)	0.73
Positive predictive value (PPV)	0.12 (0.11, 0.13)	0.91	0.15 (0.14, 0.15)	0.72
Negative predictive value (NPV)	1 (0.99, 1)	0.92	0.98 (0.97, 0.98)	0.72

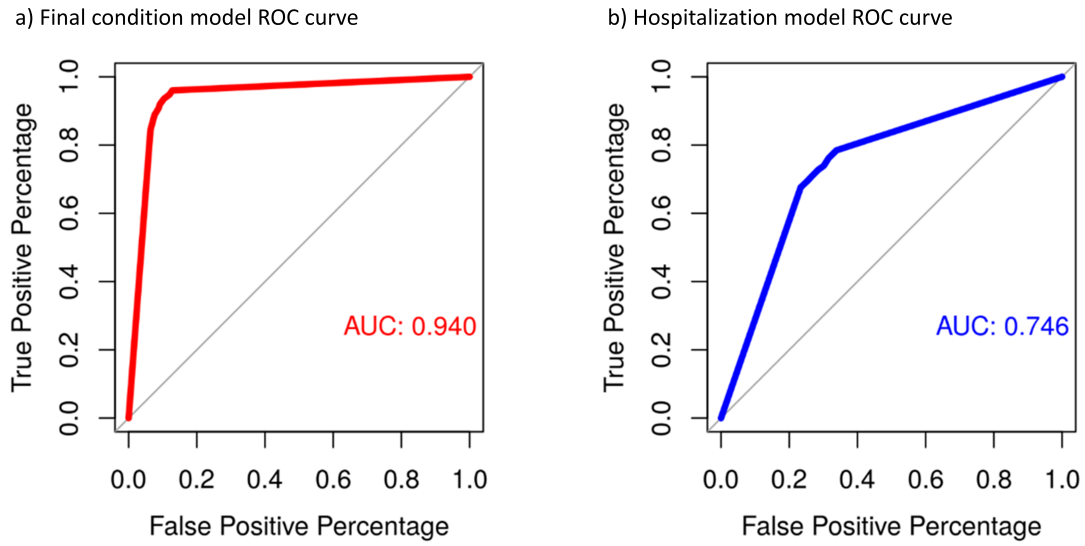


Figura 5.7: Curvas ROC de los modelos predictivos.

logística (figura 5.6 parte b). La necesidad de hospitalización pudo predecirse con el modelo LR-IPIP con una precisión equilibrada entre 0,71 y 0,73 (se puede observar en la tabla 5.5) para los conjuntos de datos desequilibrados frente a 0,72 para los conjuntos de datos equilibrados. De forma similar al otro modelo, el coeficiente Kappa de Cohen es mayor para el conjunto de datos equilibrado (0,44 frente a 0,16). Además, el ROC-AUC de este modelo para los conjuntos de datos desequilibrados fue de 0,746 (sección b de la figura 5.7). Por último, la importancia de las características obtenidas en dicho modelo mostró que la edad (FI: 1,0) era también la característica más relevante, seguida del sexo (FI: 0,26), la insuficiencia renal (FI: 0,12), el número de enfermedades crónicas (FI: 0,11) y la depresión (FI: 0,1) (figura 5.8 parte b).

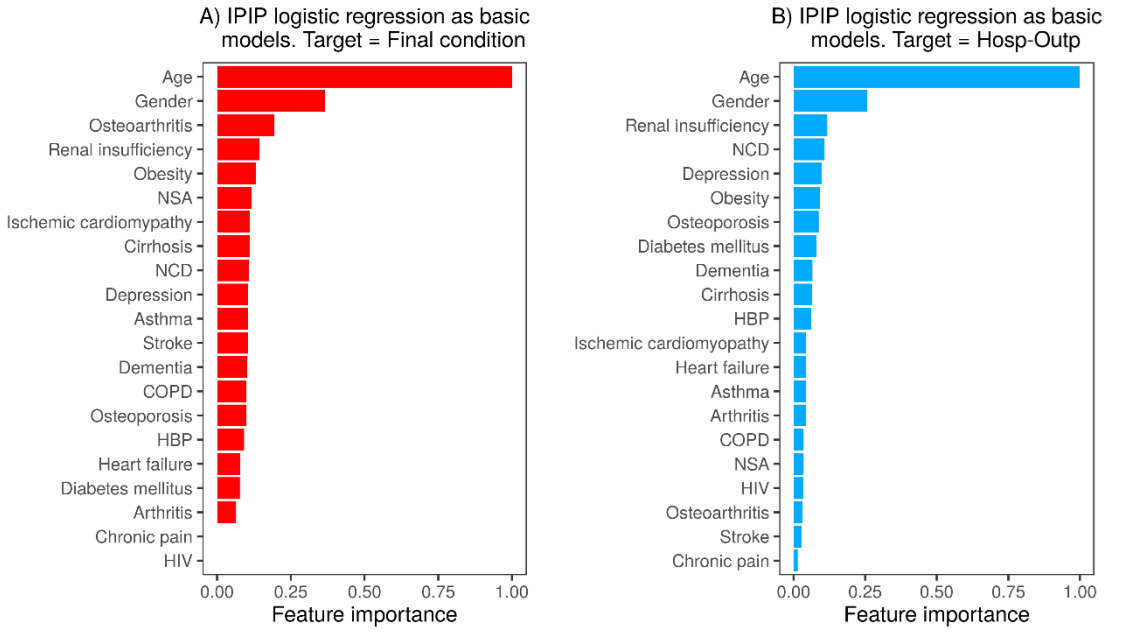


Figura 5.8: Importancia de las variables en el modelo predictivo.

5.4 Conclusiones

En este estudio, hemos analizado los diferentes tipos de pacientes COVID-19 en el sureste de España (n = 86.867). A diferencia de la mayoría de los estudios de COVID-19 que han desarrollado modelos predictivos en la literatura que manejan menos de 5000 pacientes [148, 149, 150, 151, 152]. Además, hemos presentado una técnica especialmente diseñada para tratar problemas de desequilibrio (IPIP), con la que hemos desarrollado modelos de aprendizaje automático para predecir el estado final del paciente y la necesidad de hospitalización de los mismos. Hemos entrenado y evaluado los modelos con y sin IPIP, que gestiona eficientemente el desequilibrio en los datos según nuestros resultados (figura 5.6).

En cuanto a la caracterización de los diferentes tipos de pacientes COVID-19 prototípicos, en esta región, el tipo de paciente COVID-19 más común es el que no requirió hospitalización, normalmente mujer de 38 años, con 2 patologías crónicas, mientras que el prototipo de paciente COVID-19 hospitalizado es un hombre de 62 años, con 5 patologías crónicas. Identificamos la edad, el sexo y el número de comorbilidades como factores importantes para distinguir entre pacientes ambulatorios y hospitalizados. Varios estudios también han encontrado que los pacientes COVID-19 hospitalizados son más comúnmente mayores, varones y están asociados con más comorbilidades como obesidad, diabetes mellitus e hipertensión [156, 157]. Además, pudimos encontrar diferencias

estadísticamente significativas para la edad ($p < 8{,}0 \times 10^{-3}$), el número de comorbilidades ($p < 2{,}5 \times 10^{-3}$) y el sexo ($p < 2{,}2 \times 10^{-16}$) entre los pacientes de la ICU y los no hospitalizados, aunque esas diferencias son menores que entre los pacientes ambulatorios y los hospitalizados. Los pacientes de la ICU eran alrededor de un año más jóvenes que los pacientes hospitalizados fuera de la ICU y ligeramente menos comorbilidades (figura 5.4). Por lo tanto, planteamos la hipótesis de que los clínicos incluyeron en la ICU a pacientes con más probabilidades de sobrevivir debido al número limitado de plazas disponibles en la ICU o al riesgo asociado al sexo masculino. También detectamos aún más diferencias para esas características entre los supervivientes (pacientes dados de alta) y los no supervivientes (pacientes fallecidos) (figura 5.3). En nuestra región, el prototipo del paciente que no fallece es una mujer de 39 años, con 2 patologías crónicas, mientras que el prototipo de paciente fallecido es un hombre de 83 años, con 8 patologías crónicas. De acuerdo con diversos estudios, nuestros resultados muestran que los pacientes de mayor edad tienen más probabilidades de fallecer [158, 159, 160] y también los pacientes varones tienen más probabilidades de fallecer (OR = 2,41; CI 95%: 2,11; 2,75) (tabla 5.4, figura 5.5) [161, 162]. En lo que respecta a las comorbilidades, encontramos que el asma, la osteoporosis y la osteoartritis no están asociadas con la muerte relacionada con la COVID-19. Un buen número de estudios informan de que los pacientes con asma no tienen más riesgo de COVID-19 grave [163, 164]. Para la asociación de la osteoartritis con la muerte relacionada por COVID-19 encontramos un estudio que informó de una OR similar = 0,84 (CI del 95%: 0,65-1,08) [165]. En cuanto a la osteoporosis, se sabe que las mujeres tienen más riesgo de desarrollar osteoporosis que los hombres [166]. Parece que algunos tipos particulares de complicaciones de la osteoporosis se asocian con más riesgo de COVID-19 exitus, sin embargo, este estudio no ajustó el riesgo por edad y sexo [167]. El resto de comorbilidades evaluadas en nuestro estudio se asociaron a un aumento del riesgo de mortalidad. Estas comorbilidades o patologías son la diabetes mellitus, la demencia, la obesidad, la insuficiencia cardíaca, la COPD, la hipertensión arterial, la miocardiopatía isquémica, el ictus, la insuficiencia renal, la cirrosis y la artritis. Varios estudios obtienen los mismos resultados para esas comorbilidades [165, 168, 169]. Con respecto a la depresión, en línea con nuestros resultados, un meta-análisis identificó que la depresión se asocia con más muertes relacionadas con COVID-19 [170]. Todos los resultados mencionados anteriormente son importantes para asegurar que las características y comorbilidades de nuestra población no fueran únicas. Además, creemos que debido a la similitud con otros estudios de COVID-19, nuestros datos podrían ser útiles para desarrollar modelos predictivos.

Desde el comienzo de la pandemia, se han realizado muchos estudios que han informado algunas características clínicas importantes (predictores) de mortalidad en pacientes con COVID-19 a través del desarrollo de modelos basados en ML. Las características seleccionadas utilizadas como entradas para el desarrollo de estos modelos incluyeron datos demográficos, síntomas clínicos, comorbilidad asociada y resultados de análisis o pruebas. Sin embargo, estos estudios tienen dos problemas fundamentales: el bajo número de pacientes debido a que la cantidad de parámetros estudiados, lo que puede restringir mucho el número de individuos en la cohorte, y los datos fuertemente desequilibrados entre clases superviviente y fallecido. Para superar estos inconvenientes, en este trabajo probamos diferentes modelos de aprendizaje automático teniendo en cuenta los datos básicos fácilmente accesibles en un entorno de atención de emergencia y basados en datos clínicos de EHR para ayudar durante la clasificación temprana de pacientes. Definitivamente obtuvimos resultados prometedores al predecir la condición final del paciente usando el modelo LR-IPIP (precisión balanceada de 0.92, ROC-AUC = 0.94). En cuanto a la importancia de las variables, ML detecta la edad (FI: 1,0), el sexo (FI: 0,366), la artrosis (FI: 0,194), la insuficiencia renal (FI: 0,144), la obesidad (FI: 0,123) y el número de sistemas afectados (FI: 0,117) como las variables más importantes para predecir el exitus. El modelo también detectó comorbilidades como demencia, diabetes mellitus y COPD. Estas características están asociadas con un mayor riesgo de muerte relacionada con COVID-19 según nuestro modelo. En consonancia con nuestros resultados, estas comorbilidades se asocian con manifestaciones clínicas graves observadas en pacientes adultos mayores [171, 172].Las comorbilidades como las enfermedades cardiovasculares, la hipertensión y la diabetes, aunque son muy prevalentes en los adultos mayores, se han asociado con peores resultados en la COVID-19 [165, 168, 169]. Los estudios que se basan en las comorbilidades para predecir la muerte en función de modelos generados con ML suelen situar la edad como una de las variables más influyentes [173, 174], de hecho, un meta-análisis con 611.583 pacientes demuestra un aumento de la mortalidad relacionado con la edad. Así, la mayor mortalidad se da en pacientes con más de 80 años, en los que fue 6 veces más probable el fallecimiento que en pacientes más jóvenes [175]. De manera similar, el género es una característica importante para varios estudios basados en ML [173, 176], nuestro modelo identificó que los pacientes varones tienen más probabilidades de morir, algo que ya se detectó en el análisis estadístico de nuestros datos (OR = 2.41, 95% CI: 2.11, 2.75), lo cual está de acuerdo con trabajos previos [161, 162]. Similar a nuestro modelo, otro estudio basado en ML identificó la obesidad como una característica importante [177]. Sin embargo, hasta

donde sabemos, esta es la primera vez que un modelo informa que la osteoartritis es una característica importante. Los valores beta en el modelo de regresión mostraron que la osteoartritis está asociada con un menor riesgo de muerte relacionada con COVID-19 según el modelo de ML (se pueden consultar los valores en la tabla s2 de este link `https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-022-22547-9/MediaObjects/41598_2022_22547_MOESM1_ESM.pdf`). Esto podría estar de acuerdo con un estudio que utilizó datos del biobanco del Reino Unido (OR = 0.84, 95% CI 0.65, 1.08), aunque no es estadísticamente significativo [165]. Además, la distribución de la artrosis en nuestra población no se asocia estadísticamente con el estado final del paciente. Se debe tener en cuenta que, aunque no tenemos evidencia concluyente al respecto, los pacientes con artrosis pueden ser sometidos a medicación. Curiosamente, podríamos pensar que la medicación podría desempeñar un papel en pacientes con osteoartritis y COVID-19, sin embargo, Wong et al. informó que la medicación con medicamentos antiinflamatorios no esteroideos no está asociada con un mayor riesgo de muerte por COVID-19 para los pacientes con osteoartritis [178]. La demencia, junto con el número de sistemas afectados y el número de comorbilidades, también aparecen entre las características más relevantes, lo que concuerda con los factores mencionados en otros estudios, y en el caso de la demencia, con los resultados obtenidos en una cohorte de 12.863 personas del biobanco del Reino Unido que tenían más de 65 años (con 1814 personas mayores de 80 años) se sometieron a la prueba de COVID-19, donde se observó que todas las causas de demencia aumentaban el riesgo de muerte relacionado con COVID-19 [179]. En cuanto a la precisión, nuestro modelo LR-IPIP obtuvo una precisión equilibrada entre el 89 y el 93% (ROC-AUC = 0,94) en la predicción del estado final del paciente. Es decir, en predecir si el paciente fallecería o sobreviviría. La precisión fue similar o superior a otros modelos si comparamos nuestros resultados con varios estudios. Por ejemplo, Gao et al. informaron una precisión entre 80,6 y 96,8% [149], que es un intervalo de confianza grande, además de que utilizaron puntos de datos clínicos más complejos de obtener en el momento de ingreso. Chatterjee et al. informó una precisión equilibrada del 72% [151],una precisión moderada quizás debido a la baja cantidad de pacientes con COVID-19 en este estudio. Finalmente, otro estudio basado en ML pudo predecir el riesgo de muerte ya en el momento del diagnóstico con un ROC-AUC de 0,902 [152].

La capacidad del modelo LR-IPIP para decidir la hospitalización de nuevos pacientes no fue tan eficiente (precisión equilibrada = 0,72; ROC-AUC = 0,75). En cuanto a la importancia de las variables, el modelo de ML volvió a encontrar que la edad, el sexo y el número de comorbilidades eran importantes para realizar buenas predicciones. Entre

éstas, reaparece la obesidad, y en un lugar destacado la insuficiencia renal y la depresión. Así, se ha demostrado que el fracaso renal agudo es frecuente entre los pacientes graves hospitalizados por COVID-19 y que sólo el 30 % sobrevivió con recuperación de la función renal al alta [180].

Esta investigación presenta problemas concretos del ámbito de estudio. En primer lugar, debido al carácter altamente específico de esta cohorte y a su inevitable novedad, no pudimos obtener fácilmente una cohorte alternativa que pudiera utilizarse para replicar y validar nuestros hallazgos. Afortunadamente, esto se superó en parte por el hecho de que los individuos procedían de diversos hospitales de nuestra región con gestión compartida de datos de historias clínicas electrónicas. Al tratarse de un estudio retrospectivo, la falta de algunos datos se compensó incluyendo en el estudio sólo los datos demográficos y las comorbilidades que se habían registrado correctamente. En segundo lugar, otra dificultad se deriva del fuerte desequilibrio de datos inherente a la pregunta de investigación que formulamos. Intentamos compensarlo con el desarrollo del método IPIP. En tercer lugar, hay que señalar que los datos utilizados para construir los modelos se obtuvieron en ausencia de pautas de vacunación y de nuevas variantes del virus SARS-COV-2. Sin embargo, la metodología para construir los modelos puede adaptarse fácilmente a estos nuevos escenarios. Por último, una mejor comprensión de la contribución de los distintos síntomas o comorbilidades al diagnóstico de la enfermedad podría servir para introducir nuevas características en futuros modelos, especialmente para mejorar la predicción de los pacientes que requieren o no hospitalización.

En conclusión, este trabajo muestra el análisis y desarrollo de modelos predictivos basados en ML con uno de los mayores conjuntos de datos COVID-19 (n = 86.867) hasta la fecha que ha sido obtenido del servicio de salud de la Región de Murcia (España). Además, se ha abordado el problema del desequilibrio de clases mediante el desarrollo de un nuevo algoritmo, denominado IPIP, que se ocupa automáticamente de este problema. El modelo obtenido permite predecir con gran exactitud el estado final del paciente, y con razonable precisión qué paciente necesitará ser hospitalizado, simplemente utilizando los datos demográficos y las comorbilidades accesibles en el diagnóstico COVID-19 por los clínicos. De hecho, este modelo predictivo LR-IPIP puede utilizarse, entre otras consideraciones, para priorizar el triaje de los pacientes COVID-19 cuando los recursos del sistema sanitario son limitados, como suele ocurrir durante las diferentes oleadas de COVID-19. Para facilitar esta priorización de recursos, tanto la aplicación web correspondiente como los modelos predictivos son fácilmente accesibles en repositorios abiertos (GitHub), lo

que facilitará su adaptación a nuevos conjuntos de datos de futuras oleadas epidémicas de esta enfermedad o de otros virus respiratorios en general.

En lo que respecta a la tesis este capítulo es relevante por diversos motivos relacionados con los objetivos marcados en el capítulo 1. Hemos combinado la utilización de estadística y ML para extraer información de datos clínicos. Aplicando esto hemos generado modelos predictivos que pueden ayudar a clasificar el riesgo de mortalidad y de hospitalización de pacientes de COVID. Además, hemos facilitado la difusión y el uso de estas herramientas tanto para informáticos por medio del GitHub como de usuarios finales por medio de la web.

CAPÍTULO 6

Transcriptómica, genómica y datos clínicos para el diagnóstico temprano de Alzheimer

Este capítulo presenta la investigación realizada durante la estancia doctoral realizada en la Washington University en St. Louis, Missouri (Estados Unidos de América). Durante esta estancia la investigación estuvo dirigida por la doctora Laura Ibáñez. Gracias a ser un centro de vanguardia en el estudio de enfermedades neurodegenerativas tuvimos acceso a explotar un conjunto de datos de transcriptómica de Alzheimer único en el mundo. Dicho conjunto consiste en la secuenciación de RNA libre circulante, *cell-free RNA* (cfRNA) en inglés, procedente del plasma de pacientes de Enfermedad de Alzheimer, *Alzheimer Disease* (AD) en inglés, en distintos estadios de la enfermedad y también de individuos sanos de edad similar. El conjunto de datos con el que hemos trabajado es de una gran calidad y tenemos acceso gracias a la investigación realizada por la doctora Laura Ibáñez.

El motivo principal para desarrollar esta investigación es que se necesitan biomarcadores sanguíneos asequibles, escalables y específicos para la enfermedad de Alzheimer que puedan aplicarse a nivel poblacional. Es importante recordar que como mencionábamos en la introducción de esta tesis, un biomarcador es una característica biológica que nos índica el estado de un proceso o enfermedad. En este estudio hemos desarrollado biomarcadores sanguíneos utilizando el cfRNA de pacientes con AD. Los modelos desarrollados con estos biomarcadores están compuestos por un número escalable de

transcritos que captan los mecanismos moleculares del Azheimer incluso en las fases presintomáticas de la enfermedad, es decir, cuando el paciente no ha mostrado síntomas. Con escalable nos referimos ha que tienen pontencialidad de ser utilizados en clínica en el futuro.

Creemos que es importante detectar la enfermedad en una fase precoz o temprana, antes de mostrar síntomas de deterioro cognitivo, de conseguir esto las terapias existentes podrían ser más efectivas manteniendo la calidad de vida del paciente durante más tiempo. En nuestro caso, las precisiones obtenidas se sitúan en el rango de los biomarcadores actuales procedentes de la extracción de líquido cefaloraquídeo, *Cerebrospinal fluid* (CSF) en inglés, y las especificidades son elevadas frente a otras enfermedades neurodegenerativas. Debido a la gran variedad de enfermedades neurodegenerativas y a la dificultad para diagnosticar AD, la confirmación final es un análisis del cerebro postmorten, se dan muchos diagnósticos erróneos, es decir, los médicos se equivocan con otras enfermedades neurodegenerativas en las que el paciente muestra síntomas similares. Por todo ello es importante que el biomarcador sea específico únicamente de la enfermedad que se quiere diagnosticar. Con un diagnóstico concreto se puede tratar mejor a un paciente.

Es crucial entender que la identificación y validación de un biomarcador es un proceso complejo y que requiere de una extensa investigación. Además, es muy importante que los biomarcadores se estudien en diferentes poblaciones y se evalúen en diferentes contextos clínicos para garantizar su utilidad. En esta investigación hemos iniciado ese proceso y hemos encontrado resultados prometedores para el cfRNA como biomarcador de AD utilizando técnicas de ML para generar modelos predictivos de la enfermedad.

6.1 Introducción y estado del arte

La enfermedad de Alzheimer (AD) es un trastorno neurodegenerativo complejo caracterizado clínicamente por la pérdida gradual y progresiva de la memoria y, patológicamente por la presencia de placas seniles (depósitos de beta-amiloide) y ovillos neurofibrilares (depósitos de la proteína tau) en el cerebro [181]. El diagnóstico de AD es complejo por diferentes motivos:

- No hay una prueba confirmatoria: No se puede diagnosticar definitivamente AD mientra que el paciente esté vivo. Hay diferentes pruebas de análisis o imágenes

para poder concluir un diagnóstico probable pero no definitivo. La confirmación definitiva se da cuando el paciente fallece mediante un análisis del cerebro post-morten.

- Pueden aparecer diferentes síntomas: Diferentes síntomas en diferentes personas. Mientras que algunos pueden tener problemas de memoria prominentes, otros pueden tener dificultades más notables con el lenguaje, la visión espacial, o la toma de decisiones. Sumado a la progresión variable de la enfermedad en las etapas tempranas dificulta el reconocimiento de la enfermedad.

- Variedad de enfermedades neurodegenerativas que pueden presentar síntomas similares: Como por ejemplo la demencia con cuerpos de Lewy, la enfermedad de Parkinson, la demencia frontoremporal o la demencia vascular cerebral. En definitiva, enfermedades con síntomas similares que pueden llevar a un diagnóstico incorrecto.

Estos factores pueden conducir a diagnósticos tardíos o diagnósticos incorrectos. Por tanto puede afectar al tratamiento adecuado de los pacientes y por ende a su calidad de vida.

Desde el punto de vista económico, se ha calculado que la AD y otras demencias costaron aproximadamente 355 billones americano de dólares en 2021, un coste que se ha estimado que aumentará hasta 1,1 trillones de dólares en 2050 [182]. La disponibilidad de una herramienta de diagnóstico precoz y preciso de la AD podría ahorrar 7,9 trillones de dólares en costes médicos y asistenciales [183]. En la actualidad, se están dirigiendo muchos esfuerzos a encontrar biomarcadores rentables, fiables, específicos y no invasivos para la AD que puedan utilizarse para identificar a individuos en fase presintomática, y a pacientes en fases sintomáticas tempranas de la enfermedad (individuos con AD preclínica o deterioro cognitivo leve [184, 185].

Los biomarcadores de imagen y de CSF se utilizan habitualmente para el diagnóstico de la enfermedad de Alzheimer [186, 187, 188]. El biomarcador de CSF más utilizado y preciso es el cociente Aβ42/Aβ40, que puede diagnosticar correctamente al 82,8 % de los pacientes con Alzheimer examinados [189]. Aβ es un péptido, una molécula que consiste en varios aminoácidos enlazados y que son más pequeñas que las proteínas, procedente de la proteína APP, que está relacionado con AD pero que se cree que puede tener otras funciones. Se ha comprobado que las mediciones de Aβ42/Aβ40 en el CSF parecen ser

bastante específicas y permiten diferenciar la AD de la demencia con cuerpos de Lewy, *Lewy body dementia* (LBD) en inglés, la enfermedad de Parkinson (PD) y la demencia vascular, *vascular dementia* (VaD) en inglés [190]. Sin embargo, la estandarización de las mediciones para su uso en la práctica clínica ha sido un reto y continua presentando dificultades [191, 192]. Junto con las mediciones de Aβ, los niveles de tau fosforilada (p-tau) y tau total (t-tau) en CSF o cerebro también se utilizan para ayudar al diagnóstico de AD [193, 194]. La proteína tau se encuentra en las neuronas del sistema nervioso central y juega un papel crucial en la estabilización de las células nerviosas. Sin embargo, cuando la proteína tau se fosforila en exceso, es decir se le añaden grupos fosfato, se pueden formar ovillos neurofibrilares, que son agregados insolubles de proteína tau en las células cerebrales que pueden tener relación con la demencia. La t-tau está elevada en otras enfermedades neurodegenerativas como la demencia frontotemporal, *Frontotemporal Dementia* (FTD) en inglés, la VaD y la enfermedad de Creutzfeldt-Jacob [195]. En cambio, ciertas especies de p-tau medidas en CSF, como p-tau181 y p-tau231, son más específicas de la AD y muestran fuertes correlaciones con los PET de tau [196, 197]. Un PET, o Tomografía por Emisión de Positrones, en inglés *Positron Emission Tomograph*, es una técnica de imagen médica que permite visualizar funciones biológicas en el cuerpo y que se utiliza como técnica de imagen en AD o en cáncer. Para mejorar el diagnóstico de la AD, el marco de Neurodegeneración (N) Amiloide (A) Tau (T) propuso una clasificación biológica de la AD en ocho perfiles según la positividad/negatividad de los tres biomarcadores, Aβ(A), p-tau (T) y t-tau (N) [198]. Bajo el marco de Neurodegeneración (N) Amiloide (A) Tau (T), marco ATN [199], se acepta que dar positivo para Aβ significa el comienzo del continuo de la enfermedad de Alzheimer [200, 201]. El aumento del número de biomarcadores positivos para los criterios de la ATN se correlaciona con una patología más avanzada, y se asocia con un mayor riesgo de demencia y deterioro cognitivo [200, 201]. Uno de los principales retos de los criterios ATN es la definición de los valores de corte de los biomarcadores, especialmente para la selección de individuos presintomáticos de AD [202, 203, 204].

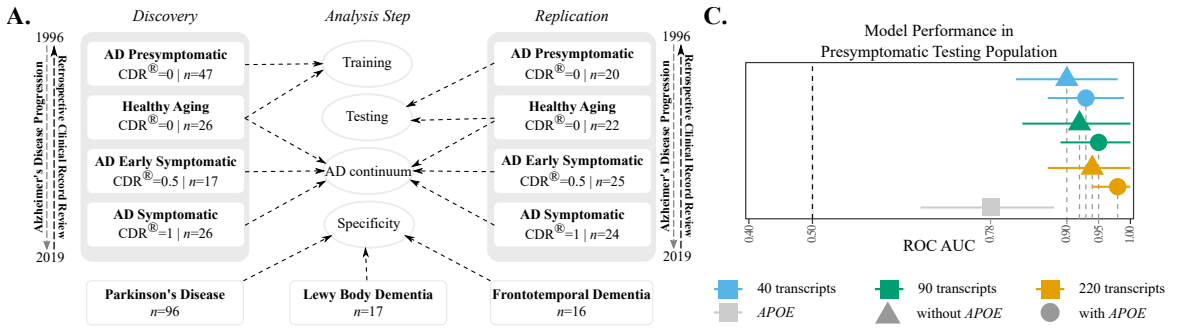
Dado que los biomarcadores de CSF y de imagen son invasivos y caros, en la última década se ha intensificado el estudio de los biomarcadores sanguíneos. Estos son menos invasivos y pueden proporcionar una precisión comparable a las medidas de CSF y de imagen [205, 206]. Por ejemplo, se ha observado que la t-tau plasmática es mayor en las fases avanzadas de la AD [206, 207], pero, al igual que ocurre con las mediciones tomadas del CSF, no parece ser específica de la AD [208]. Nuevas pruebas sugieren que las especies fosforiladas de tau, especialmente p-tau 217, son específicas de la AD, con

valores que aumentan progresivamente de individuos sanos a individuos con deterioro cognitivo y AD más avanzado [209, 210, 211]. La relación Aβ42/Aβ40 en plasma se correlaciona altamente con la amiloidosis cerebral [212], especialmente cuando se mide con técnicas de alta precisión y se combina con el genotipo APOE, el principal factor de riesgo genético para la el Alzheimer [213]. El gen APOE está implicado en el metabolismo de las grasas en el cuerpo. Existen tres variantes principales del gen, llamadas ε2, ε3 y ε4. Cada individuo hereda una copia de un gen APOE de su madre y otra de su padre, por lo que puede tener cualquier combinación de dos de estas variantes. La variante ε4 del gen APOE se ha asociado con un mayor riesgo de desarrollar AD. Al contrario, la variante ε2 se ha asociado a menor riesgo [214]. Además,la amiloidosis cerebral es una acumulación de proteínas amiloides en los vasos sanguíneos del cerebro, no es un proceso único de AD. Cuando las diferentes proteínas amiloides y péptidos se agregan se pueden formar placas amiloides y perjudicar al tejido cerebral . Este proceso puede darse en diferentes demencias y no es único de AD o asociado a enfermedad, por ejemplo, también están presentes en casi un tercio de la gente mayor de 70 años y solo el 25% de estos presenta AD [215]. En definitiva, por los factores mencionados junto con que la rentabilidad y la escalabilidad no son óptimas es difícil la implantación definitiva de estos biomarcadores.

Como hemos comentado en el anterior párrafo, la mayoría de los estudios de biomarcadores en sangre o CSF miden los niveles de proteínas, además, presentan los problemas descritos anteriormente; sin embargo, los ácidos nucleicos también pueden utilizarse como biomarcadores y estos han sido pocos estudiados hasta ahora en el campo. Este tipo de biomarcadores ha mostrado resultados prometedores, la prueba diagnóstica de ADN libre de células (cfADN) permite detectar trastornos genéticos y anomalías cromosómicas durante el embarazo y revolucionó el cribado prenatal al evitar los riesgos de aborto espontáneo relacionados con el procedimiento anterior [216]. El plasma también contiene ácido ribonucleico en su forma libre (cfRNA), fragmentos de ARN que se encuentran en las células, que tiene el potencial de captar procesos temporales ya que parece que procede de la muerte celular que constantemente se da en el organismo [217]. Es obvio que el mejor tejido para estudiar esto sería el cerebro pero es evidente que no se pueden tomar muestras del tejido cerebral de los pacientes cuando están vivos, además, la extracción de CSF también es invasiva, por lo tanto el plasma puede ser una muy buena opción en AD. Se han investigado intensamente varias especies de cfRNA como biomarcadores del cáncer [218, 219], del desarrollo fetal [220] e incluso en Alzheimer [221, 222, 223]. Aunque varios estudios propusieron los

microARN circulantes como biomarcadores de Alzheimer [222, 224, 225, 226, 227], sólo un estudio publicado utilizó cfARN mensajeros plasmáticos para captar alteraciones transcriptómicas y lo realizaron centrados en estadios avanzados de la enfermedad [223]. En sus comparaciones entre casos de AD (n=122) y controles (n=116), identificaron 2591 transcritos con una expresión diferencial entre esas condiciones. A continuación, utilizaron la información transcriptómica para construir clasificadores que discriminaran a los pacientes con AD de los controles sanos con un área bajo la curva (AUC) de 0,83 [223]. Aunque prometedores, los modelos incluyen la mayoría de los genes expresados diferencialmente (1658), en lugar de un subconjunto de los genes más informativos, lo que mejoraría la escalabilidad y facilitaría la traslación a un entorno clínico. Ningún estudio hasta ahora ha evaluado el uso de cfRNA como un enfoque potencial para desarrollar biomarcadores de Alzheimer clínicamente útiles para fases presintomáticas de la enfermedad.

Aquí, a diferencia de publicaciones previas [223, 228], aprovechamos el cfRNA plasmático de participantes presintomáticos de AD para capturar los cambios tempranos causados por la patología de AD y para construir modelos que sean capaces de mostrar un buen desempeño con un número escalable de transcritos para facilitar su potencial aplicación en la clínica. También evaluamos las capacidades de clasificación de los modelos presintomáticos de AD propuestos en el contexto del espectro de la AD, PD, la LBD y la FTD para asegurar que los modelos capturaban selectivamente los cambios asociados de la AD, es decir, que aseguren un diagnóstico diferencial específico. Utilizamos estas otras enfermedades neurodegenerativas para ejemplificar la variedad de enfermedades que pueden presentar síntomas similares al AD. Esto, junto con la dificultad para confirmar el diagnóstico, el cual solo se sabe al 100 % en un análisis del cerebro postmorten, produce diagnósticos erróneos de estas enfermedades. Por todo ello es importante que el biomarcador sea lo mas específico posible de la enfermedad que se quiere diagnosticar. En la figura 6.1A se resume el diseño y el enfoque analítico del estudio. Además, se muestra la selección de los individuos, basado en revisión retrospectiva de historias clínicas, los grupos y subgrupos incluidos en el descubrimiento y la replicación, y el enfoque con otras enfermedades neurodegenerativas. Creemos que es crucial encontrar biomarcadores para pacientes presintomáticos de AD, antes de mostrar síntomas visibles de deterioro, ya que las terapias existentes podrían ser más efectivas. Por tanto, se mantendría la calidad de vida del paciente durante más tiempo y se ahorraría en costes económicos asociados.



B.

Characteristic	Discovery Dataset (Training) Control	Discovery Dataset (Training) Presymptomatic	Replication Dataset (Testing) Control	Replication Dataset (Testing) Presymptomatic
N	26	47	22	20
Male (N) (%)	13 (50.00%)	19 (40.43%)	7 (31.82%)	9 (45%)
Age at draw median (IQR)	77.00 (73.25, 82.75)	83.00 (78.50, 87.00)	71.87 (70.55, 73.03)	73.25 (70.76, 81.29)
Age at onset median (IQR)	-	88.00 (81.50, 91.00)	-	78.50 (72.25, 83.25)
Disease duration median (IQR)	-	-4.00 (-6.00, -2.00)	-	-2.31 (-4.06, -1.27)

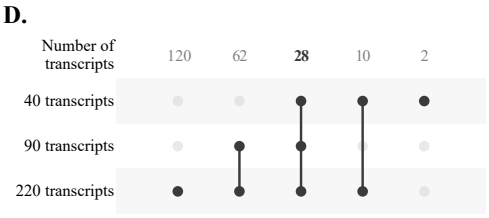


Figura 6.1: Planteamiento, características de la población y resultados generales del estudio de cfRNA como biomarcador de AD.

6.2 Métodos

6.2.1 Diseño del estudio

El RNA se extrajo de muestras de plasma de participantes con Alzheimer, controles y otras enfermedades neurodegenerativas de dos cohortes independientes, tanto del Knight-ADRC como de la Clínica de Trastornos del Movimiento de la Washington University en Saint Louis. Después de la preparación de la librería, la secuenciación y el riguroso control de calidad, comparamos a los participantes con AD presintomáticos con los controles de la cohorte de descubrimiento para identificar cuales eran los transcritos que se expresaban diferencialmente. Comparamos nuestros resultados con los publicados anteriormente en plasma [223] y con los transcritos identificados que se expresan diferencialmente en el cerebro [229] para comprender el origen potencial de los transcritos. Luego, aprovechamos las herramientas de aprendizaje automático para crear modelos predictivos que diferencian entre los participantes con AD presintomáticos y los controles con una cantidad escalable de genes que se replicaron en una cohorte independiente. Finalmente, calculamos el valor predictivo de los modelos en AD sintomática utilizando la escala de clasificación clínica de demencia, *Clinical Dementia Rating* (CDR) en inglés, (CDR=0,5 y CDR=1) para probar si los modelos eran útiles en estadios clínicos de las enfermedades y en otras enfermedades neurodegenerativas como la PD, la LBD, y la FTD, para probar su especificidad el Alzheimer (Fig. 6.1A).

6.2.2 Participantes en el estudio

Las muestras de plasma se obtuvieron de los repositorios Knight-ADRC y MDC por la Washington University en Saint Louis. Estas son cohortes de pacientes con fenotipos conocidos, tanto clínica como molecularmente, con datos longitudinales y una buena cantidad de muestras disponibles. Para el estudio incluimos 48 muestras de participantes de control sanos sin demencia, 67 muestras de participantes con AD presintomática (Clasificación clínica de demencia [230] (CDR) = 0 en el momento de toma de la muestra y con un diagnostico clínico posterior de AD), 42 muestras de participantes con AD sintomática temprana (CDR = 0,5 en extracción y diagnóstico actual de AD) y 50 muestras de AD sintomática (CDR = 1 en extracción, diagnóstico de AD en extracción y diagnóstico actual de AD) (Fig. 6.1A-B y Fig. 6.6A).

Todos los participantes con AD debían tener:

- Evidencia de depósito de Aβ (CSF Aβ<500ng/μL)
- Exploración PET positiva y/o evidencia de empeoramiento clínico medido por CDR desde el momento de la extracción hasta la última visita clínica.

Para 71 participantes, las mediciones de biomarcadores de CSF fueron recogidas a lo largo del tiempo de seguimiento y en el momento de la extracción.

También incluimos en el estudio participantes de otras enfermedades neurodegenerativas: 17 participantes con LBD, 16 participantes con FTD y 96 participantes con PD (Fig. 6.1A y Fig. 6.8A). Esta investigación se llevó a cabo de acuerdo con los protocolos recomendados. Se obtuvo el consentimiento informado por escrito de todos los participantes o sus familiares. La junta institucional de la Washington University en Saint Louis aprobó el estudio (IRB ID 201701124 y 202004010).

6.2.3 Procedimientos de extracción y secuenciación del RNA

Las muestras de plasma se recogen como parte del protocolo de investigación cada dos años para todos los participantes. Una vez obtenida la muestra de sangre, se centrifuga en 20 minutos a 1500 rpm, revoluciones por minuto, durante 10 minutos para obtener plasma y se almacena a -80°C hasta su análisis [231]. Las muestras de plasma seleccionadas de los participantes en el estudio que cumplían los criterios de inclusión se descongelaron en hielo y se centrifugaron a 2000 rpm durante 5 min antes de la

extracción del RNA para evitar la contaminación del RNA celular. Las muestras se procesaron en dos lotes. Para los participantes del lote de entrenamiento con un número total de muestras (N=245), el cfRNA plasmático total se extrajo de 0,5 mL de plasma utilizando el kit Maxwell RSC miRNA from plasma or serum (Ambion) y se le extrajo el RNA ribosomal (NEBNext rRNA Depletion Kit). Para el lote de prueba se obtuvieron las siguientes muestras (N=91), el cfRNA total se extrajo de 1mL de plasma utilizando el kit QIAmp Circulating Nucleic Acid (QIAGEN) seguido de una digestión DNAseI (New England Biolabs). En ambos casos, las librerías se generaron utilizando el NEBNext Ultra II Directional RNA Library Prep Kit para Illumina (New England Biolabs) utilizando 1ng de RNA como entrada. Las librerías se limpiaron en busca de posibles dímeros adaptadores. Se seleccionaron 40 millones de lecturas de un solo extremo de 100 pares de bases para cada muestra utilizando un Illumina NovaSeq 6000 para el lote de entrenamiento, y 15 millones de lecturas de un solo extremo de 100 pares de bases Illumina HiSeq 2500 para las pruebas.

6.2.4 Procesamiento de datos y control de calidad

Utilizamos FastQC (v0.11.7) para evaluar la calidad de secuenciación de cada muestra. A continuación, utilizamos STAR (v2.7.1a) [232] para obtener los archivos BAM y alinearlos con el genoma humano de referencia GRCh38. A continuación, utilizamos PICARD (v2.26) y SamTools para evaluar la calidad de las secuencias y el alineamiento. Por último, utilizamos Salmon (v0.11.3) [233] para cuantificar la expresión de los transcritos. Se utilizó MultiQC (v1.9) [234] para recopilar medidas de control de calidad. Se aplicó un rigurosos control de calidad. Brevemente, tras eliminar todos los genes con menos de diez lecturas en más del 90% de los individuos, calculamos el Análisis de Componentes Principales, *Principal Component Analysis* (PCA) en inglés, del transcriptoma y buscamos correlaciones con variables técnicas y metodológicas para detectar posibles sesgos. Observamos una fuerte correlación con las lecturas totales y las bases codificantes; por lo tanto, eliminamos las muestras con menos del 10% de bases codificantes y menos de 1.000.000 lecturas totales que formaban parte de la misma ronda de secuenciación. También eliminamos muestras atípicas basándonos en el PCA del transcriptoma y las distancias de Cook. Las muestras de plasma se han almacenado durante largos periodos de tiempo antes de su uso (hasta 20 años), en consecuencia, para abordar la degradación, utilizamos DESeq2 (v1.22.2) [235] para encontrar genes asociados con el tiempo de almacenamiento en los participantes de control. Todos los genes nominalmente ($p<0,05$) asociados se eliminaron de los análisis (n=2.580). Por último, utilizamos DESeq2 para

ajustar la complejidad de la librerías y normalizar los recuentos mediante transformación logarítmica para los genes restantes (n=19.830) y obtuvimos la población final y los genes con los transcritos con la suficiente calidad que utilizamos para los presentes análisis.

6.2.5 Análisis de expresión diferencial y rutas biológicas

Los análisis de expresión diferencial se realizaron utilizando DESeq2 [235]. Todos los análisis se ajustaron por sexo y edad en el momento de la extracción. Se utilizó la corrección Benjamini-Hochberg (FDR) para corregir las pruebas múltiples. Los valores p de FDR inferiores a 0,05 se consideraron significativos. Para replicar nuestros hallazgos, utilizamos los genes diferencialmente expresados identificados en cfRNA por Toden et al.[223]. Además, para evaluar si esos genes también eran diferencialmente expresados en los cerebros de los participantes con Alzheimer, utilizamos un conjunto de datos RNAseq interno de cerebros de participantes del Knight-ADRC [229]. Para caracterizar funcionalmente los genes diferencialmente expresados, llevamos a cabo un análisis de enriquecimiento gen-ontológico utilizando ToppGene Suite [236], un análisis de solapamiento de vías de enfermedad utilizando KEGG [237] y un análisis de redes de coexpresión génica utilizando CoExp Web [238] con los datos del Religious Orders Study and Rush Memory and Aging Project (ROSMAP) [239] como matriz de fondo. Para estos análisis, también utilizamos FDR para corregir las pruebas múltiples. Los valores p corregidos inferiores a 0,05 se consideraron significativos.

6.2.6 Construcción y evaluación de los modelos predictivos

Diseñamos diversos modelos predictivos utilizando ML con la finalidad de producir un clasificador adecuado para identificar casos presintomáticos de Alzheimer basado en la expresión génica y utilizando dos conjuntos de datos independientes. En este caso, como en la mayoría de los estudios en biología o medicina que pretenden utilizar ML, tenemos muchas variables o predictores, genes o transcritos en los que tenemos la expresión por medio del cfRNA en este caso, y pocos sujetos, pacientes con AD presintomática en nuestro caso. En estos casos que tienes muchos predictores y pocas muestras o datos de entrenamiento, es decir pacientes, es fácil crear un modelo que se ajuste demasiado bien a los datos de entrenamiento, pero que funcione mal con datos nuevos o de test, es decir que no generalice. Este es un fenómeno conocido como sobreajuste, *Overfitting* en inglés. Además, cuando tenemos este problema de alta dimensionalidad, es difícil visualizar las relaciones y los datos tienden a estar muy dispersos. Lo que hace que sea complicado

para los algoritmos encontrar patrones y hacer predicciones precisas. Pensando en esto hemos diseñado unos procesos, una pipeline, para tener esto en cuenta con dos objetivos principales: (1) Reducir la dimensionalidad para encontrar patrones y hacer el proceso reproducible; y (2) mantener el overfitting bajo control analizando los resultados en el conjunto de datos de test (replication).

El proceso desarrollado para aplicar a nuestro datos ha partido de una normalización rlog [235] después de un proceso de control de calidad de las muestras y la expresión de los transcritos secuenciada del cfRNA de los indiviuos. Primero, escalamos los conjuntos de datos de entrenamiento (Discovery) y test (Replication) utilizando puntuaciones z (Zscore) y generamos un modelo lineal comparando los 47 casos presintomáticos de Alzheimer y los 26 controles en el conjunto de datos de entrenamiento. Hicimos lo mismo para el conjunto de datos de prueba (20 casos de Alzheimer presintomáticos y 22 controles). Sólo mantuvimos los transcritos que tenían la misma dirección del efecto, independientemente del valor p, es decir que se mostraban más altos o más bajos que los controles de forma consistente en los dos conjuntos. Con los genes o transcritos restantes, calculamos la divergencia de Kullback-Leibler (KLD) entre el conjunto de datos de entrenamiento y el de prueba para cada gen, utilizando el paquete R entropy (v1.3.1). La KLD es una medida que cuantifica la diferencia entre dos distribuciones de probabilidad. Hay que tener en cuenta que la divergencia KL no cumple la propiedad de simetría (la divergencia KL de P a Q no es necesariamente igual que la divergencia KL de Q a P). Por tanto, para utilizarla de la forma más objetiva posible hemos calculado la distribución de probabilidad de cada transcrito para cada conjunto de datos utilizando 5 particiones de la distribución y sumando las dos divergencias, es decir, la del transcrito X en el conjunto de entrenamiento E (XE), y la del transcrito X en el conjunto de test T (XT), calculando la KLD en ambos sentidos y sumándola, de XE a XT y de XT a XE. Los transcritos que presenten una KLD más baja tendrán distribuciones de probabilidad parecidas y, por tanto. se mostrarán consistentes. Hemos utilizado esta métrica para reducir dimensionalidad y combatir el problema de sesgo entre experimentos. Por otro lado, utilizamos el valor absoluto del tamaño del efecto del modelo lineal Ridge, también conocida como regularización L2. Esta es una técnica utilizada en ML con la finalidad de prevenir el sobreajuste en un modelo predictivo, especialmente en los casos donde el modelo tiene más características que observaciones y cuando las características están altamente correlacionadas. Con esto conseguimos unos coeficientes que determinan cuanta importancia han tenido en el modelo los predictores, en nuestro caso los transcritos. Por lo que nos quedamos con los predictores con coeficientes más grandes en términos

absolutos, es decir, más importantes. Con estos dos elementos, generamos subconjuntos de 40, 65, 90, 120, 150, 180, 220 y 250 genes o transcritos. Para cada subconjunto, generamos un modelo utilizando umbrales KLD entre 0,06 y 0,36 por incrementos de 0,02 y el paquete R glmnet (v2.0.16) con sus correspondientes algoritmos de regresión lineal [240]. Todo este proceso puede verse resumido en la figura 6.2. Entrenamos un total de 272 modelos lineales con regularización L2. Seleccionamos los mejores basándonos en el error de validación cruzada producido por el algoritmo en el conjunto de datos de entrenamiento Fig. 6.3. Una vez seleccionados los modelos en base al conjunto de entrenamiento (Discovery) se estima su precisión y comportamiento en un conjunto de test (Replication).

Para comprender la biología asociada a los modelos predictivos, realizamos el análisis de las rutas y la biología siguiendo el enfoque descrito anteriormente. Para añadir solidez a estos análisis, cada conjunto de transcritos de los modelos se amplió para incluir transcritos significativamente correlacionados ($p<0,05$ y $r>0,95$) con los transcritos en cada uno de los modelos predictivos.

6.2.7 Evaluación de los factores de riesgo asociados al Alzheimer

La amiloidosis cerebral es el biomarcador de referencia para la AD. Para evaluar si los modelos predictivos estaban correlacionados con la amiloidosis cerebral, junto con otros factores de riesgo de AD conocidos, utilizamos la correlación de Spearman entre el riesgo estimado proporcionado por el clasificador y los niveles medidos en CSF de Aβ42, tau, p-tau. En este análisis sólo incluimos a aquellos individuos con mediciones de CSF disponibles en los siete años anteriores o posteriores a la fecha de extracción (n=72). Además de la correlación, también utilizamos los valores del CSF para clasificar a los participantes utilizando los criterios de la ATN y, a continuación, probamos el desempeño de los modelos para predecir los criterios de la ATN. Probamos el rendimiento de los modelos de transcriptoma cfRNA para diferenciar entre positividad A y positividad AT. No había datos disponibles para los criterios N de estas muestras.

6.2.8 Evaluación de la sensitividad y especificidad

Para evaluar la AD a diferentes estadios, es decir, en su continuum, calculamos el rendimiento de los modelos predictivos en los participantes sintomáticos tempranos con (CDR=0,5) y sintomáticos (CDR=1) (Fig. 6.6A). Escalamos los recuentos de genes al rango de la población de entrenamiento calculando la puntuación z (Zscore) utilizando la media

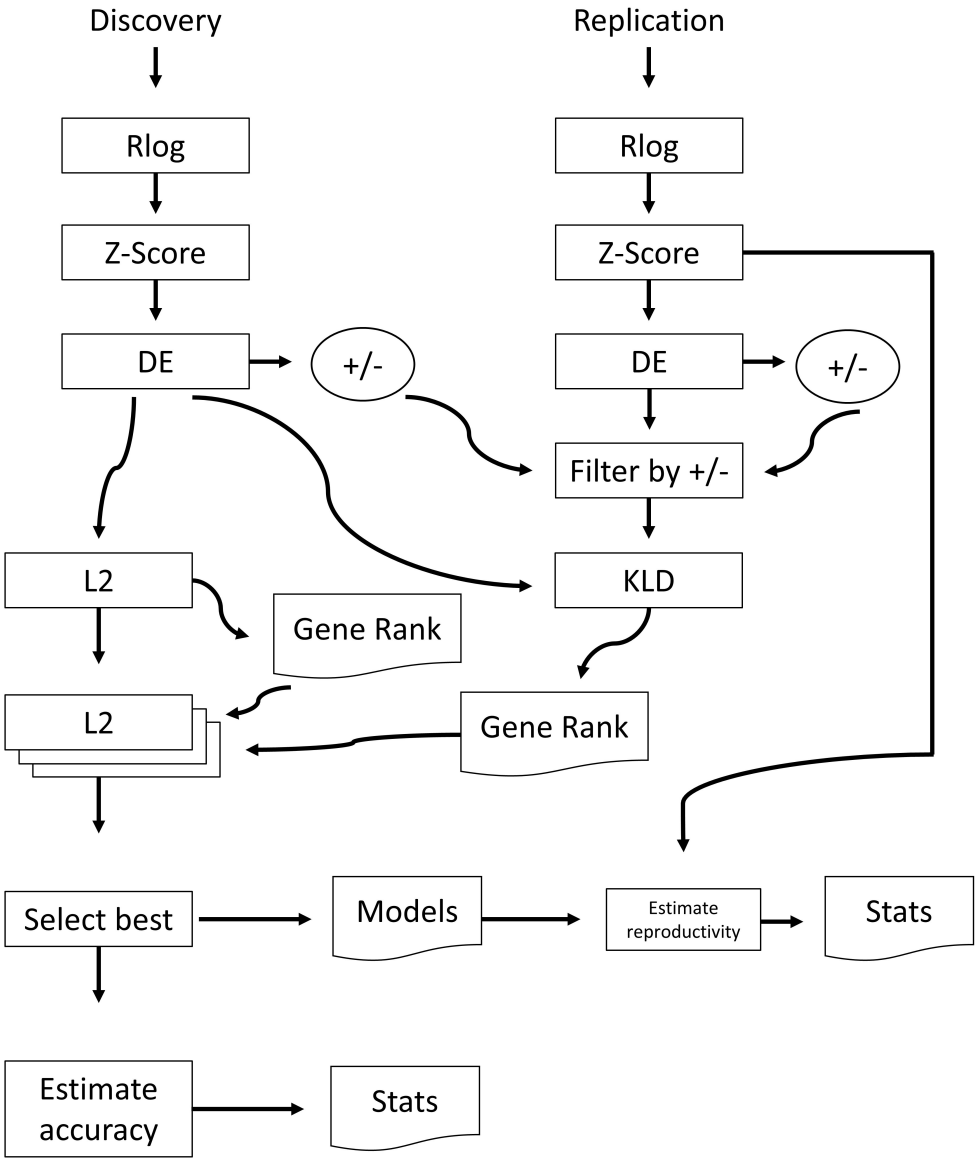


Figura 6.2: Esquema del enfoque utilizado para minimizar el efecto particular de cada experimento de secuenciación de cfRNA aplicado a obtener modelos predictivos.

y la desviación estándar de la población de entrenamiento. A continuación, calculamos la puntuación de riesgo para cada individuo utilizando la fórmula de regularización L2. Las puntuaciones superiores a 0,50 se consideraron casos. Para calcular la curva ROC, se comparó el estado predicho con el verdadero para cada grupo. Debido al solapamiento clínico y patológico entre las enfermedades neurodegenerativas, uno de los retos en

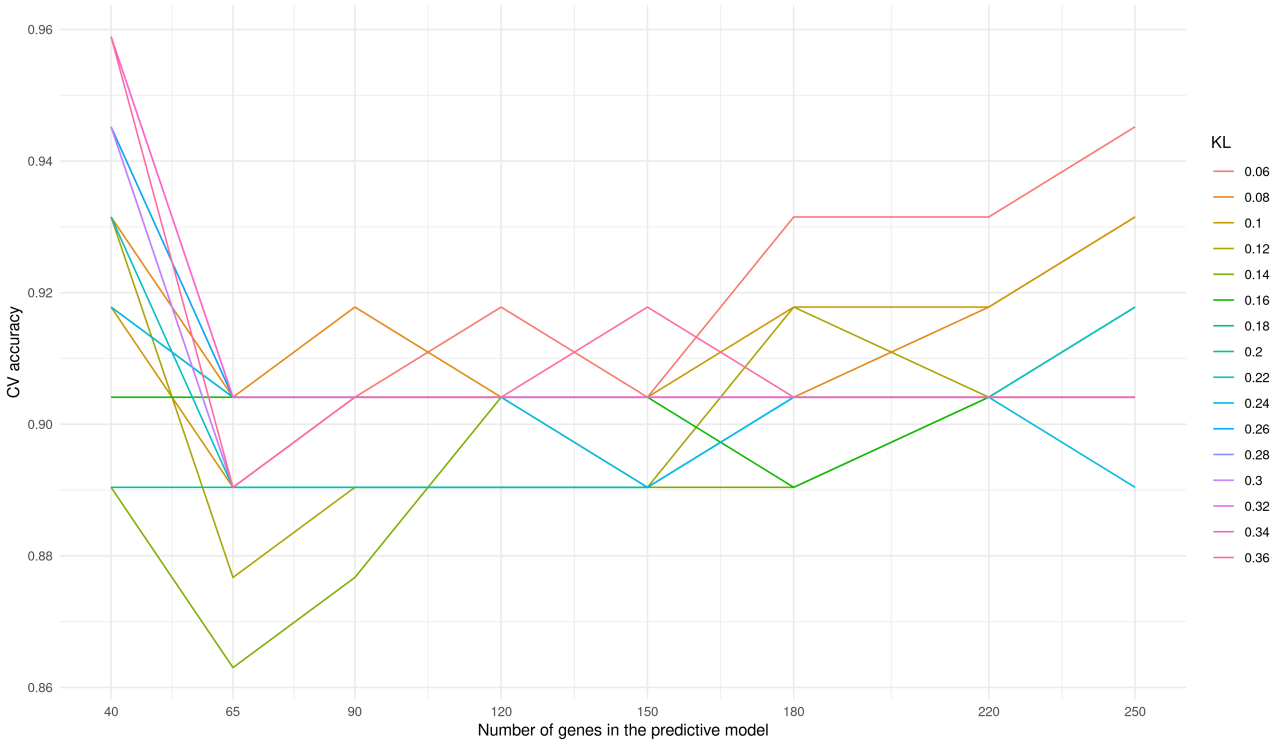


Figura 6.3: Precisión en los experimentos de CV con diferentes valores de KL para seleccionar lo modelos.

el desarrollo de biomarcadores para la neurodegeneración es la especificidad de la enfermedad. Para evaluar el desempeño de los modelos predictivos en el contexto de otras enfermedades neurodegenerativas, calculamos el valor predictivo de riesgo en 96 individuos con PD, 16 con LBD y 17 con FTD (Fig. 6.8A) y calculamos las curvas ROC como se describió anteriormente. Adicionalmente, también calculamos la curva ROC utilizando muestras de AD en lugar de controles como grupo de comparación. Por último, evaluamos si el modelo mostraba alguna mejora al añadir el genotipo APOE al predictor cfRNA. APOE es el factor de riesgo genético más importante. Por lo tanto, para comprender si el predictor captaba el efecto de APOE, incluimos el genotipo APOE en el modelo codificado por dos variables que representaban el número de alelos $\epsilon 2$ y alelos $\epsilon 4$.

6.3 Resultados

6.3.1 Concordancia entre los transcritos desregulados en el cfRNA plasmático y el cerebro de participantes con AD

Analizamos el cfRNA plasmático de participantes presintomáticos con AD para capturar los cambios tempranos causados por la patología de la AD y construir clasificadores que contengan la expresión de un número escalable de genes. Todos los participantes presintomáticos de AD debían tener una muestra antes del inicio de los síntomas (momento de la extracción), y evidencia de depósito de Aβ (LCR Aβ<500ng/L o PET positivo) y/o evidencia de empeoramiento clínico medido por CDR en la última visita clínica en comparación con el momento de la extracción (Fig. 6.1A). Para generar dos conjuntos de datos independientes, llevamos a cabo la selección retrospectiva de muestras en dos veces, dando lugar al dataset de descubrimiento y al de replicación, a partir de una cohorte caracterizada clínicamente y con suficientes datos longitudinales. Para ambos conjuntos de datos, extrajimos y secuenciamos RNA en diferentes fechas para cada conjunto de datos de muestras de plasma de 67 participantes presintomáticos de AD (ndescubrimiento=47; nreplicación=20) y 47 controles (ndescubrimiento=26; nreplicación=22) (Fig. 6.1A-B).

Tras un riguroso control de calidad, realizamos el análisis de expresión diferencial comparando participantes con AD presintomática y controles utilizando DESeq2. Se identificaron 190 transcritos diferencialmente expresados una vez se corregía este análisis con el sexo y la edad en el momento de la extracción (Fig. 6.4). Con la finalidad de ver cuantos de estos transcritos detectados habían sido detectados previamente en otros estudios utilizamos transcritos diferencialmente expresados identificados previamente en plasma de AD sintomático avanzado del único estudio disponible [223] y se replicaron 37 de nuestros hallazgos, que mostraron significación estadística en un análisis de solapamiento (p=0,01). Por otra parte, y lo que es más importante, queríamos saber si el cfRNA captaba potencialmente los cambios que tenían lugar en el cerebro. Utilizando un conjunto de datos publicado [229], descubrimos que 23 de los 190 transcritos estaban diferencialmente expresados tanto en el cerebro como en el plasma de los participantes con AD (Fig. 6.4). El solapamiento fue estadísticamente significativo (p=0,03) con un enriquecimiento de 1,6 veces el esperado. Además, los tamaños del efecto de los 23 genes en el cerebro y el plasma estaban altamente correlacionados (cor=0,83; p=7,55 x 10^{-7}). Finalmente si tomamos los tres estudios en su conjunto, replicamos siete transcritos de 190 tanto en el cfRNA del plasma como en el cerebro (MBOAT2, SLC9A9, RHOBTB3,

RUNX1M, POC1B, SRBD1 e HIPK3). Para investigar más a fondo si los 190 transcritos identificados en este estudio se expresaban en el cerebro, accedimos al portal GTEx y descubrimos que 176 de los 190 genes se expresaban en el tejido de la corteza cerebral, lo que añade pruebas de que el cerebro es una fuente potencial de los transcritos diferencialmente expresados obtenidos en cfRNA plasmático.

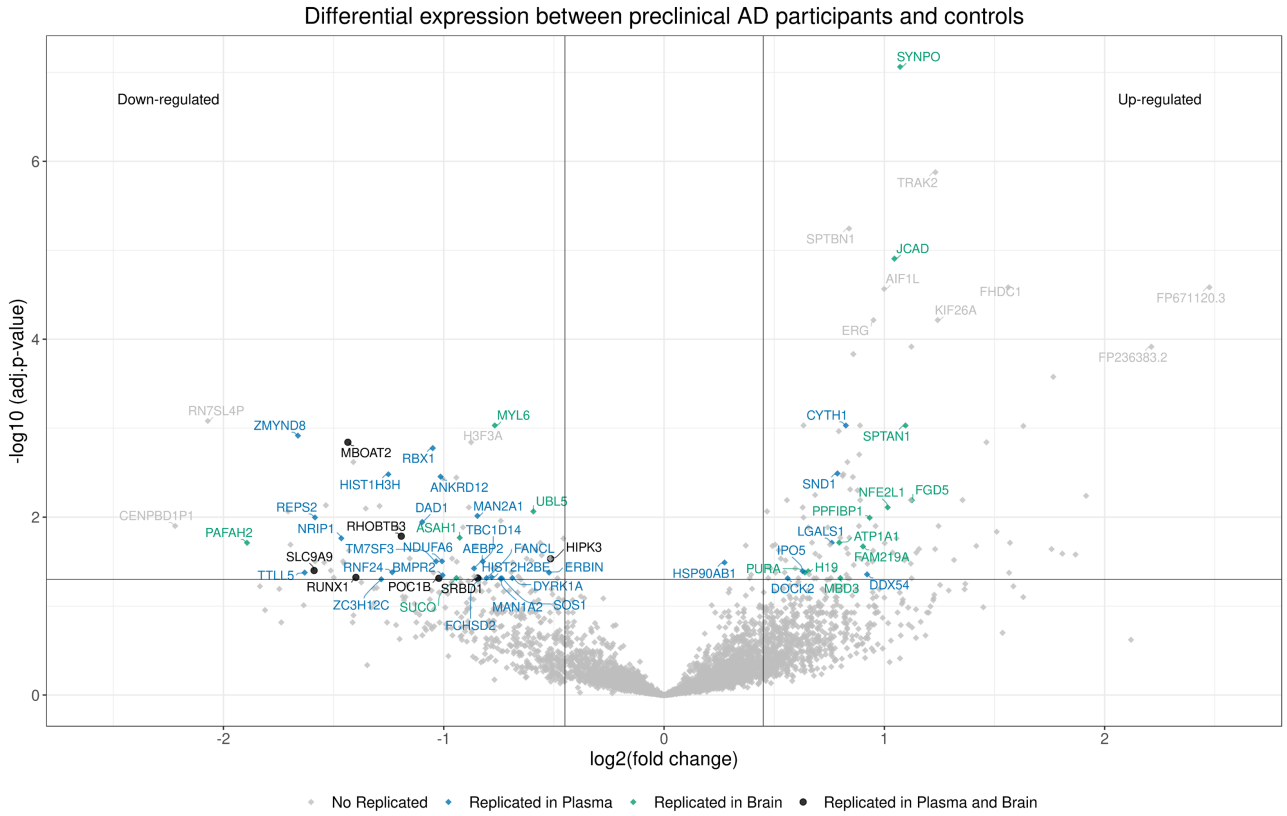


Figura 6.4: Volcano plot que muestra la expresión diferencial de transcritos en pacientes con AD presintomático respecto a controles.

Para evaluar la relevancia biológica potencial de los 190 transcritos, exploramos la Enciclopedia Kyoto de Genes y Genomas (KEGG) y encontramos que los transcritos identificados estaban enriquecidos y se solapaban significativamente con la ruta de AD (un total de nueve genes, p = 8,92 × 10^{-3}). También utilizamos la herramienta ToppFun de ToppGene Suite y descubrimos que los 190 transcritos concordaban con los transcritos regulados al alza en los cerebros de pacientes con AD (p = 1,40 × 10^{-4}). También identificamos un enriquecimiento en los términos de la ontología génica, *Gene Ontology* (GO) en inglés, para el componente celular sinapsis neuronal (p = 6,69 × 10^{-3}) y postsinapsis (p = 1,58 × 10^{-2}). Por último, realizamos un análisis de co-expresión utilizando redes de la corteza frontal de casos de AD de ROSMAP en CoExpWeb [238].

Encontramos un solapamiento estadísticamente significativo entre los 190 transcritos y dos módulos de coexpresión (thistle1 y darkgrey). El módulo thistle1 ($p = 2{,}00 \times 10^{-4}$), se asoció con los oligodendrocitos en la corteza mientras que el módulo darkgrey ($p = 0{,}03$) se asoció con el proceso de vascularización y las células endoteliales-externas. En conjunto, estos resultados sugieren que el cfRNA plasmático podría estar captando los cambios transcripcionales que tienen lugar en el cerebro de los participantes con AD presintomática.

6.3.2 El cfRNA recapitula una firma transcriptómica correspondiente a las etapas presintomáticas de AD

Para aprovechar todos los datos de RNA disponibles, hemos desarrollado un nuevo enfoque que permite el uso de dos experimentos independientes de secuenciación de RNA como descubrimiento y replicación para herramientas de aprendizaje automático (Fig. 6.2). Normalizamos utilizando el método rlog y un zscore [235]. Redujimos la dimensionalidad de los dos conjuntos de datos reteniendo los transcritos que mostraban la misma dirección del efecto en la comparación caso-control. A continuación, calculamos el solapamiento de la distribución de cada transcrito dentro de los dos conjuntos de datos utilizando la divergencia de Kullback-Leibler (KLD). Por último, utilizamos los valores absolutos de KLD para clasificar los transcritos y generar ocho subconjuntos con diversos números de genes. Dentro de cada subconjunto utilizamos umbrales KLD (de 0,06 a 0,36 por incrementos de 0,02) y modelos lineales de regularización L2 (regresión ridge) para predecir la AD presintomática en el conjunto de datos de entrenamiento. A continuación, evaluamos el rendimiento en el conjunto de datos de prueba.

Generamos un total de 272 modelos con distintos números de transcritos y seleccionamos los tres mejores basándonos en los experimentos de validación cruzada (Fig. 6.3). Los mejores modelos contenían 40, 90 y 220 transcritos con un área bajo la curva ROC (AUC-ROC) en el conjunto de datos de prueba de 0,90, 0,92 y 0,94 respectivamente (Fig. 6.1C; Tabla 6.1). Observamos que el solapamiento de los transcritos entre los modelos era significativo ($p < 2.16 \times 10^{-16}$ - Fig. 6.1D), lo que sugiere que la estrategia de estandarización y selección de características implementada aquí tiende a seleccionar buenos predictores de forma consistente. De hecho, los 28 transcritos comunes a los tres modelos tenían un AUC de 0,92. Tras extraer los valores beta, es decir, la importancia de los predictores de cada gen en cada uno de los modelos predictivos (Fig. 6.5), observamos que el gen SYNPO (el gen con mayor diferencia de expresión en el análisis de expresión

Tabla 6.1: Comportamiento de los tres modelos predictivos escogidos en individuos presintomáticos de AD para los conjuntos de datos de entrenamiento y test.

Model	Status	Balanced Accuracy	Cohen's Kappa	Sensitivity	Specificity	AUC
40 transcripts model	Training	0.957	0.885	1.000	0.915	-
	Testing	0.859	0.715	0.818	0.900	0.900 (0.819, 0.981)
	Testing + APOE	0.809	0.618	0.818	0.800	0.934 (0.874, 0.994)
90 transcripts model	Training	0.926	0.803	1.000	0.851	-
	Testing	0.905	0.809	0.909	0.900	0.916 (0.834, 0.998)
	Testing + APOE	0.857	0.714	0.864	0.850	0.950 (0.892, 1.000)
220 transcripts model	Training	0.936	0.830	1.000	0.872	-
	Testing	0.952	0.905	0.955	0.950	0.941 (0.871, 1.000)
	Testing + APOE	0.902	0.808	0.955	0.850	0.975 (0.942, 1.000)

diferencial) era la característica más relevante para los modelos con 90 y 220 genes.

En biomarcadores plasmáticos previamente publicados, la inclusión del genotipo APOE en el modelo mejoró el rendimiento. En nuestro caso, la adición del genotipo APOE no cambió el poder predictivo de los modelos, lo que puede implicar que ya estamos capturando el riesgo asociado con el genotipo APOE (Fig. 6.1C- Tabla 6.1). Además de la capacidad de diferenciar entre participantes presintomáticos de AD y controles, también probamos el rendimiento de los modelos dentro del continuum de AD. Evaluamos la precisión de los tres modelos (con y sin genotipo APOE) en AD sintomática temprana (CDR=0,5, n=42) y sintomática (CDR=1, n=50). En todos los casos, el AUC fue superior a 0,90 (Fig. 6.6B), lo que sugiere que a medida que la AD progresa, sus firmas moleculares cambian pero no drásticamente.

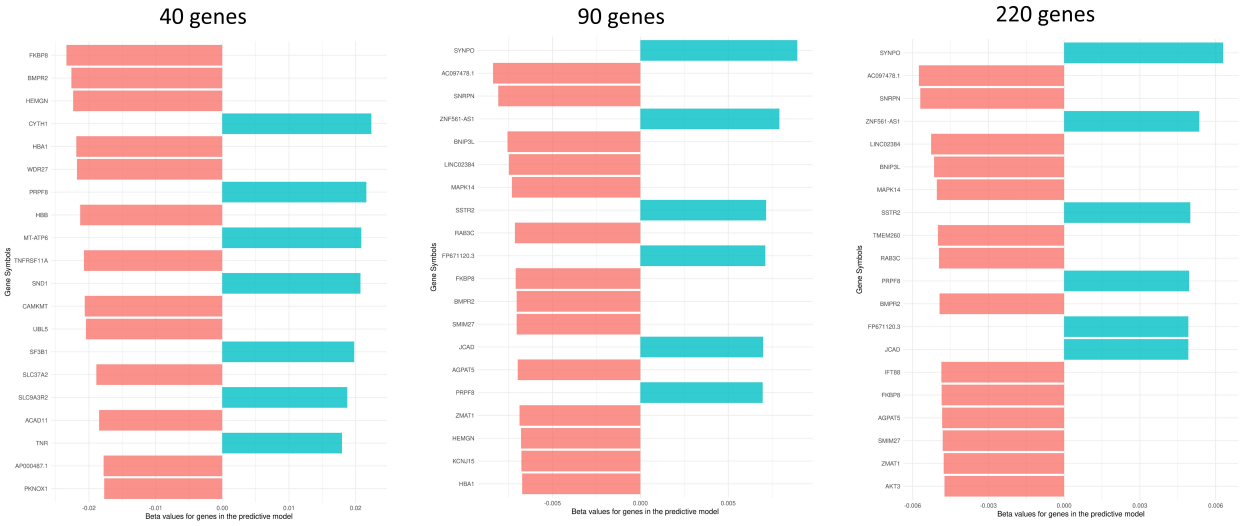


Figura 6.5: Volcano plot que muestra la expresión diferencial de transcritos en pacientes con AD presintómatico respecto a controles.

6.3.3 Los modelos predictivos de cfRNA están enriquecidos en rutas relacionadas con AD en fases tempranas de la patobiología de la enfermedad

Para comprender la relación entre los transcritos incluidos en los modelos predictivos y su posible implicación en la patobiología de AD, realizamos un análisis de enriquecimiento génico para cada uno de los modelos por separado. Dado el número limitado de transcritos incluidos en cada uno de los tres modelos, y con el fin de añadir solidez a los análisis de enriquecimiento, ampliamos cada conjunto de transcritos para incluir transcritos que mostraran una correlación significativa ($p<0,05$ y $r>0,95$) con los transcritos de cada modelo predictivo. Así, los conjuntos aumentaron a 844, 1054 y 2436 transcritos para los modelos predictivos de 40, 90 y 220 transcritos respectivamente. Según esperábamos debido al numero de transcritos originales compartidos estos los conjutos, los conjutos expandidos de transcritos también comparten un número significativo de transcritos ($p<0.05$). Se identificaron 1201, 1111 y 494 términos GO sobrerrepresentados. Los términos relevantes que se sabe que están asociados con AD, como las vías y procesos relacionados con el sistema inmune (ID de términos GO: 0002218, 0002753, 0002757 y

A.

Characteristic	CDR 0.5	CDR 1
N	54	64
Male (N)	27	36
(%)	(50.00%)	(56.25%)
Age at draw median	76.00	76.00
(IQR)	(71.00, 80.52)	(71.00, 82.00)
Age at onset median	72.00	70.00
(IQR)	(65.00, 76.00)	(64.00, 75.00)
Disease duration	4.00	7.00
median (IQR)	(2.73, 6.75)	(5.00, 9.00)

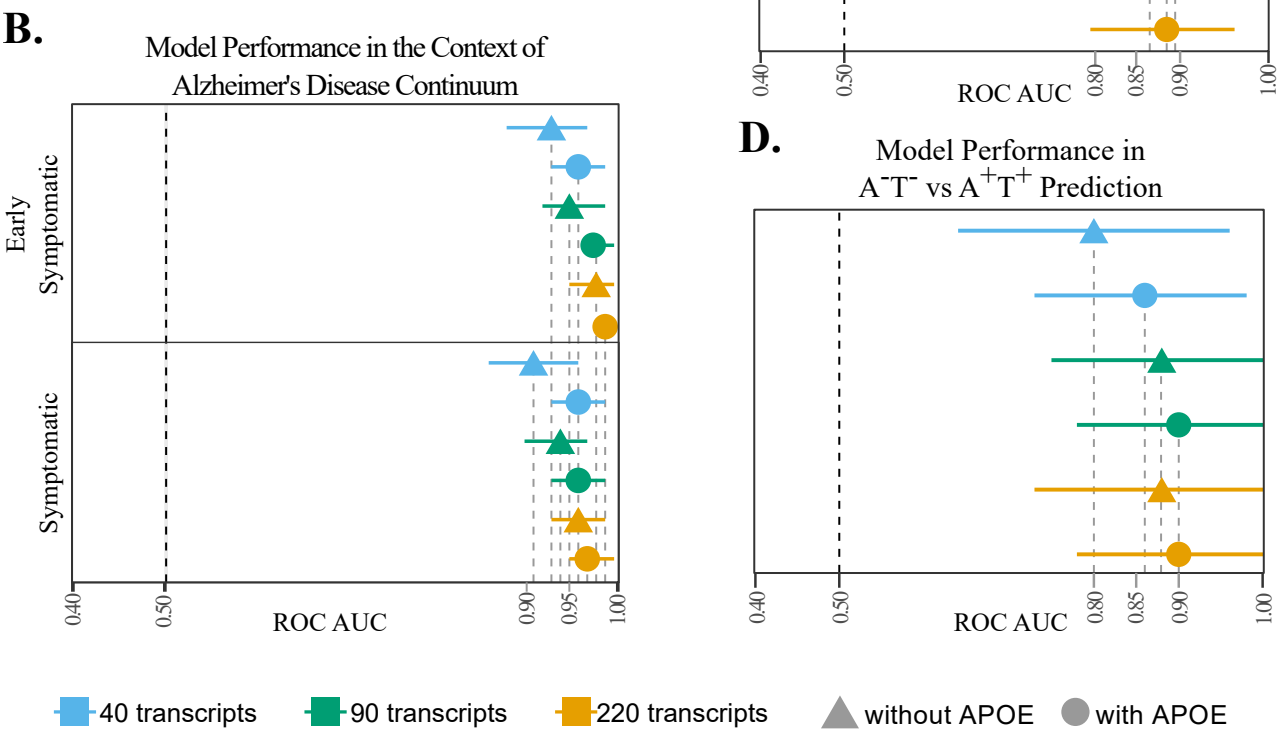


Figura 6.6: Evaluación del modelo en las diferentes etapas de AD y en el contexto del marco ATN.

0002764), o el lisosoma (ID de términos GO: 0005765 y 0005766) fueron significativos en los tres análisis. Identificamos un enriquecimiento significativo en términos relacionados con la regulación de la apoptosis y la muerte neuronal (GO term IDs: 0043523, 0051402, 0070997, 1901214, 1901215 y 1901216) en los tres análisis, lo que apoya la captura de los procesos neuropatológicos tempranos que tienen lugar en el cerebro por los modelos predictivos. Del mismo modo, los análisis de enriquecimiento KEGG identificaron 78, 68 y 40 términos significativamente sobrerrepresentados para cada uno de los conjuntos generados para cada modelo predictivo. Entre otros, las enfermedades neurodegenerati-

vas incluyendo AD y PD fueron significativamente enriquecidas sugiriendo que de hecho estamos capturando procesos relacionados con la biología conocida de la enfermedad neurodegenerativa en una etapa temprana de la enfermedad.

6.3.4 Los modelos predictivos entrenados con participantes presintomáticos de AD pueden predecir con exactitud la positividad amiloide

Los biomarcadores actuales evalúan los niveles de Aβ42 en CSF o plasma para predecir la amiloidosis cerebral. Se investigó si el riesgo estimado de AD calculado utilizando los tres modelos generados aquí (representado por un número de cero a uno) se correlacionaba con los niveles de Aβ42 en CSF. Para aquellos controles (n=43) y participantes presintomáticos de AD (n=28) con mediciones de CSF disponibles en el momento de la extracción de sangre, probamos si el riesgo de AD calculado utilizando los tres modelos se correlacionaba con Aβ42, tau y p-tau en CSF (Fig. 6.7). Encontramos asociaciones significativas con los niveles de Aβ42 en CSF, especialmente para el modelo que contenía 220 transcritos (r^2 = -0,54; p = 1,27 × 10^{-6}), pero no con otros biomarcadores de CSF o factores de riesgo de AD (Fig. 6.7). Las asociaciones con Aβ42 en CSF tuvieron una dirección negativa, como era de esperar. Por último, clasificamos estas muestras siguiendo los criterios de la ATN. De las 72 muestras, 49 eran A-, y 23 A+, mientras que 23 eran T- y 49 T+. Utilizando los tres modelos transcriptómicos, predecimos el estado de positividad A con AUCs de 0,89, 0,88 y 0,86 para los modelos con 220, 90 y 40 transcritos respectivamente (Fig. 6.6C). Al incluir APOE en los modelos, no observamos cambios en las AUC, lo que demuestra que el modelo transcriptómico capta los cambios relacionados con la enfermedad de Alzheimer y su patología principal. También probamos el rendimiento predictivo para A+T+ en comparación con A-T-, aunque el tamaño de la muestra era limitado (n=13 participantes en cada grupo). El modelo con 40 transcritos fue el de peor rendimiento con un AUC de 0,80, mientras que los modelos con 90 y 220 transcritos tuvieron un AUC de 0,88 (Fig. 6.6D). En este caso, la inclusión de APOE mejoró el AUC en todos los casos, 0,86 para el modelo con 40 transcritos y 0,90 para los modelos que incluían 90 y 220 transcritos.

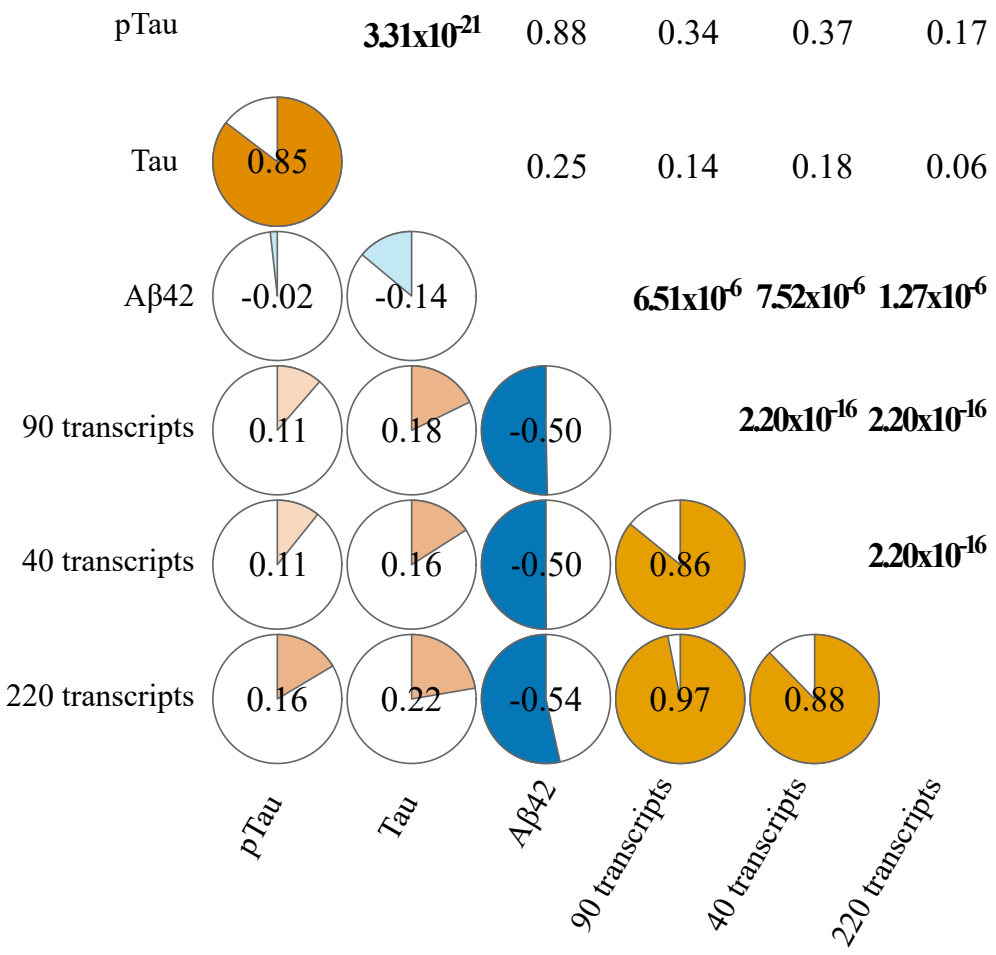


Figura 6.7: Correlación entre los niveles medidos en CSF de biomarcadores de AD con los modelos predictivos de cfRNA.

6.3.5 Los modelos predictivos entrenados con participantes con AD presintomática también pueden predecir la AD en las fases sintomáticas de la enfermedad

Además de la capacidad de diferenciar entre los participantes presintomáticos y los controles, también comprobamos el rendimiento de los modelos dentro del continuo de la AD. Evaluamos la precisión de los tres modelos en AD sintomática temprana (CDR=0,5, n=42) y sintomática (CDR=1, n=50) en comparación con los controles (n=48) (Fig. 6.6A). Para los participantes con ad sintomática temprana, el AUC de los modelos compuestos por 40, 90 y 220 transcritos fue de 0,93, 0,95 y 0,98 respectivamente, mientras que para la AD sintomática el AUC fue de 0,91, 0,94 y 0,96 (Fig. 6.6B). A diferencia de nuestros resultados con el grupo presintomático, la adición del genotipo APOE mejoró la

precisión de los tres modelos en el continuo de AD (Fig. 6.6B). Sin embargo, la mejora no fue estadísticamente significativa, lo que sugiere que estamos capturando parte del efecto del genotipo APOE en el transcriptoma plasmático o que el nivel de precisión del modelo ya es lo bastante alto. En general, el modelo con 40 transcritos mostró menor poder predictivo que los modelos que incluían transcritos adicionales. Nuestros resultados sugieren que, a medida que progresa la AD, las firmas moleculares cambian, lo que repercute en la precisión de los modelos predictivos. Sin embargo, la adición de transcritos a la firma produce un mejor AUC, lo que sugiere que para algunos transcritos, los cambios se acentúan con la progresión de la enfermedad, aumentando la capacidad predictiva de los modelos.

6.3.6 Los modelos predictivos entrenados con participantes presintomáticos de AD tienen una capacidad limitada para predecir otras enfermedades neurodegenerativas

Por último, queríamos evaluar si los modelos eran específicos de la AD. Evaluamos el rendimiento de nuestros modelos en muestras de PD (n=96), LBD (n=17) y FTD (n=16) (Fig. 6.8A). Probamos la especificidad utilizando dos enfoques, en primer lugar, preguntamos si los modelos podían clasificar correctamente cada una de estas enfermedades en comparación con los controles (Fig. 6.8B), y en segundo lugar, si podían distinguir entre estas enfermedades y AD (Fig. 6.8C). Los modelos tuvieron un bajo poder predictivo para diferenciar la PD de los controles (AUC <0.72), mientras que el rendimiento para la FTD y la LBD varió y dependió del número de transcritos en los modelos (0.64 <AUC <0.93), sugiriendo que los modelos son específicos para la AD, pero podríamos estar capturando parte del mismo proceso biológico en enfermedades con alto solapamiento como la LBD. En este caso, la adición del genotipo APOE no disminuyó el poder predictivo, por tanto, no aumentó la especificidad (Fig. 6.8B). De forma similar, al diferenciar entre AD y otras enfermedades neurodegenerativas, los modelos tuvieron un alto poder predictivo para diferenciar AD de PD (0,77 <AUC <0,85) y la FTD (0,75 <AUC <0,86), pero no tanto para la LBD (0,55 <AUC <0,69). En contraste con las secciones anteriores, la adición del genotipo APOE mejoró la diferenciación de la AD de otras enfermedades neurodegenerativas con AUC >0.70 en todos los casos (Fig. 6.8B-C).

El modelo con 220 transcritos pudo diferenciar a los participantes con PD de los con AD con un AUC de 0,81, mientras que la diferenciación de LBD o FTD (AUC <0,76) fue menos precisa. Para los modelos que incluían un número menor de transcritos (90 y

A.

Characteristic	Parkinson's Disease	Lewy Body Dementia	Frontotemporal Dementia
N	96	17	16
Male (N) (%)	61 (63.54%)	11 (64.71%)	11 (68.75%)
Age at draw median (IQR)	72.00 (67.00, 77.00)	79 (74.00, 83.00)	62.50 (59.50, 67.25)
Age at onset median (IQR)	64.00 (60.00, 70.00)	71.00 (69.00, 73.00)	57.50 (54.50, 62.25)
Disease duration median (IQR)	7.00 (4.00, 10.00)	6.00 (3.00, 10.00)	4.50 (3.00, 7.25)

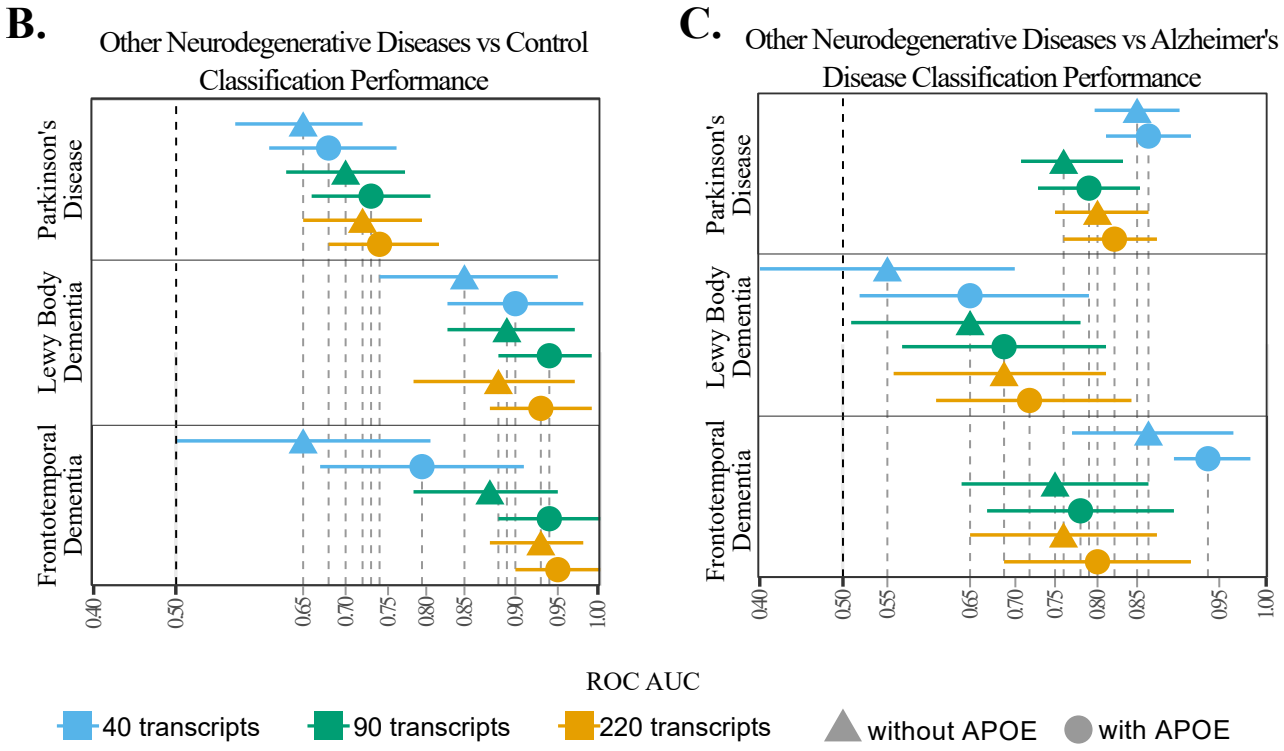


Figura 6.8: Evaluación de la especificidad del modelo como biomarcador para diferenciar enfermedades neurodegenerativas.

40), podían diferenciar AD de PD (AUC >0,81), pero apenas podían hacerlo de la LBD o FTD (AUC <0,69), lo que sugiere que hay varias transcritos que están comúnmente desreguladas en las tres enfermedades y por lo tanto no son tan útiles para la tarea de la diferenciación. De hecho, cuando evaluamos los patrones de expresión en cada enfermedad neurodegenerativa en comparación con los controles de todos los transcritos incluidos en los tres modelos, observamos que los mismos transcritos estaban desregulados en las tres enfermedades, pero en diferentes direcciones en comparación con los controles. La LBD fue la que presentó las diferencias más llamativas en la desregulación de los trans-

critos seleccionados en los modelos predictivos, lo que sugiere que la LBD tiene varios genes desregulados comúnmente con la AD, pero en direcciones y proporciones diferentes. En consecuencia, es posible pensar que la LBD tiene, no sólo más características clínicas compartidas con la AD, sino también más vías moleculares que aquellas compartidas con PD o la FTD, haciéndola la más difícil de diferenciar por los modelos de cfRNA.

6.4 Conclusiones

Este es el primer estudio que utiliza cfRNA plasmático para crear modelos predictivos basados en aprendizaje automático capaces de identificar la AD en las fases presintomáticas en dos conjuntos de datos independientes. Hemos identificado transcritos en plasma que parecen recapitular los cambios que tienen lugar en el cerebro de los participantes en la AD, lo que sugiere que los cambios que tienen lugar en el cerebro se filtran al torrente sanguíneo, muy probablemente debido a la ruptura de la barrera hematoencefálica (BHE) [241, 242]. También hemos construido modelos predictivos que clasifican correctamente a los participantes presintomáticos de AD y controles con un número razonable de transcritos, y que mostraron una alta precisión y especificidad para la AD. Dado el reducido número de transcritos que incluyen los modelos, son potencialmente aplicables al ámbito clínico si nuevas pruebas respaldan su uso beneficioso. Se necesitan estudios con muestras de mayor tamaño para mejorar el rendimiento de los modelos predictivos, sin embargo, aquí demostramos por primera vez que el cfRNA no sólo puede utilizarse como herramienta de predicción precoz, sino que también capta los cambios patológicos tempranos.

Hemos investigado los cambios tempranos en el plasma mediante la cuantificación del cfARN y hemos identificado un solapamiento significativo con los publicados anteriormente [223]. Además, también hemos comprobado que los cambios tempranos en el plasma de cfRNA podrían estar originándose en el cerebro, ya que varios transcritos también están diferencialmente expresados en cerebros de individuos con AD [229, 243]. Además, identificamos que los transcritos de cfRNA desregulados forman parte de módulos de co-expresión ya identificados en la corteza de los casos de AD y enriquecidos en términos GO asociados con el cerebro como sinapsis y postsinapsis. También encontramos una asociación con los oligodendrocitos y el citoesqueleto para los transcritos identificados. La organización del citoesqueleto parece desempeñar un papel clave en la proliferación de oligodendrocitos. Por ejemplo, TRAK2, un transcrito diferencialmente expresado en plasma de participantes presintomáticos en AD, se asocia

con oligodendrocitos participando en la regulación de la organización del citoesqueleto de actina y con el transporte mitocondrial, todos procesos que pueden estar contribuyendo a la AD [244, 245, 246]. Varios estudios sugieren que existe una ruptura temprana de la barrera hematoencefálica debido al inicio de la enfermedad [241, 242, 247], hecho que apoya el origen en el sistema nervioso central de nuestros hallazgos, y los de otros [223].

Hasta donde sabemos, el estudio de Toden et al. [223] fue el único que evaluó el cfRNA plasmático como biomarcador de la AD. Sin embargo, presenta importantes diferencias con el presente estudio que hacen que nuestro diseño de estudio sea más adecuado para construir modelos predictivos. En primer lugar, no incluyó participantes presintomáticos de AD, por lo que su modelo con 1658 transcritos sólo es aplicable a fases clínicas. En segundo lugar, no realizaron análisis de especificidad con otras enfermedades neurodegenerativas, por lo que se desconoce la especificidad del modelo. En tercer lugar, utilizaron todos los transcritos diferencialmente expresados (1658) para construir el modelo, lo que, dado el elevado número de transcritos, dificulta su traslación a un entorno clínico. Además, al utilizar transcritos diferencialmente expresados, el modelo puede ser redundante, estar sobreajustado y, por tanto, no ser generalizable. Aquí hemos adoptado un enfoque más conservador, hemos desarrollado varios modelos predictivos para comprender cómo se comporta el cfRNA en el contexto de la AD. También hemos incluido diferentes enfermedades neurodegenerativas para calcular la especificidad y reconocer el solapamiento entre enfermedades. Por último, hemos estudiado la correlación entre los biomarcadores de la AD en el CSF y los criterios de la ATN para compararlos con las herramientas utilizadas en entornos clínicos reales.

Hasta la fecha, los biomarcadores del CSF han demostrado ser el enfoque más eficaz para clasificar la AD. Las combinaciones de biomarcadores del CSF clasifican la AD clínica y los controles con precisiones que oscilan entre el 0,6 y el 0,95 dependiendo de la edad [248, 249]. El biomarcador de CSF más utilizado y preciso es el cociente Aβ42/Aβ40, que puede diagnosticar correctamente al 82,8 % de los pacientes con AD examinados [189]. Además, las combinaciones de CSF mostraron también una buena precisión en la detección de AD incipiente en participantes con deterioro cognitivo leve [190]. Los biomarcadores plasmáticos actuales muestran una precisión similar a la del CSF. La combinación de Aβ42/Aβ40 plasmáticos, la edad y el estado APOE ϵ4 está altamente correlacionada con la positividad de la PET amiloide y, por lo tanto, podría utilizarse para detectar individuos antes de la punción lumbar, la PET o pruebas adicionales [213]. El modelo cfRNA arroja una precisión similar a la del plasma Aβ42/Aβ40, sin incluir

otras variables. De hecho, demostramos que la adición de APOE no mejora los modelos. La principal ventaja del cfRNA frente a las mediciones proteicas actuales, es su potencial para ser trasladado a una PCR en tiempo real, que es una técnica más rentable que puede ser implementada en todos los entornos clínicos, incluso en aquellos que son remotos. Además, dada la independencia de Aβ, el cfRNA podría utilizarse para la monitorización de la terapia cuando evaluamos fármacos dirigidos a la proteína Aβ.

Los modelos predictivos construidos utilizando enfoques de aprendizaje automático en enfermedades neurodegenerativas tienden a contener un gran número de características [223, 228], o contienen sólo los genes, transcritos o proteínas diferencialmente expresados identificados [250, 251]. En este caso, nos centramos en modelos con un número relativamente bajo de transcritos sin comprometer la precisión para maximizar su potencial de traslación a la clínica. Los modelos anteriores que utilizaban cfRNA informaron de un AUC de 0,83 para la AD clínica; al reducir el número de transcritos, hemos aumentado el AUC hasta 0,94 para la AD presintomática y utilizando sólo 220 transcritos. Para las fases clínicas, el modelo con 220 genes mostró una precisión estable, superando al publicado anteriormente. También demostramos que el modelo predictivo cfRNA es específico para la AD, ya que no es capaz de predecir la PD, o la FTD. Aunque el modelo es capaz de predecir la LBD con una precisión razonable, esperamos cierto grado de solapamiento en la predicción de estas enfermedades debido al solapamiento clínico y patológico existente [252, 253, 254, 255].

Este estudio tiene varias limitaciones. El tamaño de la muestra de AD presintomática es muy limitado. Sin embargo, creemos que hemos accedido a la muestra más amplia de AD esporádica presintomática con datos clínicos retrospectivos. Debido a la estrategia de selección de la muestra, las muestras se han almacenado en el congelador durante largos periodos de tiempo, lo que podría afectar a nuestros hallazgos. Hemos eliminado cualquier transcrito que mostrara una degradación selectiva para minimizar este efecto. El uso de técnicas de ARN-seq es muy sensible a los sesgos, especialmente al uso posterior del aprendizaje automático. Aunque los dos conjuntos de datos son metodológicamente independientes según el campo del aprendizaje automático, proceden del mismo sitio, lo que aumenta el posible efecto de sesgo del presente estudio. No obstante, a efectos de modelización, la generación de dos conjuntos de datos independientes y el uso de enfoques matemáticos han mitigado la presencia de un posible sesgo metodológico. Además, hemos propuesto un nuevo enfoque para integrar conjuntos de datos RNAseq aplicable en estudios que utilizan múltiples conjuntos de datos RNAseq. Por último, se ne-

cesitan muestras de mayor tamaño para todos los estadios de la AD y otras enfermedades neurodegenerativas para confirmar nuestros hallazgos de transcritos diferencialmente expresados y generalizar y mejorar la precisión y especificidad del modelo. No obstante, creemos que este estudio sirve como prueba de que el cfRNA tiene el potencial de detectar cambios relacionados con la patobiología de la AD, incluso antes de la aparición de los síntomas.

En conclusión, el presente estudio sirve como prueba de que el cfRNA tiene el potencial de detectar cambios relacionados con la patobiología de la AD, incluso antes de la aparición de los síntomas. En conclusión, la presente investigación sirve como prueba de que el cfRNA puede tener la capacidad de identificar alteraciones relacionadas con la patobiología de la AD, antes de la manifestación de los síntomas clínicos. Aunque el tamaño de la muestra para la AD presintomática es una limitación del estudio actual, es el mayor y, hasta donde sabemos, el único disponible hasta la fecha basado en la progresión clínica y la información retrospectiva de biomarcadores. Nos ha permitido modelar y replicar en un conjunto de datos independiente un predictor que puede identificar la AD presintomática. Además, el predictor se ha diseñado independientemente del Aβ42, lo que lo convierte en un candidato excelente para monitorizar posibles terapias modificadoras de la enfermedad". El uso del cfRNA plasmático como biomarcador podría ser muy ventajoso debido a su rentabilidad en comparación con las medidas actuales de CSF y plasma y al hecho de que los modelos de cfRNA tienen el potencial de transformarse en paneles de PCR en tiempo real, por lo que el cfRNA podría implementarse sin problemas en la clínica sin equipos ni formación adicionales, también en entornos remotos, algo imposible con las herramientas actuales. En general, creemos que se necesitan más estudios longitudinales con muestras de mayor tamaño para confirmar el uso del cfRNA como biomarcador, pero los resultados actuales muestran un potencial sin precedentes.

CAPÍTULO 7

DISCUSIÓN

En este capítulo vamos a poner en perspectiva los resultados obtenidos a la par que la consecución de objetivos (sección 7.1). Después, vamos a tratar las limitaciones, lo cual es un punto crucial de reconocer, entender y aprender de cara al futuro (sección 7.2). Posteriormente, vamos a resumir las conclusiones del trabajo (sección 7.3). Finalmente, señalaremos posibles vías futuras que surgen de las limitaciones o de posibles avances que se puedan implementar (sección 7.5).

7.1 Revisión, interpretación e implicaciones de los resultados

Hemos enfocado la investigación a la aplicación y desarrollo de herramientas bioinformáticas para tratar diferentes problemas que tienen implicaciones en biología y medicina. El objetivo ha sido ayudar a la comunidad científica a validar relaciones gen-fenotipo, facilitar las decisiones clínicas y mejorar el diagnóstico de enfermedades. Recapitulando, hemos integrado y usado datos médicos, genómicos y transcriptómicos con la finalidad de estudiar las relaciones gen-fenotipo (capítulo 4), de ayudar en el manejo de pacientes con COVID-19 (capítulo 5) y de encontrar biomarcadores para el diagnóstico de enfermedades neurodegenerativas como el Alzheimer (capítulo 6).

En el capítulo 4, presentamos PhenoExam, una innovadora herramienta bioinfor-

mática diseñada para utilizar la información de bases de datos de fenotipo, facilitando así el análisis y la comparación de las relaciones gen-fenotipo a través de listas de genes. Esta herramienta identifica de manera eficiente los términos fenotípicos enriquecidos en un conjunto específico de genes. Adicionalmente, permite la comparación entre dos conjuntos de genes, proporcionando métricas de similitud, términos fenotípicos compartidos y términos diferenciales, es decir, los que caracterizan a uno de los conjuntos de genes de manera única.

PhenoExam ha demostrado su fiabilidad y precisión tanto en simulaciones como en análisis empíricos, identificando con éxito diferencias sutiles en términos fenotípicos entre diferentes conjuntos de genes. Hemos identificado el número de genes requerido en cada base de datos para asegurar un análisis confiable y minimizar la posibilidad de errores que puedan dar lugar a interpretaciones erróneas de los resultados.

Como ejemplo de la utilidad de PhenoExam, hemos detectado los fenotipos significativos y diferenciales de varios conjuntos de genes, como se ilustra en los paneles de PD y EOD y en los conjuntos de genes de la epilepsia. En el caso de PD y EOD, se observó que a pesar de las similitudes a nivel fenotípico, existen diferencias fenotípicas potencialmente relevantes. Esto puede ayudar a diferenciar entre estas enfermedades y ayudar a asociar nuevos genes que se vinculen con algunos de los fenotipos que las hacen diferentes.

Uno de los aspectos más destacados de PhenoExam, y que lo distingue de otras herramientas, es su capacidad para realizar lo que hemos definido como análisis condicionados. Un ejemplo es el de los genes del panel de epilepsia con genes predichos mediante inteligencia artificial que podrían estar asociados a la epilepsia. Los resultados mostraron términos de fenotipo comunes con el conjunto de referencia de la epilepsia, lo cual es valioso como primer paso para la validación de los genes de enfermedad predichos de forma computacional.

En resumen, PhenoExam es una herramienta eficaz para explorar las conexiones entre los términos fenotípicos a través de la integración de diversas bases de datos de anotación. Cumplimos con uno de nuestros objetivos principales al desarrollar el método propuesto como una herramienta accesible tanto en el lenguaje de programación R como a través de una interfaz web.

La relación entre los genes y el fenotipo es uno de los principales objetivos de las investigaciones en biología y medicina con un enfoque en genética [59, 60, 61, 62,

63, 64]. Un recurso interesante en este contexto es la existencia de paneles de genes [256, 257, 258], que se utilizan como referencia para ayudar a diagnosticar ciertas enfermedades. En definitiva, una enfermedad puede asociarse a un fenotipo específico y a genes concretos [65, 66, 67].

Dada la importancia de este asunto, existen herramientas para realizar un análisis de fenotipo sobre un conjunto de genes como HPOsim [85]. También hay herramientas para realizar estudios sobre dos conjuntos de genes como modPhEA [87]. PhenoExam también realiza esto, permitiendo diferentes métodos estadísticos y un amplio rango de bases de datos para el análisis. Además, es la única herramienta que conocemos hasta la fecha que puede realizar un análisis condicionado, como el mencionado entre los conjuntos de epilepsia, ya que fue creada específicamente para esta necesidad.

En la actualidad, los experimentos derivados de los progresos en genómica, transcriptómica, o las predicciones bioinformáticas empleando inteligencia artificial, producen conjuntos de genes que podrían ser una fuente interesante de comparar con los conjuntos de referencia para el fenotipo o la enfermedad que se está investigando. PhenoExam puede realizar este proceso de forma eficaz y simple, convirtiéndose en un filtro inicial para las investigaciones subsiguientes que busquen confirmar la asociación entre genes y fenotipos, ahorrando tiempo y recursos en la investigación.

Por ejemplo, en el análisis de los genes predichos de epilepsia, recuperamos fenotipos asociados a la epilepsia de manera estadísticamente significativa. Estas relaciones entre genes y fenotipos provienen de genes específicos dentro de ese conjunto. Con los resultados obtenidos, sería interesante profundizar en la investigación de todos los genes de ese conjunto como posibles asociaciones a dichos fenotipos en el laboratorio o en las clínicas de diagnóstico genético.

En conclusión, los resultados obtenidos con PhenoExam tienen implicaciones para la mejora de los paneles de genes, para ayudar a distinguir enfermedades o nuevas asociaciones gen-fenotipo, y para enfocar el estudio detallado de los nuevos genes predichos en relación con el fenotipo buscado. Todas estas áreas de acción pueden contribuir a facilitar el diagnóstico de enfermedades, que es uno de los objetivos de esta tesis.

En el capítulo 5, aprovechando la disponibilidad de datos de la pandemia de COVID-19, nos propusimos facilitar la gestión de pacientes mediante técnicas de análisis de datos médicos inteligente y ML. Para ello contamos con datos de más de 86000 pacientes facilitados por el servicio murciano de salud. Inicialmente, analizamos las características

de los pacientes que fallecían, sobrevivían, eran hospitalizados o no. Empleamos métodos estadísticos para confirmar la asociación del sexo, la edad y ciertas comorbilidades con los distintos sucesos estudiados. Por ejemplo, determinamos que una mayor edad conlleva un mayor riesgo de muerte, que los hombres están en mayor riesgo que las mujeres, y que algunas comorbilidades como la insuficiencia renal o la enfermedad cardíaca también están asociadas a un mayor riesgo. Estos descubrimientos concuerdan con los de otros investigadores [156, 157, 158, 159, 160].

Posteriormente, utilizando el algoritmo IPIP que diseñamos para tratar el desequilibrio de los datos, generamos modelos predictivos para clasificar a los pacientes de COVID-19. El objetivo era hacerlo solo con los datos que los médicos tienen en el momento del diagnóstico, como la edad, el sexo y el historial de comorbilidades del paciente. De esta manera, podemos clasificar al paciente desde el inicio y contribuir en el triaje de los pacientes durante períodos de alta demanda hospitalaria. Por un lado, el modelo que clasifica la probabilidad de que un paciente sobreviva o fallezca demostró una exactitud del 92%. Por otro lado, el modelo que clasifica la probabilidad de hospitalización de un paciente mostró una exactitud del 72%. Obteniendo unos resultados superiores o iguales a los de otros modelos que necesitan de datos más complejos y que, por tanto, no se podían aplicar en el momento del diagnóstico [149, 151, 152]. Estos modelos están disponibles en GitHub[1] y también en una web mediante shiny app[2]. En resumen, esta investigación es un ejemplo de aplicación y desarrollo de técnicas de análisis inteligente de datos, utilizando estadística y ML, para extraer información de datos médicos. Además, también hemos trabajado en el objetivo propuesto de facilitar el uso de los modelos haciéndolos disponibles en GitHub y creando una web.

En el capítulo 6, nos propusimos estudiar los cfRNA como biomarcadores para el diagnóstico de Alzheimer en fases tempranas. Hemos desarrollado modelos predictivos con la información transcriptómica del cfRNA en plasma para el diagnóstico de Alzheimer en fases tempranas. Obtuvimos una precisión del 95% en la tarea de diferenciar un paciente con Alzheimer presintomático de un individuo sano utilizando el modelo que recogía la información de 220 transcritos. Además, hemos estudiado el comportamiento de los modelos en los diferentes estadios de la enfermedad y hemos analizado su especificidad comparando su desempeño en pacientes de otras enfermedades neurodegenerativas.

Para comprender la relevancia de los resultados es necesario poner en perspectiva

[1] `https://github.com/antoniogt/ipip`
[2] `https://alejandrocisterna.shinyapps.io/PROVIA/`

la dificultad de diagnosticar Alzheimer. El diagnostico definitivo solo se puede realizar analizando tejido cerebral una vez que el paciente fallece [259]. Los primeros signos de la enfermedad de Alzheimer pueden ser sutiles y difíciles de distinguir de los cambios normales relacionados con la edad. Además, los síntomas que aparecen suele ser variables lo que dificulta el diagnóstico. Para realizar un diagnóstico de posible Alzheimer los clínicos realizan pruebas de imagen, analizan proteínas en el CSF y realizan cuestionarios al paciente para evaluar su nivel cognitivo. Una vez se realiza y se revisan los resultados por expertos el diagnóstico acaba siendo preciso en un 77% de los casos [260] y estos son datos para pacientes que ya presentan síntomas de deterioro cognitivo (CDR >0.5). En este trabajo, utilizando la información transcriptómica del cfRNA del plasma somos capaces de distinguir con una alta precisión (95%) a individuos que no presentan síntomas visibles pero que acabaran desarrollando Alzheimer (individuos presintomáticos). Es decir, somos capaces de determinar que una persona aparentemente sana, que no muestra síntomas de deterioro cognitivo (CDR=0), va a sufrir Alzheimer hasta siete años antes de que aparezcan los síntomas visibles. Además, estos resultados se obtienen realizando análisis del plasma, mucho menos invasivo que la extracción del CSF.

El diagnóstico temprano y preciso de Alzheimer puede ser muy importante por numerosos motivos. Una de las más interesante es beneficiarse antes de que aparezcan los síntomas de deterioro cognitivo de los beneficios de los posibles nuevos tratamientos [261]. Hasta la fecha solo hay un fármaco aprobado concreto indicado para el tratamiento del Alzheimer, el Aducanumab. El Aducanumab es un anticuerpo monoclonal dirigido hacia Aβ que ha obtenido resultados prometedores aunque también ha levantado controversia por la mínima mejora que produce [262, 263]. La existencia de un método de detección de Alzheimer temprano puede suponer un mayor efecto de este o futuros tratamientos reduciendo la velocidad en el deterioro cognitivo de los pacientes. Más allá de los tratamientos dirigidos, el tener un diagnóstico temprano puede posibilitar cambios en el estilo de vida y un manejo mejor de la enfermedad por el paciente, los médicos y los familiares. En definitiva, mejorar la calidad de vida del paciente, alargar su esperanza de vida y ahorrar pruebas, malestar y costes.

Actualmente otros biomarcadores o combinaciones de ellos como tau, Aβ40 o Aβ42 han demostrado una precisión entre el 70-95% para detectar individuos con deterioro cognitivo y Alzheimer, es decir, pacientes a partir de un CDR de medio punto [264]. Lo cual difiere de lo que estudiamos nosotros con pacientes asintomáticos con CDR igual a cero. Centrándonos en el paso concreto desde un individuo presintomático hacia

el deterioro cognitivo y al Alzheimer tenemos un estudio interesante para poner en contexto nuestros resultados [265]. En este estudio comenzaron con 720 participantes de los cuales el 22% tenía biomarcadores positivos de AD al inicio del estudio. El 34% por ciento de aquellos con biomarcadores positivos desarrollaron deterioro cognitivo después de un seguimiento medio de 4 años, en comparación con solo el 8% de aquellos con biomarcadores negativos. Entre aquellos con progresión clínica de demencia, se asignó un diagnóstico clínico de AD en el 80% de aquellos con biomarcadores positivos. En cuanto a método, lo más similar son los modelos propuestos por Toden et al [223], los cuales también utilizan cfRNA, aunque obteniendo menos AUC (entorno al 0.83), centrandose en individuos sintomáticos (CDR =>1) y utilizando demasiados transcritos como para llevar esos biomarcadores a la clínica.

En definitiva, en este trabajo se han analizado datos utilizando y desarrollando métodos estadísticos y técnicas de Machine Learning (ML). Por otro, se han utilizado estos nuevos métodos para ayudar al diagnóstico utilizando información clínica, genómica y transcriptómica. Hemos tratado desde enfermedades concretas como el Alzheimer o el COVID-19 hasta problemas más básicos, como el descubrimiento o análisis de las relaciones gen-fenotipo. Siempre tratando de cumplir los objetivos planteados, facilitando la difusión y el uso de las herramientas y técnicas propuestas.

7.2 Limitaciones

La investigación desarrollada durante esta tesis no está exenta de limitaciones. Toda investigación está sujeta a limitaciones de diferente tipo: en la metodología, en el diseño del estudio, en la muestra, o en la interpretación de los resultados. En los siguientes párrafos vamos a tratar las limitaciones generales y específicas de cada parte de la investigación desarrollada.

En el capítulo 4 presentamos PhenoExam una herramienta bioinformática para el estudio de las relaciones gen-fenotipo. Esta herramienta tiene algunas limitaciones que deben considerarse:

- La metodología empleada por PhenoExam para determinar estadísticamente la semejanza entre un conjunto de genes y otro, en base a sus términos fenotípicos, se fundamenta principalmente en pruebas de aleatorización. De este modo, presuponemos que al generar un gran número de pruebas, estamos examinando variados

conjuntos de genes. Este proceso podría derivar en errores si se solicitan pocas pruebas para determinar el nivel de significancia estadística. Por ello, sugerimos que el usuario genere al menos 1000 pruebas de aleatorización. Además, este método presupone que la aleatorización conduce a la formación de conjuntos de genes que pueden poseer fenotipos semejantes a los del conjunto que se desea comparar. No obstante, esto no es equivalente a la obtención de un conjunto de genes a través de un experimento ya orientado a un resultado específico. En resumen, recomendamos efectuar comparaciones tanto de manera aleatoria como con conjuntos de genes similares para poner en contexto las similitudes o diferencias.

- PhenoExam requiere listas de genes como datos de entrada. Es responsabilidad del usuario garantizar la calidad de estos datos. Si los conjuntos de genes están incompletos o contienen errores, los resultados podrían ser inexactos. Aunque PhenoExam proporciona algunos mensajes de información, es prácticamente seguro que no se han tenido en cuenta todas las posibles excepciones.
- PhenoExam integra numerosas bases de datos para realizar su análisis. Si estas bases de datos no están actualizadas o contienen información errónea, esto podría afectar la precisión de los resultados de PhenoExam. PhenoExam necesitará actualizaciones periódicas para incorporar estos avances en su marco de trabajo.
- Aunque PhenoExam puede proporcionar información valiosa, la interpretación de los resultados requiere de un conocimiento experto. Los resultados deben ser interpretados cuidadosamente y en el contexto de la literatura existente y del objetivo del análisis.
- Por el diseño de la herramienta PhenoExam no unifica la nomenclatura. Sirve para consultar las bases de datos pero no para unificarlas.
- PhenoExam no trabaja con datos de transcriptómica, lo cual puede ser interesante cuando se está estudiando la expresión de los genes.

En el capítulos 5 estudiamos los tipos de pacientes de COVID-19 y desarrollamos modelos predictivos para predecir su estado final. Los resultados y modelos de esta investigación tiene algunas limitaciones que deben considerarse:

- Los datos de los pacientes que utilizamos provienen exclusivamente de la Región de Murcia. Esto podría generar limitaciones a la hora de generalizar los resultados

a otras comunidades o a datos recopilados por otros servicios de salud. Sin embargo, hemos observado que muchos de nuestros descubrimientos están en sintonía con los de otros investigadores. Además, hemos sugerido métodos para manejar tanto el desbalanceo de datos como el manejo de datos sencillos, con el fin de promover su uso y generalización.

- Los datos con los que trabajamos fueron obtenidos antes de la implementación de la vacunación masiva, o en un contexto con una tasa de vacunación aún muy baja. Esto se debe a la fecha en que se realizó el estudio. No obstante, actualmente estamos trabajando con datos más recientes que sí incluyen el efecto de las vacunas.

- Al evaluar conjuntos de datos desbalanceados, que posiblemente se asemejen más a la naturaleza de los datos a los que los modelos se encontrarán en una situación real, los resultados para predecir la clase positiva (en un caso, la muerte y en otro, el ingreso hospitalario) demuestran que estos escenarios son los más difíciles de identificar con precisión. Es decir, cuando se pronostica uno de estos eventos no es tan seguro que vaya a ocurrir. Por tanto, los usuarios deben tomar con cautela estas predicciones y usarlas de forma conveniente.

En el capítulo 6, determinamos los transcritos más relacionados con el Alzheimer temprano y desarrollamos modelos predictivos con la información proporcionada por el cfRNA para el diagnóstico de Alzheimer en fases tempranas. Los resultados obtenidos y modelos propuestos no están exentos de limitaciones que deben considerarse:

- El número de muestras con el que contamos. Pese a ser uno de los mayores conjuntos de datos con individuos presintomáticos disponibles, el número es todavía bajo para analizar con una mayor potencia y confianza estadística.

- El centro de procedencia de las muestras. En nuestro caso todas provienen del mismo centro para el Alzheimer, esto puede influir introduciendo sesgos propios de esa población o centro de toma de las muestras. Sería necesario en el futuro una validación con datos de otros centros.

- La dificultad para determinar el efecto de la etnia. Debido al número reducido de muestras que tenemos no se puede considerar estudiar el efecto de la etnia en la información transcriptómica.

- El tiempo que las muestras tienen que estar en el congelador. Por el diseño del estudio tratamos con muestras que pueden llevar hasta 20 años en un congelador. Hemos intentado limitar este problema tomando el tiempo en el congelador como una covariable e identificar los transcritos asociados a una posible degradación.
- La dificultad de compararse en métodos o resultados con otras investigaciones.

7.3 Conclusiones

A continuación vamos a detallar las conclusiones finales que se pueden extraer de los resultados y discusión de cada investigación:

- Hemos desarrollado y validado una herramienta (PhenoExam) capaz de realizar análisis de fenotipo en y entre conjuntos de genes.
- Hemos puesto a disposición esta herramienta en R desde el GitHub `https://github.com/alexcis95/PhenoExam` y una interfaz web `https://alejandrocisternа.shinyapps.io/phenoexamweb/`. La herramienta tiene su propio tutorial de uso en `https://raw.githack.com/alexcis95/PhenoExamWebTutorials/main/tutorial.html`.
- Con PhenoExam hemos demostrado que los genes predichos de epilepsia muestran fenotipos similares a los del panel de epilepsia de referencia. En definitiva, se han propuesto nuevos genes relacionados con la epilepsia.
- Hemos encontrado asociación estadística entre la edad, el sexo y las comorbilidades de los pacientes de COVID-19 con su estado final. Esto ha permitido describir el perfil de los pacientes más comunes respecto a su estado.
- Hemos aplicado y desarrollado un algoritmo (IPIP) capaz de combatir el desbalanceo de los datos, tan usual en datos clínicos.
- Utilizando la edad, el sexo y las comorbilidades previas de pacientes COVID-19 hemos creado modelos predictivos capaces de clasificar con gran exactitud el estado final del paciente (sobrevive o fallece) y la necesidad de ingreso (hospitalario o externos).
- Hemos puesto a disposición de la comunidad en una web los modelos de clasificación de los pacientes COVID-19 que hemos generado.

- Hemos detectado un total de 190 transcritos diferencialmente expresados entre individuos con Alzheimer presintomáticos e individuos sanos de edad similar.
- Hemos demostrado que estos transcritos detectados en plasma y asociados al Alzheimer temprano tienen similitudes y comparten funciones biológicas a los de estudios realizados en cerebro.
- Hemos desarrollado modelos que predicen el desarrollo de Alzheimer en base a la información transcriptómica procedente del cfRNA. Estos modelos muestran una precisión de entre el 85 y el 95 % para predecir que un individuo sano asintomático desarrollará Alzheimer. Además, son modelos sencillos, con un número limitados de transcritos con opciones de ser trasladados a la práctica clínica.
- Hemos comprobado que los modelos de cfRNA para Alzheimer muestran una alta especificidad comprobando con enfermedades como el Parkinson, la demencia frontotemporal y la demencia por cuerpos de Lewy.

7.4 Conclusions

Next, we will detail the final conclusions that can be drawn from the results and discussion of each investigation:

- We have developed and validated a tool (PhenoExam) capable of performing phenotype analysis between sets of genes.
- We have made this tool available in R from GitHub `https://github.com/alexcis95/PhenoExam` and a web interface `https://alejandrocisterna.shinyapps.io/phenoexamweb/`. The tool has its own tutorial at `https://raw.githack.com/alexcis95/PhenoExamWebTutorials/main/tutorial.html`
- Using PhenoExam we have demonstrated that the predicted epilepsy genes show phenotypes similar to those of the reference epilepsy panel. In short, new genes related to epilepsy have been proposed.
- We have found a statistical association between age, sex, and comorbidities of COVID-19 patients and their final condition. This has allowed us to describe the profile of the most common patients in relation to their final condition.
- We have developed an algorithm (IPIP) capable of dealing with data imbalance.

- Using the age, sex, and pre-existing comorbidities of COVID-19 patients, we have developed predictive models capable of accurately classifying the final status of the patient (survives or dies) and the necessity of hospital admission.
- We have made the COVID-19 patient classification models that we have generated available to the community via a website.
- We have detected a total of 190 differentially expressed transcripts between presymptomatic Alzheimer's individuals and healthy individuals of a similar age.
- We have demonstrated that these transcripts, detected in plasma and associated with early Alzheimer's, have similarities and share biological functions with those in brain studies.
- We have developed models that predict the development of Alzheimer's based on transcriptomic information from cfRNA. These models show an accuracy between 85 % and 95 % to predict that a healthy asymptomatic individual will develop Alzheimer's. Additionally, these are simple models, with a limited number of transcripts, with the potential to be transferred to clinical practice.
- We have verified that the cfRNA models for Alzheimer's show high specificity when tested against diseases like Parkinson's, frontotemporal dementia, and dementia with Lewy bodies.

7.5 Posibles vías futuras de la investigación

En esta sección vamos a tratar las posibles mejoras o vías futuras de la investigación presentada.

En cuanto al capítulo 4 en el que presentamos PhenoExam, podemos señalar las siguientes posibles mejoras y vías de trabajo futuras:

- Posibilitar los análisis con datos transcriptómicos directamente.
- Integrar más bases de datos relacionadas con los fenotipos.
- Utilizar o desarrollar métricas de similitud entre fenotipos de distintas bases de datos.
- Actualizar periódicamente las bases de datos a las que unifica el acceso PhenoExam.

- Continuar investigando a través de datos de la empresa NIMGenetics la idoneidad de los genes de epilepsia propuestos.

Refiriéndonos al capítulo 5 en el que presentamos los modelos predictivos del estado final de pacientes con COVID-19, podemos señalar las siguientes posibles mejoras y vías de trabajo futuras:

- Sería conveniente probar y ajustar los modelos de predicción elaborados para el manejo de pacientes COVID-19 con datos procedentes de otros centros sanitarios distintos a los de la Región de Murcia.
- Estudiar el efecto de la vacunación en el resto de variables estudiadas. Generar nuevos modelos con datos de pacientes que incluyan la información de vacunación.
- Diseñar estrategias para centrarnos en aumentar el PPV de los modelos. Puede ser interesante segmentar por franjas de edad para diseñar más de un modelo prestando atención en subir el PPV.

Refiriéndonos al capítulo 6 en el cual desarrollamos modelos predictivos con la información proporcionada por el cfRNA para el diagnóstico de Alzheimer en fases tempranas, podemos señalar las siguientes posibles mejoras y vías de trabajo futuro:

- Validar los hallazgos utilizando datos de pacientes procedentes de otros centros.
- Mejorar y precisar los modelos utilizando más datos.
- Trabajar en llevar los niveles detectados en el RNAseq a otras técnicas más sencillas y baratas como una reacción en cadena de la polimerasa.
- Determinar si la etnia u otros factores no analizados pueden influenciar en los resultados de los experimentos.

CAPÍTULO 8

CONTRIBUCIONES DE LA TÉSIS

En este capítulo se resumen las contribuciones a la comunidad científica y los proyectos relacionados en los que se ha trabajado a lo largo del desarrollo de la tesis. Este capítulo sirve como resumen de las acciones realizadas para cumplir el objetivo de difusión de la investigación propuesto en el capítulo 1.

8.1 Publicaciones y difusión de resultados

8.1.1 Publicaciones como primer autor

En esta sección resumiremos los dos trabajos científicos publicados como primer autor:

- **Cisterna, A.**, González-Vidal, A., Ruiz, D. et al. **PhenoExam: gene set analyses through integration of different phenotype databases**. **BMC Bioinformatics** 23, 567 (2022). `https://doi.org/10.1186/s12859-022-05122-x`.

  Este es el artículo publicado con la herramienta de análisis y comparación de fenotipos en conjuntos de genes descrita en el capítulo 4. Según los datos de SJR (ranking de Scimago) la revista es Q1 para los topics de applied mathematics, biochemistry y computer science applications. Además, es Q2 para los topics de molecular biology y structural biology. La revista tiene un factor de impacto (IF) de 3,3.

- **Cisterna-García, A**., Guillén-Teruel, A., Caracena, M. et al. **A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study**. **Scientific Rep. Nature** 12, 18126 (2022). https://doi.org/10.1038/s41598-022-22547-9.

  Este es el artículo donde se presenta la investigación realizada para clasificar si el paciente de COVID-19 fallece o sobrevive o si será necesaria su hospitalización. Los resultados han sido descritos en el capítulo 5. Según los datos de SJR la revista es Q1 para el topic multidisciplinary. La revista tiene un factor de impacto de 4,99.

A continuación se especifican otros dos trabajos como primer autor que están en proceso de revisión:

- Cisterna-García A., Beric A., Ali M. et al. **Plasma Cell-Free RNA Signatures Predict Alzheimer's Disease**. Under revision (**Science Advances**).

  Este es el artículo donde se presenta la investigación realizada durante la estancia doctoral en la Washington University en St. Louis, Missouri (Estados Unidos de América). En este trabajo se utiliza RNA libre en plasma de pacientes con Alzheimer y de sujetos sanos para generar modelos que predigan el desarrollo de la enfermedad en etapas tempranas. Los resultados han sido descritos en el capítulo 6.

- **Cisterna-García A.**, Bustos B., Bandres-Ciga S. et al. **Genome-wide epistasis analysis in Parkinson's disease between populations with different genetic ancestry reveals significant variant-variant interactions**;**medRxiv** 2022.07.29.22278162; doi: https://doi.org/10.1101/2022.07.29.22278162.

  Este trabajo propone una nueva herramienta para investigar la asociación de las interacciones genéticas, interacción entre variantes o epistasis, con el Parkinson. Se ha desarrollado un paquete de R con una pipeline para el análisis de las interacciones más prometedores, además se han descrito 14 interacciones significativas y se han replicado dos de ellas en otras poblaciones. Este artículo se encuentra en proceso de revisión para mandarlo a una revista.

8.1.2 Publicaciones como coautor

En esta sección señalaremos los trabajos en los que se ha participado como coautor, cuatro de ellos han sido publicados en revistas y otros dos se encuentran en revisión:

- García-Ruiz S, Gil-Martínez AL, **Cisterna A**, Jurado-Ruiz. F, Reynolds RH, Cookson MR, Hardy J, Ryten M and Botía JA (2021) **CoExp: A Web Tool for the Exploitation of Co-expression Networks**. **Front. Genet.** 12:630187. doi: https://doi.org/10.3389/fgene.2021.630187. Según los datos de SJR la revista es Q2 para los topic de genetics y molecular medicine. La revista tiene un factor de impacto de 4,77.

- Shaul Lerner, Raya Eilam, Lital Adler, Julien Baruteau, Topaz Kreiser, Michael Tsoory, Alexander Brandis, Tevie Mehlman, Mina Ryten, Juan A. Botia, Sonia Garcia Ruiz, **Alejandro Cisterna García**, Carlo Dionisi-Vici, Giusy Ranucci, Marco Spada, Ram Mazkereth, Robert McCarter, Rima Izem, Thomas J. Balmat, Rachel Richesson, Members of the UCDC, Ehud Gazit, Sandesh C. S. Nagamani and Ayelet Erez. **ASL expression in ALDH1A1+ neurons in the substantia nigra metabolically contributes to neurodegenerative phenotype**. **Hum Genet** 140, 1471–1485 (2021). https://doi.org/10.1007/s00439-021-02345-5. Según los datos de SJR la revista es Q1 para el topic de genetics y clinical genetics. La revista tiene un factor de impacto de 5,88.

- Juan A Sánchez, Ana L Gil-Martinez, **Alejandro Cisterna**, Sonia García-Ruíz, Alicia Gómez-Pascual, Regina H Reynolds, Mike Nalls, John Hardy, Mina Ryten and Juan A Botía. **Modeling multifunctionality of genes with secondary gene co-expression networks in human brain provides novel disease insights**, **Bioinformatics**, Volume 37, Issue 18, 15 September 2021, Pages 2905–2911, https://doi.org/10.1093/bioinformatics/btab175. Según los datos de SJR la revista es Q1 para los topic de biochemistry, computational mathematics, computational theory and mathematics, computer science applications, molecular biology y statistics and probability. La revista tiene un factor de impacto de 6,93.

- Chen Z, Tucci A, Cipriani V, Gustavsson EK, Ibañez K, Reynolds RH, Zhang D, Vestito L, **Cisterna García A**, Sethi S, Brenton JW, García-Ruiz S, Fairbrother-Browne A, Gil-Martinez AL; Genomics England Research Consortium Nick Wood; Hardy JA, Smedley D, Houlden H, Botía J, Ryten M. **Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia**. **Brain**. 2023 Jan 10:awad009. doi: https://doi.org/10.1093/brain/awad009. Epub ahead of print. PMID: 36624280. Según los datos de SJR la revista es Q1 para los topic de medicine y clinical neurology. La revista tiene un factor de impacto de 15,25.

- A. Gómez-Pascual, A. Martirosyan, K. Hebestreit, C. Mameffe, S. Poovathingal, T. G. Belgard, C. A. Altar, A. Kottick, M. Holt, V. Hanson-Smith, **A. Cisterna**, M. Mighdoll, R. Scannevin, S. Guelfi, J. A. Botía. **Single-nucleus co-expression networks of dopaminergic neurons support iron accumulation as a plausible explanation to their vulnerability in Parkinson's disease**; **bioRxiv** 2022.12.13.514863; doi: `https://doi.org/10.1101/2022.12.13.514863`

- Konstantin Senkevich, Sara Bandres-Ciga, **Alejandro Cisterna-García**, Eric Yu, Bernabe I. Bustos, Lynne Krohn, Steven J. Lubbe, Juan A. Botía, the International Parkinson's Disease Genomics Consortium (IPDGC), Ziv Gan-Or. **Genome-wide association study stratified by MAPT haplotypes identifies potential novel loci in Parkinson's disease**; **medRxiv** 2023.04.14.23288478; doi: `https://doi.org/10.1101/2023.04.14.23288478`

8.1.3 Comunicaciones a congresos como primer autor

- **Alejandro Cisterna Garcia**; Hsiang-Han Chen; Joanne Norton; Jen Gentsch; Kristy Bergmann; Fengxian Wang; John Budde; Carlos Cruchaga; Juan Antonio Botia Blaya; Laura Ibañez. ; **Preclinical Alzheimer's Disease accurate prediction using plasma cell-free RNA sequences**; **Alzheimer's Association International Conference**, 2022; San Diego, Estados Unidos de América

- **Alejandro Cisterna Garcia**; Hsiang-Han Chen; Oscar Harari; Carlos Cruchaga; Juan Antonio Botia Blaya; **Predicción de estadios tempranos de Alzheimer usando RNA libre circulante del plasma**; **VII Jornadas Doctorales 2022 Campus Mare Nostrum** (Universidad de Murcia y Universidad Politécnica de Cartagena)

- **Alejandro Cisterna Garcia**; Hsiang-Han Chen; Carlos Cruchaga; Juan Antonio Botia Blaya; Laura Ibañez; **Plasma Cell-Free RNA as Non-Invasive Biomarker for Parkinson's Disease; The PSG Thirty-Fourth Annual Symposium on Etiology, Pathogenesis, and Treatment of Parkinson's Disease and Other Movement Disorders**; Phoenix, Estados Unidos de América

- **Alejandro Cisterna García**; Hsiang-Han Chen; Oscar Harari; Carlos Cruchaga; Juan Antonio Botia Blaya; Laura Ibañez.; **PREDICTION OF EARLY STAGES OF ALZHEIMER DISEASE USING PLASMA RNA SEQUENCES; AD/PD™**

2022, Alzheimer's & Parkinson's Diseases Conference; Barcelona, Cataluña, España

- **Alejandro Cisterna García**; Irene Díez; Paolo Maietta; Sara Álvarez; Alvaro Sánchez Ferrer; Juan Antonio Botía.; **In silico association of the seizures phenotype to iron-induced non-apoptotic cell death**; **European Society of Human Genetics ESHG 2021**; Viena, Austria

- **Alejandro Cisterna García**; Aurora González Vidal; Daniel Ruiz Villa; Jordi Ortiz; Ana Sabater Aguado; John A Hardy; Irene Díez García-Prieto; Paolo Maietta; Mina Ryten; Sara Álvarez de Ándres; Juan A. Botía Blaya.; **Switching between mendelian genes space and clinical phenotypes space contributes to comparative Mendelian gene sets analysis**; **European Bioconductor Meeting** 2020

- **Alejandro Cisterna**; Ana Sabater; David Zhang; John Anthony Hardy; Mina Ryten; Irene Díez; Paolo Maietta; Sara Álvarez; Juan Antonio Botía. **Towards a more efficient genetic diagnosis of epilepsy: switching between a mendelian genes space and a clinical phenotypes space**. **European Society of Human Genetics ESHG** 2020

8.1.4 Proyecto adicional colaboración con NIMGenetics

Además de las publicaciones descritas, es importante mencionar el resto de tareas realizadas asociadas a la generación de la tesis.

Los resultados obtenidos a lo largo de la tesis han posibilitado un proyecto de colaboración entre la Universidad de Murcia y NIMGenetics. Este proyecto se centra en la **detección y priorización de las variantes causantes de enfermedad en casos clínicos reales utilizando el fenotipo del paciente e AI**. El nombre del proyecto y el identificador es 37070 SOFTWARE PARA PRIORIZACIÓN DE VARIANTES (SPV).

Hemos integrado y utilizado distintas fuentes de datos como HPO, OMIM, Clinvar y la información clínica y genómica del paciente en una herramienta. Dicha herramienta es capaz de señalar y priorizar determinadas variantes de entre todas las del sujeto. Un individuo suele presentar unas 100000 variantes, el software es capaz de encontrar la variante o variantes probablemente patogénica señalando únicamente 80 variantes en el

96% de los casos en los que se ha evaluado. La evaluación se ha llevado a cabo con más de 400 casos clínicos reales.

8.1.5 Participación en otros proyectos de investigación durante el doctorado

Participación en el proyecto para investigación de COVID-19: 00007/COVI/20 PREDICCIÓN DE PRONÓSTICO EN PACIENTES DE COVID-19 BASADO EN INTELIGENCIA ARTIFICIAL (PROVIA).

Proyectos de financiación internacional ID: 37128 CELL-FREE RNA AS ALZHEIMER'S DRUG DISEASE NON-INVASIVE BIOMARKER. ENTIDAD: UNIVERSIDAD DE MURCIA COMIENZO: 01/04/2022 ,FIN: 31/12/2022 INVESTIGADOR PRINCIPAL: Botía Blaya, Juan A. OTROS INVESTIGADORES: CISTERNA GARCIA, A.

BIBLIOGRAFÍA

[1] A. Bayat. Science, medicine, and the future: Bioinformatics. *BMJ*, 324(7344):1018–1022, April 2002.

[2] Mihir R. Atreya and Hector R. Wong. Precision medicine in pediatric sepsis. *Current Opinion in Pediatrics*, 31(3):322–327, June 2019.

[3] Iuliia Branco and Altino Choupina. Bioinformatics: new tools and applications in life science and personalized medicine. *Applied Microbiology and Biotechnology*, 105(3):937–951, February 2021.

[4] Christian Castaneda, Kip Nalley, Ciaran Mannion, Pritish Bhattacharyya, Patrick Blake, Andrew Pecora, Andre Goy, and K Stephen Suh. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, 5(1):4, December 2015.

[5] Laura M. Hack, Gabriel R. Fries, Harris A. Eyre, Chad A. Bousman, Ajeet B. Singh, Joao Quevedo, Vineeth P. John, Bernhard T. Baune, and Boadie W. Dunlop. Moving pharmacoepigenetics tools for depression toward clinical use. *Journal of Affective Disorders*, 249:336–346, April 2019.

[6] J. Lejeune, M. Gautier, and R. Turpin. [Study of somatic chromosomes from 9 mongoloid children]. *Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences*, 248(11):1721–1722, March 1959.

[7] Melina Claussnitzer, Judy H. Cho, Rory Collins, Nancy J. Cox, Emmanouil T. Dermitzakis, Matthew E. Hurles, Sekar Kathiresan, Eimear E. Kenny, Cecilia M. Lindgren, Daniel G. MacArthur, Kathryn N. North, Sharon E. Plon, Heidi L. Rehm, Neil Risch, Charles N. Rotimi, Jay Shendure, Nicole Soranzo, and Mark I. McCarthy. A brief history of human disease genetics. *Nature*, 577(7789):179–189, January 2020.

[8] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong,

Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, February 2001.

[9] W. Richard McCombie, John D. McPherson, and Elaine R. Mardis. Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, 9(11):a036798, November 2019.

[10] Mario Vailati-Riboni, Valentino Palombo, and Juan J. Loor. What Are Omics Sciences? In Burim N. Ametaj, editor, *Periparturient Diseases of Dairy Cows*, pages 1–7. Springer International Publishing, Cham, 2017.

[11] E.A. Milward, A. Shahandeh, M. Heidari, D.M. Johnstone, N. Daneshi, and H. Hondermarck. Transcriptomics. In *Encyclopedia of Cell Biology*, pages 160–165. Elsevier, 2016.

[12] Li-Rong Yu, Nicolas A. Stewart, and Timothy D. Veenstra. Proteomics. In *Essentials of Genomic and Personalized Medicine*, pages 89–96. Elsevier, 2010.

[13] Clary B. Clish. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies*, 1(1):a000588, October 2015.

[14] T.A. Turunen, M.-A. Väänänen, and S. Ylä-Herttuala. Epigenomics. In *Encyclopedia of Cardiovascular Research and Medicine*, pages 258–265. Elsevier, 2018.

[15] N. M. R. Sales, P. B. Pelegrini, and M. C. Goersch. Nutrigenomics: Definitions and Advances of This New Science. *Journal of Nutrition and Metabolism*, 2014:1–6, 2014.

[16] Kui Yang and Xianlin Han. Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. *Trends in Biochemical Sciences*, 41(11):954–969, November 2016.

[17] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, December 2017.

[18] Konrad J. Karczewski and Michael P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310, May 2018.

[19] Cornelis J.M. Melief, RenéE.M. Toes, Jan Paul Medema, Sjoerd H. Van Der Burg, Ferry Ossendorp, and Rienk Offringa. Strategies for immunotherapy of cancer. In *Advances in Immunology*, volume 75, pages 235–282. Elsevier, 2000.

[20] Maura Abbott and Yelena Ustoyev. Cancer and the Immune System: The History and Background of Immunotherapy. *Seminars in Oncology Nursing*, 35(5):150923, October 2019.

[21] Junzo Hamanishi, Masaki Mandai, Noriomi Matsumura, Kaoru Abiko, Tsukasa Baba, and Ikuo Konishi. PD-1/PD-L1 blockade in cancer treatment: perspectives and issues. *International Journal of Clinical Oncology*, 21(3):462–473, June 2016.

[22] Kheng Newick, Shaun O'Brien, Edmund Moon, and Steven M. Albelda. CAR T Cell Therapy for Solid Tumors. *Annual Review of Medicine*, 68(1):139–152, January 2017.

[23] Zhenguang Wang, Zhiqiang Wu, Yang Liu, and Weidong Han. New development in CAR-T cell therapy. *Journal of Hematology & Oncology*, 10(1):53, December 2017.

[24] Tzen S. Toh, Frank Dondelinger, and Dennis Wang. Looking beyond the hype: Applied AI and machine learning in translational medicine. *EBioMedicine*, 47:607–615, September 2019.

[25] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, January 2022.

[26] Pratik Shah, Francis Kendall, Sean Khozin, Ryan Goosen, Jianying Hu, Jason Laramie, Michael Ringel, and Nicholas Schork. Artificial intelligence and machine learning in clinical development: a translational perspective. *npj Digital Medicine*, 2(1):69, July 2019.

[27] Julie Abimanyi-Ochom, Shalika Bohingamu Mudiyanselage, Max Catchpool, Marnie Firipis, Sithara Wanni Arachchige Dona, and Jennifer J. Watts. Strategies to reduce diagnostic errors: a systematic review. *BMC Medical Informatics and Decision Making*, 19(1):174, December 2019.

[28] Stéphanie Nguengang Wakap, Deborah M. Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2):165–173, February 2020.

[29] Stephanie Chan, Vidhatha Reddy, Bridget Myers, Quinn Thibodeaux, Nicholas Brownstone, and Wilson Liao. Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. *Dermatology and Therapy*, 10(3):365–386, June 2020.

[30] Sana Ansari, Imran Shafi, Aiza Ansari, Jamil Ahmad, and Syed Ismail Shah. Diagnosis of liver disease induced by hepatitis virus using Artificial Neural Networks. In *2011 IEEE 14th International Multitopic Conference*, pages 8–12, Karachi, Pakistan, December 2011. IEEE.

[31] Zodwa Dlamini, Flavia Zita Francies, Rodney Hull, and Rahaba Marima. Artificial intelligence (AI) and big data in cancer and precision oncology. *Computational and Structural Biotechnology Journal*, 18:2300–2311, 2020.

[32] Dow-Mu Koh, Nickolas Papanikolaou, Ulrich Bick, Rowland Illing, Charles E. Kahn, Jayshree Kalpathi-Cramer, Celso Matos, Luis Martí-Bonmatí, Anne Miles, Seong Ki Mun, Sandy Napel, Andrea Rockall, Evis Sala, Nicola Strickland, and Fred Prior. Artificial intelligence and machine learning in cancer imaging. *Communications Medicine*, 2(1):133, October 2022.

[33] Amanda H. Gonsalves, Fadi Thabtah, Rami Mustafa A. Mohammad, and Gurpreet Singh. Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, pages 51–56, Xiamen China, July 2019. ACM.

[34] Alexandra-Maria Tăuţan, Bogdan Ionescu, and Emiliano Santarnecchi. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artificial Intelligence in Medicine*, 117:102081, July 2021.

[35] Andrea R. Horvath, Sarah J. Lord, Andrew StJohn, Sverre Sandberg, Christa M. Cobbaert, Stefan Lorenz, Phillip J. Monaghan, Wilma D.J. Verhagen-Kamerbeek, Christoph Ebert, and Patrick M.M. Bossuyt. From biomarkers to medical tests: The changing landscape of test evaluation. *Clinica Chimica Acta*, 427:49–57, January 2014.

[36] Robert M Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, February 2018.

[37] Yang Nan, Javier Del Ser, Simon Walsh, Carola Schönlieb, Michael Roberts, Ian Selby, Kit Howard, John Owen, Jon Neville, Julien Guiot, Benoit Ernst, Ana Pastor, Angel Alberich-Bayarri, Marion I. Menzel, Sean Walsh, Wim Vos, Nina Flerin, Jean-Paul Charbonnier, Eva van Rikxoort, Avishek Chatterjee, Henry Woodruff, Philippe Lambin, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, Francisco Herrera, and Guang Yang. Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Information Fusion*, 82:99–122, June 2022.

[38] Vasileios C. Pezoulas, Themis P. Exarchos, and Dimitrios Ioannou Fotiadis. *Medical data sharing, harmonization and analytics*. Academic Press, London, United Kingdom ; San Diego, CA, 2020. OCLC: on1119620876.

[39] Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, January 2015.

[40] Ségolène Aymé and J. Schmidtke. Networking for rare diseases: a necessity for Europe. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 50(12):1477–1483, December 2007.

[41] Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*, 84(4):524–533, April 2009.

[42] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, and Peter N Robinson. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, January 2021.

[43] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand,

Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013.

[44] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, November 2012.

[45] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, November 2021.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[47] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, January 2021.

[48] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, January 2011.

[49] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[50] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007.

[51] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing : Imprint: Springer, Cham, 2nd ed. 2016 edition, 2016.

[52] Max Kuhn. Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, 28(5), 2008.

[53] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. dplyr: A Grammar of Data Manipulation, 2023.

[54] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, August 2005.

[55] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, April 2015.

[56] François Chollet and others. Keras, 2015.

[57] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[58] TensorFlow Developers. TensorFlow, March 2023.

[59] Gerardo Jimenez-Sanchez, Barton Childs, and David Valle. Human disease genes. *Nature*, 409(6822):853–855, February 2001.

[60] Paweł Stankiewicz and James R. Lupski. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, 61(1):437–455, February 2010.

[61] Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9):1748–1759, September 2012.

[62] C. J. Shaw. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human Molecular Genetics*, 13(90001):57R–64, January 2004.

[63] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren A Kibbe, Lihua Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10(S1):S6, July 2009.

[64] Peter N. Robinson, Sebastian Köhler, Anika Oellrich, Sanger Mouse Genetics Project, Kai Wang, Christopher J. Mungall, Suzanna E. Lewis, Nicole Washington, Sebastian Bauer, Dominik Seelow, Peter Krawitz, Christian Gilissen, Melissa Haendel, and Damian Smedley. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2):340–348, February 2014.

[65] Larissa K. F. Temple, Robin S. McLeod, Steven Gallinger, and James G. Wright. Defining Disease in the Genomics Era. *Science*, 293(5531):807–808, August 2001.

[66] Jackie Leach Scully. What is a disease?: Disease, disability and their definitions. *EMBO reports*, 5(7):650–653, July 2004.

[67] Marylyn D. Ritchie, Joshua C. Denny, Dana C. Crawford, Andrea H. Ramirez, Justin B. Weiner, Jill M. Pulley, Melissa A. Basford, Kristin Brown-Gentry, Jeffrey R. Balser, Daniel R. Masys, Jonathan L. Haines, and Dan M. Roden. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *The American Journal of Human Genetics*, 86(4):560–572, April 2010.

[68] David J. Hunter. Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298, April 2005.

[69] Seng H. Cheng, Richard J. Gregory, John Marshall, Sucharita Paul, David W. Souza, Gary A. White, Catherine R. O'Riordan, and Alan E. Smith. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell*, 63(4):827–834, November 1990.

[70] Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, May 2008.

[71] Torben Hansen. Type 2 diabetes mellitus–a multifactorial disease. *Annales Universitatis Mariae Curie-Sklodowska. Sectio D: Medicina*, 57(1):544–549, 2002.

[72] Carol J Bult, Judith A Blake, Cynthia L Smith, James A Kadin, Joel E Richardson, the Mouse Genome Database Group, A Anagnostopoulos, R Asabor, R M Baldarelli, J S Beal, S M Bello, O Blodgett, N E Butler, K R Christie, L E Corbani, J Creelman, M E Dolan, H J Drabkin, S L Giannatto, P Hale, D P Hill, M Law, A Mendoza, M McAndrews, D Miers, H Motenko, L Ni, H Onda, M Perry, J M Recla, B Richards-Smith, D Sitnikov, M Tomczuk, G Tonorio, L Wilming, and Y Zhu. Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806, January 2019.

[73] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, January 2019.

[74] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143, January 2021.

[75] Heidi L. Rehm, Jonathan S. Berg, Lisa D. Brooks, Carlos D. Bustamante, James P. Evans, Melissa J. Landrum, David H. Ledbetter, Donna R. Maglott, Christa Lese Martin, Robert L. Nussbaum, Sharon E. Plon, Erin M. Ramos, Stephen T. Sherry, and Michael S. Watson. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine*, 372(23):2235–2242, June 2015.

[76] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple,

Arianna Tucci, Helen Brittain, Anna de Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019.

[77] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, Carmen de Torres, Rodrigo Dienstmann, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25, December 2018.

[78] Alba Gutiérrez-Sacristán, Solène Grosdidier, Olga Valverde, Marta Torrens, Àlex Bravo, Janet Piñero, Ferran Sanz, and Laura I. Furlong. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics*, 31(18):3075–3077, September 2015.

[79] Ruilin Tian, Anthony Abarientos, Jason Hong, Sayed Hadi Hashemi, Rui Yan, Nina Dräger, Kun Leng, Mike A. Nalls, Andrew B. Singleton, Ke Xu, Faraz Faghri, and Martin Kampmann. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24(7):1020–1034, July 2021.

[80] Tomasz Zemojtel, Sebastian Köhler, Luisa Mackenroth, Marten Jäger, Jochen Hecht, Peter Krawitz, Luitgard Graul-Neumann, Sandra Doelken, Nadja Ehmke, Malte Spielmann, Nancy Christine Øien, Michal R. Schweiger, Ulrike Krüger, Götz Frommer, Björn Fischer, Uwe Kornak, Ricarda Flöttmann, Amin Ardeshirdavani, Yves Moreau, Suzanna E. Lewis, Melissa Haendel, Damian Smedley, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine*, 6(252), September 2014.

[81] Raj Kalaria. Similarities between Alzheimer's disease and vascular dementia. *Journal of the Neurological Sciences*, 203-204:29–34, November 2002.

[82] ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson,

Mark J Daly, Alkes L Price, and Benjamin M Neale. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, November 2015.

[83] H. Lohi, J. Turnbull, X. C. Zhao, S. Pullenayegum, L. Ianzano, M. Yahyaoui, M. A. Mikati, N. P. Quinn, S. Franceschetti, F. Zara, and B. A. Minassian. Genetic diagnosis in Lafora disease: Genotype–phenotype correlations and diagnostic pitfalls. *Neurology*, 68(13):996–1001, March 2007.

[84] Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *The American Journal of Human Genetics*, 85(4):457–464, October 2009.

[85] Yue Deng, Lin Gao, Bingbo Wang, and Xingli Guo. HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology. *PLOS ONE*, 10(2):e0115692, February 2015.

[86] Jiajie Peng, Hansheng Xue, Weiwei Hui, Junya Lu, Bolin Chen, Qinghua Jiang, Xuequn Shang, and Yadong Wang. An online tool for measuring and visualizing phenotype similarities using HPO. *BMC Genomics*, 19(S6):571, August 2018.

[87] Meng-Pin Weng and Ben-Yang Liao. modPhEA: model organism Phenotype Enrichment Analysis of eukaryotic gene sets. *Bioinformatics*, 33(21):3505–3507, November 2017.

[88] Jiguang Wang, Qiang Huang, Zhi-Ping Liu, Yong Wang, Ling-Yun Wu, Luonan Chen, and Xiang-Sun Zhang. NOA: a novel Network Ontology Analysis method. *Nucleic Acids Research*, 39(13):e87–e87, July 2011.

[89] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking : Bringing Order to the Web. In *The Web Conference*, 1999.

[90] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, page gkz1021, November 2019.

[91] Bryony Braschi, Paul Denny, Kristian Gray, Tamsin Jones, Ruth Seal, Susan Tweedie, Bethan Yates, and Elspeth Bruford. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Research*, 47(D1):D786–D792, January 2019.

[92] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[93] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995.

[94] John Alroy. A new twist on a very old binary similarity coefficient. *Ecology*, 96(2):575–586, February 2015.

[95] Stefania Salvatore, Knut Dagestad Rand, Ivar Grytten, Egil Ferkingstad, Diana Domanska, Lars Holden, Marius Gheorghe, Anthony Mathelier, Ingrid Glad, and Geir Kjetil Sandve. Beware the Jaccard: the choice of **similarity measure** is important and non-trivial in genomic colocalisation analysis. *Briefings in Bioinformatics*, 21(5):1523–1530, September 2020.

[96] Winston Chang, Joe Cheng, J. J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*. 2023.

[97] Nicki Niemann and Joseph Jankovic. Juvenile parkinsonism: Differential diagnosis, genetics, and treatment. *Parkinsonism & Related Disorders*, 67:74–89, October 2019.

[98] Xandra O. Breakefield, Anne J. Blood, Yuqing Li, Mark Hallett, Phyllis I. Hanson, and David G. Standaert. The pathophysiological basis of dystonias. *Nature Reviews Neuroscience*, 9(3):222–234, March 2008.

[99] A. Berardelli. Pathophysiology of bradykinesia in Parkinson's disease. *Brain*, 124(11):2131–2146, November 2001.

[100] Pei-Hao Chen, Rong-Long Wang, De-Jyun Liou, and Jin-Siang Shaw. Gait Disorders in Parkinson's Disease: Assessment and Management. *International Journal of Gerontology*, 7(4):189–193, December 2013.

[101] Stephane Hunot and E. C. Hirsch. Neuroinflammatory processes in Parkinson's disease. *Annals of Neurology*, 53(S3):S49–S60, 2003.

[102] Gerard W. O'Keeffe and Aideen M. Sullivan. Evidence for dopaminergic axonal degeneration as an early pathological process in Parkinson's disease. *Parkinsonism & Related Disorders*, 56:9–15, November 2018.

[103] A. Albanese, M. Di Giovanni, and S. Lalli. Dystonia: diagnosis and management. *European Journal of Neurology*, 26(1):5–17, January 2019.

[104] A. Brashear, M. R. Farlow, I. J. Butler, E. J. Kasarskis, and W. B. Dobyns. Variable phenotype of rapid-onset dystonia-parkinsonism. *Movement Disorders*, 11(2):151–156, March 1996.

[105] Raffaella Romano, Alessandro Bertolino, Angelo Gigante, Davide Martino, Paolo Livrea, and Giovanni Defazio. Impaired cognitive functions in adult-onset primary cranial cervical dystonia. *Parkinsonism & Related Disorders*, 20(2):162–165, February 2014.

[106] Shinichi Furuya, Kenta Tominaga, Fumio Miyazaki, and Eckart Altenmüller. Losing dexterity: patterns of impaired coordination of finger movements in musician's dystonia. *Scientific Reports*, 5(1):13360, August 2015.

[107] Anna Castagna, Serena Frittoli, Maurizio Ferrarin, Francesca Del Sorbo, Luigi M. Romito, Antonio E. Elia, and Alberto Albanese. Quantitative gait analysis in parkin disease: Possible role of dystonia: Quantitative Gait Analysis in Parkin Disease. *Movement Disorders*, 31(11):1720–1728, November 2016.

[108] H. Booth, W. D. Hirst, and R. Wade-Martins. The role of astrocyte dysfunction in Parkinson's disease pathogenesis. *Trends Neurosci*, 40, 2017.

[109] C. Y. Kim, T. Wirth, and C. Hubsch. Early-onset parkinsonism is a manifestation of the PPP2R5D p. E200K mutation. *Ann Neurol*, 88, 2020.

[110] F. L. Muiswinkel, R. A. I. Vos, and J. G. J. M. Bol. Expression of NAD (P) H: quinone oxidoreductase in the normal and Parkinsonian substantia nigra. *Neurobiol Aging*, 25, 2004.

[111] C. Zarow, S. A. Lyness, and J. A. Mortimer. Neuronal loss is greater in the locus coeruleus than nucleus basalis and substantia nigra in Alzheimer and Parkinson diseases. *Arch Neurol*, 60, 2003.

[112] T. Ziemssen and H. Reichmann. Cardiovascular autonomic dysfunction in Parkinson's disease. *J Neurol Sci*, 289, 2010.

[113] D. Aarsland and M. W. Kurz. The epidemiology of dementia associated with Parkinson disease. *J Neurol Sci*, 289, 2010.

[114] K. Rosenkranz, A. Williamon, and K. Butler. Pathophysiological differences between musician's dystonia and writer's cramp. *Brain*, 128, 2005.

[115] M. H. Ibrahim, A. Fadhil, and S. S. Ali. Could dystonia be initial presentation of corpus callosum infarction in young age patients? A case report study. *Neurosci Med*, 6, 2015.

[116] C. Colosimo, P. Pantano, and V. Calistri. Diffusion tensor imaging in primary cervical dystoniaJournal of Neurology. *Neurosurg Psychiatry*, 76, 2005.

[117] S. A. Schneider, A. E. Lang, and E. Moro. Characteristic head drops and axial extension in advanced chorea-acanthocytosis. *Mov Disord*, 25, 2010.

[118] K. M. Gorman, E. Meyer, and M. A. Kurian. Review of the phenotype of early-onset generalised progressive dystonia due to mutations in KMT2B. *Eur J Paediatr Neurol*, 22, 2018.

[119] K. Lohmann and C. Klein. Update on the genetics of dystonia. *Curr Neurol Neurosci Rep*, 17, 2017.

[120] A. J. Groffen, T. Klapwijk, and A. F. Rootselaar. Genetic and phenotypic heterogeneity in sporadic and familial forms of paroxysmal dyskinesia. *J Neurol*, 260, 2013.

[121] Botía, J. A., Guelfi, S., Zhang, D., et al. (2018). G2P: Using machine learning to understand and predict genes causing rare neurological disorders. bioRxiv, 288845.

[122] C. E. Stafstrom and L. Carmant. Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harb Perspect Med*, 5, 2015.

[123] H. Ishiura, K. Doi, and J. Mitsui. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet*, 50, 2018.

[124] E. Trinka, J. Höfler, and A. Zerbs. Causes of status epilepticus. *Epilepsia*, 53, 2012.

[125] G. M. Abdel-Salam, A. A. Halász, and A. E. Czeizel. Association of epilepsy with different groups of microcephaly. *Dev Med Child Neurol*, 42, 2000.

[126] M. D. C. Carvalho, R. A. Ximenes, and U. R. Montarroyos. Early epilepsy in children with Zika-related microcephaly in a cohort in Recife, Brazil: Characteristics, electroencephalographic findings, and treatment response. *Epilepsia*, 61, 2020.

[127] A. Ricobaraza, L. Mora-Jimenez, and E. Puerta. Epilepsy and neuropsychiatric comorbidities in mice carrying a recurrent Dravet syndrome SCN1A missense mutation. *Sci Rep*, 9, 2019.

[128] P. Parisi, R. Moavero, and A. Verrotti. Attention deficit hyperactivity disorder in children with epilepsy. *Brain Develop*, 32, 2010.

[129] B. H. Lee, T. Smith, and A. R. Paciorkowski. Autism spectrum disorder and epilepsy: disorders with a shared biology. *Epilepsy Behav*, 47, 2015.

[130] R. Tuchman and I. Rapin. Epilepsy in autism. *Lancet Neurol*, 1, 2002.

[131] Epi25 Collaborative. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. Am J Hum Genet. 2019;105(2):267–282. https://doi.org/10.1016/j.ajhg.2019.05.020.

[132] B. Hu, H. Guo, P. Zhou, and Z. . L. Shi. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.*, 19, 2021.

[133] WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.

[134] Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C. & Di Napoli, R. Features, Evaluation, and Treatment of Coronavirus (COVID-19). in StatPearls (StatPearls Publishing, 2021).

[135] D. Hornuss. Anosmia in COVID-19 patients. *Clin. Microbiol. Infect.*, 26, 2020.

[136] Z. Zhou. Effect of gastrointestinal symptoms in patients with COVID-19. *Gastroenterology*, 158, 2020.

[137] H. . Y. Wang. Potential neurological symptoms of COVID-19. *Ther. Adv. Neurol. Disord.*, 13, 2020.

[138] G. Pascarella. COVID-19 diagnosis and management: A comprehensive review. *J. Intern. Med.*, 288, 2020.

[139] Carfì, A., Bernabei, R., Landi, F., & for the Gemelli against COVID-19 Post-Acute Care Study Group. Persistent symptoms in patients after acute COVID-19. JAMA324, 603 (2020).

[140] Z. Zheng. Risk factors of critical and mortal COVID-19 cases: A systematic literature review and meta-analysis. *J. Infect.*, 81, 2020.

[141] Lopez Bernal, J. et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: Test negative case-control study. BMJ n1088 (2021). https://doi.org/10.1136/bmj.n1088.

[142] Ioannou, G. N. et al. COVID-19 Vaccination effectiveness against infection or death in a national U.S. health care system: A target trial emulation study. Ann. Intern. Med. M21-3256 (2021). https://doi.org/10.7326/M21-3256.

[143] Johnson, A. G. et al. COVID-19 Incidence and death rates among unvaccinated and fully vaccinated adults with and without booster doses during periods of delta and omicron variant emergence—25 U.S. Jurisdictions, April 4–December 25, 2021. MMWR Morb. Mortal. Wkly. Rep.71, 132–138 (2022).

[144] S. Sanche. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.*, 26, 2020.

[145] B. Sen-Crowe, M. Sutherland, M. McKenney, and A. Elkbuli. A closer look into global hospital beds capacity and resource shortages during the COVID-19 pandemic. *J. Surg. Res.*, 260, 2021.

[146] E. Mannucci, G. A. Silverii, and M. Monami. Saturation of critical care capacity and mortality in patients with the novel coronavirus (COVID-19) in Italy. *Trends Anaesth. Crit. Care*, 33, 2020.

[147] A. Olivas-Martínez. In-hospital mortality from severe COVID-19 in a tertiary care center in Mexico City; causes of death, risk factors and the impact of hospital saturation. *PLoS ONE*, 16, 2021.

[148] L. Yan. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.*, 2, 2020.

[149] Y. Gao. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat. Commun.*, 11, 2020.

[150] C. Ryan. Predicting severe outcomes in Covid-19 related illness using only patient demographics, comorbidities and symptoms. *Am. J. Emerg. Med.*, 45, 2021.

[151] A. Chatterjee. Can predicting COVID-19 mortality in a European cohort using only demographic and comorbidity data surpass age-based prediction: An externally validated study. *PLoS ONE*, 16, 2021.

[152] E. Jimenez-Solem. Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. *Sci. Rep.*, 11, 2021.

[153] L. Breiman. Random forest. *Mach. Learn.*, 45, 2001.

[154] Kuhn, M. Building predictive models in R using the caret package. J. Stat. Softw.28, (2008).

[155] J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20, 1960.

[156] M. E. Killerby. Characteristics associated with hospitalization among patients with COVID-19—Metropolitan Atlanta, Georgia, March–April 2020. *MMWR Morb. Mortal. Wkly. Rep.*, 69, 2020.

[157] G. M. Vahey. Risk factors for hospitalization among persons with COVID-19—Colorado. *PLoS ONE*, 16, 2021.

[158] F. K. Ho. Is older age associated with COVID-19 mortality in the absence of other risk factors? General population cohort study of 470,034 participants. *PLoS ONE*, 15, 2020.

[159] Mahase, E. Covid-19: Why are age and obesity risk factors for serious disease? BMJ m4130. https://doi.org/10.1136/bmj.m4130 (2020).

[160] M. Biswas, S. Rahaman, T. K. Biswas, Z. Haque, and B. Ibrahim. Association of sex, age, and comorbidities with mortality in COVID-19 patients: A systematic review and meta-analysis. *Intervirology*, 64, 2021.

[161] N. T. Nguyen. Male gender is a predictor of higher mortality in hospitalized adults with COVID-19. *PLoS ONE*, 16, 2021.

[162] H. Peckham. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nat. Commun.*, 11, 2020.

[163] P. A. Franco, S. Jezler, and A. A. Cruz. Is asthma a risk factor for coronavirus disease-2019 worse outcomes? The answer is no, but .... *Curr. Opin. Allergy Clin. Immunol.*, 21, 2021.

[164] D. T. Timberlake, K. Strothman, and M. H. Grayson. Asthma, severe acute respiratory syndrome coronavirus-2 and coronavirus disease 2019. *Curr. Opin. Allergy Clin. Immunol.*, 21, 2021.

[165] R. K. Topless. Gout, rheumatoid arthritis, and the risk of death related to coronavirus disease 2019: An analysis of the UK Biobank. *ACR Open Rheumatol.*, 3, 2021.

[166] N. Salari. The global prevalence of osteoporosis in the world: A comprehensive systematic review and meta-analysis. *J. Orthop. Surg.*, 16, 2021.

[167] G. Hampson, M. Stone, J. R. Lindsay, R. K. Crowley, and S. H. Ralston. Diagnosis and management of osteoporosis during COVID-19: Systematic review and practical guidance. *Calcif. Tissue Int.*, 109, 2021.

[168] J. E. Peña. Hypertension, diabetes and obesity, major risk factors for death in patients with COVID-19 in Mexico. *Arch. Med. Res.*, 52, 2021.

[169] H. Surendra. Clinical characteristics and mortality associated with COVID-19 in Jakarta, Indonesia: A hospital-based retrospective cohort study. *Lancet Reg. Health West. Pac.*, 9, 2021.

[170] F. Ceban. Association between mood disorders and risk of COVID-19 infection, hospitalization, and death: A systematic review and meta-analysis. *JAMA Psychiat.*, 78, 2021.

[171] F. Zhou. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet*, 395, 2020.

[172] C. Huang. Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. *The Lancet*, 395, 2020.

[173] K. Ikemura. Using automated machine learning to predict the mortality of patients with COVID-19: Prediction model development study. *J. Med. Internet Res.*, 23, 2021.

[174] F. Nezhadmoghadam and J. Tamez-Peña. Risk profiles for negative and positive COVID-19 hospitalized patients. *Comput. Biol. Med.*, 136, 2021.

[175] C. Bonanad. The effect of age on mortality in patients with COVID-19: A meta-analysis with 611,583 subjects. *J. Am. Med. Dir. Assoc.*, 21, 2020.

[176] G. Halasz. A machine learning approach for mortality prediction in COVID-19 pneumonia: Development and evaluation of the piacenza score. *J. Med. Internet Res.*, 23, 2021.

[177] M. A. Dabbah. Machine learning approach to dynamic risk modeling of mortality in COVID-19: A UK Biobank study. *Sci. Rep.*, 11, 2021.

[178] A. Y. Wong. Use of non-steroidal anti-inflammatory drugs and risk of death from COVID-19: An OpenSAFELY cohort analysis based on two cohorts. *Ann. Rheum. Dis.*, 80, 2021.

[179] A. C. Tahira, S. Verjovski-Almeida, and S. T. Ferreira. Dementia is an age-independent risk factor for severity and death in COVID-19 inpatients. *Alzheimers Dement.*, 17, 2021.

[180] L. Chan. AKI in hospitalized patients with COVID-19. *J. Am. Soc. Nephrol.*, 32, 2021.

[181] Frank M. LaFerla and Salvatore Oddo. Alzheimer's disease: Aβ, tau and synaptic dysfunction. *Trends in Molecular Medicine*, 11(4):170–176, April 2005.

[182] 2021 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 17(3):327–406, March 2021.

[183] Alzheimer's Association. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, March 2018.

[184] Bruno Dubois, Harald Hampel, Howard H. Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, Karl Broich, Enrica Cavedo, Sebastian Crutch, Jean-François Dartigues, Charles Duyckaerts, Stéphane Epelbaum, Giovanni B. Frisoni, Serge Gauthier,

Remy Genthon, Alida A. Gouw, Marie-Odile Habert, David M. Holtzman, Miia Kivipelto, Simone Lista, José-Luis Molinuevo, Sid E. O'Bryant, Gil D. Rabinovici, Christopher Rowe, Stephen Salloway, Lon S. Schneider, Reisa Sperling, Marc Teichmann, Maria C. Carrillo, Jeffrey Cummings, Cliff R. Jack, and Proceedings of the Meeting of the International Working Group (IWG) and the American Alzheimer's Association on "The Preclinical State of AD"; July 23, 2015; Washington DC, USA. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323, March 2016.

[185] Maria Paraskevaidi, David Allsop, Salman Karim, Francis L. Martin, and St-John Crean. Diagnostic Biomarkers for Alzheimer's Disease Using Non-Invasive Specimens. *Journal of Clinical Medicine*, 9(6):1673, June 2020.

[186] Anja H. Simonsen, James McGuire, Oskar Hansson, Henrik Zetterberg, Vladimir N. Podust, Huw A. Davies, Gunhild Waldemar, Lennart Minthon, and Kaj Blennow. Novel Panel of Cerebrospinal Fluid Biomarkers for the Prediction of Progression to Alzheimer Dementia in Patients With Mild Cognitive Impairment. *Archives of Neurology*, 64(3):366, March 2007.

[187] R.S. Osorio, E. Pirraglia, T. Gumb, J. Mantua, I. Ayappa, S. Williams, L. Mosconi, L. Glodzik, and M.J. De Leon. Imaging and Cerebrospinal Fluid Biomarkers in the Search for Alzheimer's Disease Mechanisms. *Neurodegenerative Diseases*, 13(2-3):163–165, 2014.

[188] Peter N. E. Young, Mar Estarellas, Emma Coomans, Meera Srikrishna, Helen Beaumont, Anne Maass, Ashwin V. Venkataraman, Rikki Lissaman, Daniel Jiménez, Matthew J. Betts, Eimear McGlinchey, David Berron, Antoinette O'Connor, Nick C. Fox, Joana B. Pereira, William Jagust, Stephen F. Carter, Ross W. Paterson, and Michael Schöll. Imaging biomarkers in neurodegeneration: current and future practices. *Alzheimer's Research & Therapy*, 12(1):49, December 2020.

[189] Leonardo Biscetti, Nicola Salvadori, Lucia Farotti, Samuela Cataldi, Paolo Eusebi, Silvia Paciotti, and Lucilla Parnetti. The added value of Aβ42/Aβ40 in the CSF signature for routine diagnostics of Alzheimer's disease. *Clinica Chimica Acta*, 494:71–73, July 2019.

[190] Shorena Janelidze, Henrik Zetterberg, Niklas Mattsson, Sebastian Palmqvist, Hugo Vanderstichele, Olof Lindberg, Danielle Westen, Erik Stomrud, Lennart

Minthon, Kaj Blennow, the Swedish BioFINDER study group, and Oskar Hansson. CSF A β 42/A β 40 and A β 42/A β 38 ratios: better diagnostic markers of Alzheimer disease. *Annals of Clinical and Translational Neurology*, 3(3):154–165, March 2016.

[191] Oskar Hansson, Sylvain Lehmann, Markus Otto, Henrik Zetterberg, and Piotr Lewczuk. Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease. *Alzheimer's Research & Therapy*, 11(1):34, December 2019.

[192] Julia Kuhlmann, Ulf Andreasson, Josef Pannee, Maria Bjerke, Erik Portelius, Andreas Leinenbach, Tobias Bittner, Magdalena Korecka, Rand G. Jenkins, Hugo Vanderstichele, Erik Stoops, Piotr Lewczuk, Leslie M. Shaw, Ingrid Zegers, Heinz Schimmel, Henrik Zetterberg, and Kaj Blennow. CSF Aβ1–42 – an excellent but complicated Alzheimer's biomarker – a route to standardisation. *Clinica Chimica Acta*, 467:27–33, April 2017.

[193] Matthew R. Brier, Brian Gordon, Karl Friedrichsen, John McCarthy, Ari Stern, Jon Christensen, Christopher Owen, Patricia Aldea, Yi Su, Jason Hassenstab, Nigel J. Cairns, David M. Holtzman, Anne M. Fagan, John C. Morris, Tammie L. S. Benzinger, and Beau M. Ances. Tau and Aβ imaging, CSF measures, and cognition in Alzheimer's disease. *Science Translational Medicine*, 8(338), May 2016.

[194] K. Buerger, M. Ewers, T. Pirttila, R. Zinkowski, I. Alafuzoff, S. J. Teipel, J. DeBernardis, D. Kerkman, C. McCulloch, H. Soininen, and H. Hampel. CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. *Brain*, 129(11):3035–3041, September 2006.

[195] Argonde C. Van Harten, Maartje I. Kester, Pieter-Jelle Visser, Marinus A. Blankenstein, Yolande A.L. Pijnenburg, Wiesje M. Van Der Flier, and Philip Scheltens. Tau and p-tau as CSF biomarkers in dementia: a meta-analysis. *cclm*, 49(3):353–366, March 2011.

[196] Shorena Janelidze, Erik Stomrud, Ruben Smith, Sebastian Palmqvist, Niklas Mattsson, David C. Airey, Nicholas K. Proctor, Xiyun Chai, Sergey Shcherbinin, John R. Sims, Gallen Triana-Baltzer, Clara Theunis, Randy Slemmon, Marc Mercken, Hartmuth Kolb, Jeffrey L. Dage, and Oskar Hansson. Cerebrospinal fluid p-tau217 performs better than p-tau181 as a biomarker of Alzheimer's disease. *Nature Communications*, 11(1):1683, April 2020.

[197] Ramesh J. L. Kandimalla, Sudesh Prabhakar, Willayat Yousuf Wani, Alka Kaushal, Nidhi Gupta, Deep Raj Sharma, V. K. Grover, Neerja Bhardwaj, Kajal Jain, and Kiran Dip Gill. CSF p-Tau levels in the prediction of Alzheimer's disease. *Biology Open*, 2(11):1119–1124, November 2013.

[198] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Contributors, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, April 2018.

[199] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Contributors, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, April 2018.

[200] Jarith L. Ebenau, Tessa Timmers, Linda M.P. Wesselman, Inge M.W. Verberk, Sander C.J. Verfaillie, Rosalinde E.R. Slot, Argonde C. Van Harten, Charlotte E. Teunissen, Frederik Barkhof, Karlijn A. Van Den Bosch, Mardou Van Leeuwenstijn, Jori Tomassen, Anouk Den Braber, Pieter Jelle Visser, Niels D. Prins, Sietske A.M. Sikkes, Philip Scheltens, Bart N.M. Van Berckel, and Wiesje M. Van Der Flier. ATN classification and clinical progression in subjective cognitive decline: The SCIENCe project. *Neurology*, 95(1):e46–e58, July 2020.

[201] Koen Delmotte, Jolien Schaeverbeke, Koen Poesen, and Rik Vandenberghe. Prognostic value of amyloid/tau/neurodegeneration (ATN) classification based on diagnostic cerebrospinal fluid samples for Alzheimer's disease. *Alzheimer's Research & Therapy*, 13(1):84, December 2021.

[202] Leslie M Shaw, Teresa Waligorska, Lea Fields, Magdalena Korecka, Michal Figurski, John Q Trojanowski, Udo Eichenlaub, Sabine Wahl, Meixiang Quan, Michael J

Pontecorvo, D Richard Lachno, Jane A Talbot, Scott W Andersen, Eric R Siemers, and Robert A Dean. Derivation of cutoffs for the Elecsys® amyloid β (1–42) assay in Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:698–705, 2018.

[203] Silke Ingala, Charlotte De Boer, Lotte A Masselink, Ilaria Vergari, Luca Lorenzini, Kaj Blennow, Gael Chételat, Carol Di Perri, Michael Ewers, Wiesje M van der Flier, Nick C Fox, Juan Domingo Gispert, Sven Haller, José Luis Molinuevo, Graciela Muniz-Terrera, Henri J Mutsaerts, Craig W Ritchie, Karen Ritchie, Mark E Schmidt, Adam J Schwarz, Linda Vermunt, Adam D Waldman, Joanna Wardlaw, Alle Meije Wink, Robin Wolz, Viktor Wottschel, Philip Scheltens, Pieter Jelle Visser, and Frederik Barkhof. Application of the ATN classification scheme in a population without dementia: Findings from the EPAD cohort. *Alzheimer's & Dementia*, 17:1189–1204, 2021.

[204] Jonathan Vogelgsang, Dirk Wedekind, Caroline Bouter, Hans-Wolfgang Klafki, and Jens Wiltfang. Reproducibility of Alzheimer's Disease Cerebrospinal Fluid-Biomarker Measurements under Clinical Routine Conditions. *Journal of Alzheimer's Disease*, 62:203–212.

[205] Massimo S. Fiandaca, Dimitrios Kapogiannis, Mark Mapstone, Adam Boxer, Erez Eitan, Janice B. Schwartz, Erin L. Abner, Ronald C. Petersen, Howard J. Federoff, Bruce L. Miller, and Edward J. Goetzl. Identification of preclinical Alzheimer's disease by a profile of pathogenic proteins in neurally derived blood exosomes: A case-control study. *Alzheimer's & Dementia*, 11(6):600, June 2015.

[206] Bob Olsson, Ronald Lautner, Ulf Andreasson, Annika Öhrfelt, Erik Portelius, Maria Bjerke, Mikko Hölttä, Christoffer Rosén, Caroline Olsson, Gabrielle Strobel, Elizabeth Wu, Kelly Dakin, Max Petzold, Kaj Blennow, and Henrik Zetterberg. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology*, 15(7):673–684, June 2016.

[207] J. L. Dage, A. M. V. Wennberg, D. C. Airey, C. E. Hagen, D. S. Knopman, M. M. Machulda, R. O. Roberts, C. R. Jack, R. C. Petersen, and M. M. Mielke. Levels of tau protein in plasma are associated with neurodegeneration and cognitive function in a population-based elderly cohort. *Alzheimer's & Dementia*, 12:1226–1234, 2016.

[208] M. P. Pase, A. S. Beiser, J. J. Himali, C. L. Satizabal, H. J. Aparicio, C. DeCarli, G. Ch^ene, C. Dufouil, and S. Seshadri. Assessment of Plasma Total Tau Level as

a Predictive Biomarker for Dementia and Related Endophenotypes. *JAMA Neurol*, 76:598–606, 2019.

[209] T. K. Karikari, T. A. Pascoal, N. J. Ashton, S. Janelidze, A. L. Benedet, J. L. Rodriguez, M. Chamoun, M. Savard, M. S. Kang, J. Therriault, M. Sch.ºll, G. Massarweh, J.-P. Soucy, K. H.ºglund, G. Brinkmalm, N. Mattsson, S. Palmqvist, S. Gauthier, E. Stomrud, H. Zetterberg, O. Hansson, P. Rosa-Neto, and K. Blennow. Blood phosphorylated tau 181 as a biomarker for Alzheimer's disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts. *The Lancet Neurology*, 19:422–433, 2020.

[210] E. H. Thijssen, R. La Joie, A. Wolf, A. Strom, P. Wang, L. Iaccarino, V. Bourakova, Y. Cobigo, H. Heuer, S. Spina, L. VandeVrede, X. Chai, N. K. Proctor, D. C. Airey, S. Shcherbinin, C. D. Evans, J. R. Sims, H. Zetterberg, K. Blennow, A. M. Karydas, C. E. Teunissen, J. H. Kramer, L. T. Grinberg, W. W. Seeley, H. Rosen, B. F. Boeve, B. L. Miller, G. D. Rabinovici, J. L. Dage, J. C. Rojas, and A. L. Boxer. Diagnostic value of plasma phosphorylated tau181 in Alzheimer's disease and frontotemporal lobar degeneration. *Nat Med*, 26:387–397, 2020.

[211] S. Janelidze, D. Bali, N. J. Ashton, N. R. Barth'elemy, J. Vanbrabant, E. Stoops, E. Vanmechelen, Y. He, A. O. Dolado, G. Triana-Baltzer, M. J. Pontecorvo, H. Zetterberg, H. Kolb, M. Vandijck, K. Blennow, R. J. Bateman, and O. Hansson. Head-to-head comparison of 10 plasma phospho-tau assays in prodromal Alzheimer's disease. *Brain*, 146:1592–1601, 2023.

[212] Shannon L. Risacher, Noelia Fandos, Judith Romero, Ian Sherriff, Pedro Pesini, Andrew J. Saykin, and Liana G. Apostolova. Plasma amyloid beta levels are associated with cerebral amyloid and tau deposition. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11(1):510–519, December 2019.

[213] Suzanne E. Schindler, James G. Bollinger, Vitaliy Ovod, Kwasi G. Mawuenyega, Yan Li, Brian A. Gordon, David M. Holtzman, John C. Morris, Tammie L.S. Benzinger, Chengjie Xiong, Anne M. Fagan, and Randall J. Bateman. High-precision plasma β-amyloid 42/40 predicts current and future brain amyloidosis. *Neurology*, 93(17):e1647–e1659, October 2019.

[214] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106–118, February 2013.

[215] Anand Viswanathan and Steven M. Greenberg. Cerebral amyloid angiopathy in the elderly. *Annals of Neurology*, 70(6):871–880, December 2011.

[216] T. R. Everett and L. S. Chitty. Cell-free fetal DNA: the new tool in fetal medicine. *Ultrasound Obstet Gynecol*, 45:499–507, 2015.

[217] G. Tzimagiorgis, E. Z. Michailidou, A. Kritis, A. K. Markopoulos, and S. Kouidou. Recovering circulating extracellular or cell-free RNA from bodily fluids. *Cancer Epidemiol*, 35:580–589, 2011.

[218] Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*, 164(1-2):57–68, January 2016.

[219] Suthee Rapisuwon, Eveline E. Vietsch, and Anton Wellstein. Circulating biomarkers to monitor cancer progression and treatment. *Computational and Structural Biotechnology Journal*, 14:211–222, 2016.

[220] Winston Koh, Wenying Pan, Charles Gawad, H. Christina Fan, Geoffrey A. Kerchner, Tony Wyss-Coray, Yair J. Blumenfeld, Yasser Y. El-Sayed, and Stephen R. Quake. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences*, 111(20):7361–7366, May 2014.

[221] Subodh Kumar and P. Hemachandra Reddy. Are circulating microRNAs peripheral biomarkers for Alzheimer's disease? *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1862(9):1617–1627, September 2016.

[222] Kira S. Sheinerman, Jon B. Toledo, Vladimir G. Tsivinsky, David Irwin, Murray Grossman, Daniel Weintraub, Howard I. Hurtig, Alice Chen-Plotkin, David A. Wolk, Leo F. McCluskey, Lauren B. Elman, John Q. Trojanowski, and Samuil R. Umansky. Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases. *Alzheimer's Research & Therapy*, 9(1):89, December 2017.

[223] Shusuke Toden, Jiali Zhuang, Alexander D. Acosta, Amy P. Karns, Neeraj S. Salathia, James B. Brewer, Donna M. Wilcock, Jonathan Aballi, Mike Nerenberg, Stephen R. Quake, and Arkaitz Ibarra. Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. *Science Advances*, 6(50):eabb1654, December 2020.

[224] Jun-ichi Satoh, Yoshihiro Kino, and Shumpei Niida. MicroRNA-Seq Data Analysis Pipeline to Identify Blood Biomarkers for Alzheimer's Disease from Public Data. *Biomarker Insights*, 10:BMI.S25132, January 2015.

[225] Hyman M. Schipper, Olivier C. Maes, Howard M. Chertkow, and Eugenia Wang. MicroRNA Expression in Alzheimer Blood Mononuclear Cells. *Gene Regulation and Systems Biology*, 1:GRSB.S361, January 2007.

[226] Daniela Galimberti, Chiara Villa, Chiara Fenoglio, Maria Serpente, Laura Ghezzi, Sara M.G. Cioffi, Andrea Arighi, Giorgio Fumagalli, and Elio Scarpini. Circulating miRNAs as Potential Biomarkers in Alzheimer's Disease. *Journal of Alzheimer's Disease*, 42(4):1261–1267, October 2014.

[227] Lin Tan, Jin-Tai Yu, Qiu-Yan Liu, Meng-Shan Tan, Wei Zhang, Nan Hu, Ying-Li Wang, Lei Sun, Teng Jiang, and Lan Tan. Circulating miR-125b as a biomarker of Alzheimer's disease. *Journal of the Neurological Sciences*, 336(1-2):52–56, January 2014.

[228] Mary B. Makarious, Hampton L. Leonard, Dan Vitale, Hirotaka Iwaki, Lana Sargent, Anant Dadu, Ivo Violich, Elizabeth Hutchins, David Saffo, Sara Bandres-Ciga, Jonggeol Jeff Kim, Yeajin Song, Melina Maleknia, Matt Bookman, Willy Nojopranoto, Roy H. Campbell, Sayed Hadi Hashemi, Juan A. Botia, John F. Carter, David W. Craig, Kendall Van Keuren-Jensen, Huw R. Morris, John A. Hardy, Cornelis Blauwendraat, Andrew B. Singleton, Faraz Faghri, and Mike A. Nalls. Multi-modality machine learning predicting Parkinson's disease. *npj Parkinson's Disease*, 8(1):35, April 2022.

[229] Hsiang-Han Chen, Abdallah Eteleeb, Ciyang Wang, Maria Victoria Fernandez, John P. Budde, Kristy Bergmann, Joanne Norton, Fengxian Wang, Curtis Ebl, John C. Morris, Richard J. Perrin, Randall J. Bateman, Eric McDade, Chengjie Xiong, Alison Goate, Martin Farlow, Jasmeer Chhatwal, Peter R. Schofield, Helena Chui, Oscar Harari, Carlos Cruchaga, Laura Ibanez, and Dominantly Inherited Alzheimer Network. Circular RNA detection identifies circPSEN1 alterations in brain specific to autosomal dominant Alzheimer's disease. *Acta Neuropathologica Communications*, 10(1):29, December 2022.

[230] John C. Morris. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, 43(11):2412.2–2412–a, November 1993.

[231] Anne M. Fagan, Mark A. Mintun, Robert H. Mach, Sang-Yoon Lee, Carmen S. Dence, Aarti R. Shah, Gina N. LaRossa, Michael L. Spinner, William E. Klunk, Chester A. Mathis, Steven T. DeKosky, John C. Morris, and David M. Holtzman. Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Aβ_{42} in humans. *Annals of Neurology*, 59(3):512–519, March 2006.

[232] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[233] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, April 2017.

[234] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, October 2016.

[235] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.

[236] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(Web Server):W305–W311, July 2009.

[237] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, January 2017.

[238] Sonia García-Ruiz, Ana L. Gil-Martínez, Alejandro Cisterna, Federico Jurado-Ruiz, Regina H. Reynolds, NABEC (North America Brain Expression Consortium), Mark R. Cookson, John Hardy, Mina Ryten, and Juan A. Botía. CoExp: A Web Tool for the Exploitation of Co-expression Networks. *Frontiers in Genetics*, 12:630187, February 2021.

[239] David A. Bennett, Aron S. Buchman, Patricia A. Boyle, Lisa L. Barnes, Robert S. Wilson, and Julie A. Schneider. Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's Disease*, 64(s1):S161–S189, June 2018.

[240] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[241] Axel Montagne, Zhen Zhao, and Berislav V. Zlokovic. Alzheimer's disease: A matter of blood–brain barrier dysfunction? *Journal of Experimental Medicine*, 214(11):3151–3169, November 2017.

[242] Melanie D. Sweeney, Abhay P. Sagare, and Berislav V. Zlokovic. Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nature Reviews Neurology*, 14(3):133–150, March 2018.

[243] Eric M. Blalock, James W. Geddes, Kuey Chu Chen, Nada M. Porter, William R. Markesbery, and Philip W. Landfield. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences*, 101(7):2173–2178, February 2004.

[244] Adam R. Fenton, Thomas A. Jongens, and Erika L. F. Holzbaur. Mitochondrial adaptor TRAK2 activates and functionally links opposing kinesin and dynein motors. *Nature Communications*, 12(1):4578, July 2021.

[245] Rodrigo A. Quintanilla, Carola Tapia-Monsalves, Erick H. Vergara, María José Pérez, and Alejandra Aranguiz. Truncated Tau Induces Mitochondrial Transport Failure Through the Impairment of TRAK2 Protein and Bioenergetics Decline in Neuronal Cells. *Frontiers in Cellular Neuroscience*, 14:175, July 2020.

[246] Sónia C. Correia, George Perry, and Paula I. Moreira. Mitochondrial traffic jams in Alzheimer's disease - pinpointing the roadblocks. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1862(10):1909–1917, October 2016.

[247] Dominantly Inherited Alzheimer Network, Oliver Preische, Stephanie A. Schultz, Anja Apel, Jens Kuhle, Stephan A. Kaeser, Christian Barro, Susanne Gräber, Elke Kuder-Buletta, Christian LaFougere, Christoph Laske, Jonathan Vöglein, Johannes Levin, Colin L. Masters, Ralph Martins, Peter R. Schofield, Martin N. Rossor, Neill R. Graff-Radford, Stephen Salloway, Bernardino Ghetti, John M. Ringman, James M. Noble, Jasmeer Chhatwal, Alison M. Goate, Tammie L. S. Benzinger, John C. Morris, Randall J. Bateman, Guoqiao Wang, Anne M. Fagan, Eric M. McDade, Brian A. Gordon, and Mathias Jucker. Serum neurofilament

dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer's disease. *Nature Medicine*, 25(2):277–283, February 2019.

[248] N. Mattsson, E. Rosen, O. Hansson, N. Andreasen, L. Parnetti, M. Jonsson, S.-K. Herukka, W. M. Van Der Flier, M. A. Blankenstein, M. Ewers, K. Rich, E. Kaiser, M. M. Verbeek, M. Olde Rikkert, M. Tsolaki, E. Mulugeta, D. Aarsland, P. J. Visser, J. Schroder, J. Marcusson, M. De Leon, H. Hampel, P. Scheltens, A. Wallin, M. Eriksdotter-Jonhagen, L. Minthon, B. Winblad, K. Blennow, and H. Zetterberg. Age and diagnostic performance of Alzheimer disease CSF biomarkers. *Neurology*, 78(7):468–476, February 2012.

[249] Carles Falcon, Alan Tucholka, Gemma C. Monté-Rubio, Raffaele Cacciaglia, Grégory Operto, Lorena Rami, Juan Domingo Gispert, and José Luis Molinuevo. Longitudinal structural cerebral changes related to core CSF biomarkers in preclinical Alzheimer's disease: A study of two independent datasets. *NeuroImage: Clinical*, 19:190–201, 2018.

[250] Yuting Zhang, Upamanyu Ghose, Noel J. Buckley, Sebastiaan Engelborghs, Kristel Sleegers, Giovanni B. Frisoni, Anders Wallin, Alberto Lleó, Julius Popp, Pablo Martinez-Lage, Cristina Legido-Quigley, Frederik Barkhof, Henrik Zetterberg, Pieter Jelle Visser, Lars Bertram, Simon Lovestone, Alejo J. Nevado-Holgado, and Liu Shi. Predicting AT(N) pathologies in Alzheimer's disease from blood-based proteomic data using neural networks. *Frontiers in Aging Neuroscience*, 14:1040001, November 2022.

[251] Lihua Wang, Daniel Western, Jigyasha Timsina, Charlie Repaci, Won-Min Song, Joanne Norton, Pat Kohlfeld, John Budde, Sharlee Climer, Omar H. Butt, Daniel Jacobson, Michael Garvin, Alan R. Templeton, Shawn Campagna, Jane O'Halloran, Rachel Presti, Charles W. Goss, Philip A. Mudd, Beau M. Ances, Bin Zhang, Yun Ju Sung, and Carlos Cruchaga. Plasma proteomics of SARS-CoV-2 infection and severity reveals impact on Alzheimer's and coronary disease pathways. *iScience*, 26(4):106408, April 2023.

[252] D. Galasko, L. A. Hansen, R. Katzman, W. Wiederholt, E. Masliah, R. Terry, L. R. Hill, P. Lessin, and L. J. Thal. Clinical-Neuropathological Correlations in Alzheimer's Disease and Related Dementias. *Archives of Neurology*, 51(9):888–895, September 1994.

[253] Hiroshige Fujishiro, Eizo Iseki, Shinji Higashi, Koji Kasanuki, Norio Murayama, Takashi Togo, Omi Katsuse, Hirotake Uchikado, Naoya Aoki, Kenji Kosaka, Heii Arai, and Kiyoshi Sato. Distribution of cerebral amyloid deposition and its relevance to clinical phenotype in Lewy body dementia. *Neuroscience Letters*, 486(1):19–23, December 2010.

[254] Noritaka Wakasugi and Takashi Hanakawa. It Is Time to Study Overlapping Molecular and Circuit Pathophysiologies in Alzheimer's and Lewy Body Disease Spectra. *Frontiers in Systems Neuroscience*, 15:777706, November 2021.

[255] Alessandro Padovani, Enrico Premi, Andrea Pilotto, Stefano Gazzina, Maura Cosseddu, Silvana Archetti, Vanessa Cancelli, Barbara Paghera, and Barbara Borroni. Overlap between Frontotemporal Dementia and Alzheimer's Disease: Cerebrospinal Fluid Pattern and Neuroimaging Study. *Journal of Alzheimer's Disease*, 36(1):49–55, June 2013.

[256] Zornitza Stark, Rebecca E. Foulger, Eleanor Williams, Bryony A. Thompson, Chirag Patel, Sebastian Lunke, Catherine Snow, Ivone U.S. Leong, Arina Puzriakova, Louise C. Daugherty, Sarah Leigh, Christopher Boustred, Olivia Niblock, Antonio Rueda-Martin, Oleg Gerasimenko, Kevin Savage, William Bellamy, Victor San Kho Lin, Roman Valls, Lavinia Gordon, Helen K. Brittain, Ellen R.A. Thomas, Ana Lisa Taylor Tavares, Meriel McEntagart, Susan M. White, Tiong Y. Tan, Alison Yeung, Lilian Downie, Ivan Macciocca, Elena Savva, Crystle Lee, Ain Roesley, Paul De Fazio, Jane Deller, Zandra C. Deans, Sue L. Hill, Mark J. Caulfield, Kathryn N. North, Richard H. Scott, Augusto Rendon, Oliver Hofmann, and Ellen M. McDonagh. Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. *The American Journal of Human Genetics*, 108(9):1551–1557, September 2021.

[257] Saudi Mendeliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biology*, 16(1):134, December 2015.

[258] Lora Bean, Birgit Funke, Colleen M. Carlston, Jennifer L. Gannon, Sibel Kantarci, Bryan L. Krock, Shulin Zhang, and Pinar Bayrak-Toydemir. Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 22(3):453–461, March 2020.

[259] Michael A. DeTure and Dennis W. Dickson. The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*, 14(1):32, December 2019.

[260] Marwan N. Sabbagh, Lih-Fen Lue, Daniel Fayard, and Jiong Shi. Increasing Precision of Clinical Diagnosis of Alzheimer's Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data. *Neurology and Therapy*, 6(S1):83–95, July 2017.

[261] Jason Weller and Andrew Budson. Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Research*, 7:1161, July 2018.

[262] S. Budd Haeberlein, P.S. Aisen, F. Barkhof, S. Chalkias, T. Chen, S. Cohen, G. Dent, O. Hansson, K. Harrison, C. Von Hehn, T. Iwatsubo, C. Mallinckrodt, C.J. Mummery, K.K. Muralidharan, I. Nestorov, L. Nisenbaum, R. Rajagovindan, L. Skordos, Y. Tian, C.H. Van Dyck, B. Vellas, S. Wu, Y. Zhu, and A. Sandrock. Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease. *The Journal of Prevention of Alzheimer's Disease*, 2022.

[263] Rajesh R Tampi, Brent P Forester, and Marc Agronin. Aducanumab: evidence from clinical trial data and controversies. *Drugs in Context*, 10:1–9, October 2021.

[264] Adriane Dallanora Henriques, Andrea Lessa Benedet, Einstein Francisco Camargos, Pedro Rosa-Neto, and Otávio Toledo Nóbrega. Fluid and imaging biomarkers for Alzheimer's disease: Where we stand and where to head to. *Experimental Gerontology*, 107:169–177, July 2018.

[265] Justin M Long, Dean W Coble, Chengjie Xiong, Suzanne E Schindler, Richard J Perrin, Brian A Gordon, Tammie L S Benzinger, Elizabeth Grant, Anne M Fagan, Oscar Harari, Carlos Cruchaga, David M Holtzman, and John C Morris. Preclinical Alzheimer's disease biomarkers accurately predict cognitive and neuropathological outcomes. *Brain*, 145(12):4506–4518, December 2022.