Epidemiología oculta: exploración de las enfermedades raras en Europa y su impacto regional en España

o

Estudio descriptivo de las enfermedades raras en Europa y su impacto regional en España mediante datos públicos

## Resumen

## Introducción

¿Qué son las enfermedades raras? ¿Por qué son relevantes?

¿Cuáles son las dificultades del diagnóstico y tratamiento?

¿Cómo es su investigación? Hitos históricos.

¿Cuál es el papel del análisis de datos en este contexto?

\*\* Objetivo del trabajo: preguntas y/o hipótesis.

En Europa, se define enfermedad rara (ER) o poco frecuente como aquella que, con peligro de muerte o invalidez crónica, tiene una prevalencia menor de 5 casos por cada 10.000 habitantes1.

El Registro Estatal de Enfermedades Raras (ReeR) está formado por los registros autonómicos, encargados de transmitir los datos al primero.

Objetivo:

* Explorar y caracterizar la situación de las enfermedades raras en Europa y España a partir de datos públicos (Orphadata, ReeR, EMEA, PubMed), con el fin de identificar patrones epidemiológicos, genéticos y clínicos relevantes para su investigación y gestión sanitaria.

Preguntas e hipótesis:

* Las enfermedades raras con mayor número de fenotipos (HPO) tienden a tener mayor prevalencia, posiblemente porque son más fácilmente detectables o mejor documentadas.
* Las enfermedades raras del sistema nervioso son las más prevalentes, dado su impacto clínico y la frecuencia de estudio en investigaciones biomédicas.
* ¿Qué regiones de España presentan mayor carga de enfermedades raras según los informes ReeR?
* Las comunidades autónomas con mayor población no siempre concentran más diagnósticos, lo cual puede deberse a diferencias en el acceso al diagnóstico o registro.

## Metodología

Fuente de datos:

* API Orphadata
* European Medicines Agency (EMEA) Orphan designation
* Informes ReeR (Registro estatal enfermedades raras España) 2021 – 2024
* Publicaciones científicas de PubMed que contienen “rare disease”

Procesamiento de datos: limpieza + EDA + preprocesado

Herramientas utilizadas: Python (pandas, matplotlib/seaborn/plotly), Power BI, MySQL

Pipeline, filtros, etc.

**Fuente de datos**

**Herramientas utilizadas**

El desarrollo del proyecto se llevó a cabo utilizando Python como lenguaje de programación (versión 3.13.2), con Jupyter Notebook como framework principal para la ejecución interactiva del código y Visual Studio Code como entorno de programación.

Los análisis se realizaron empleando las siguientes bibliotecas:

* *pandas* para manipulación de datos tabulares
* *numpy* para operaciones numéricas
* *matplotlib* y *seaborn* para visualizaciones estáticas
* *plotly* para visualizaciones interactivas
* *scikit-learn* para análisis exploratorio y preprocesamiento
* *requests* y *xml.etree.ElementTree* para acceso y procesamiento de datos obtenidos por API

Todo el desarrollo se realizó en un entorno local, ejecutado sobre sistema operativo Windows 64 bits.

## Resultados

Resultados obtenidos, sin interpretación.

Tablas y gráficos: prevalencia, localización geográfica, edad, etc.

Distribución de enfermedades por región, clasificación según sistema afectado, según gen implicado, etc.

Evaluación según enfoque: epimediológicas/genéticas.

**Datos de Orphadata**

Para obtener datos de las enfermedades raras de Orphanet, se accedió a la API y se consiguió guardar nueve DataFrames en formato .csv.

Se estudió cuidadosamente el contenido de cada uno de estos DataFrames, se eliminaron las columnas innecesarias y se reorganizaron las otras, dando lugar a la siguiente base de datos conteniendo las columnas indicadas en la Figura 1

A computer screen shot of a computer

AI-generated content may be incorrect.

Figura 1. Diagrama relacional de la base de datos obtenida de Orphadata.

**Datos del registro estatal de enfermedades raras (ReeR)**

## Discusión

Análisis de los resultados.

Buscar patrones, tendencias y su significado.

Analizar si existe información suficiente para sacar conclusiones. Limitaciones del Dataset.

Analizar las implicaciones que tienen estos resultados para políticas de salud y para la investigación en este ámbito.

## Conclusión

Resumir los puntos clave de la investigación.

¿Qué aporta este trabajo?

Posibles mejoras y futuras líneas de investigación.

## Referencias

## Extra:

1. **Clasificación de enfermedades por sistema afectado (modelo supervisado)**

Objetivo: Dado un conjunto de características (por ejemplo, prevalencia, edad de aparición, tipo de herencia, síntomas), predecir a qué sistema del cuerpo afecta la enfermedad (neurológico, inmunológico, etc.).

Modelo: Árboles de decisión, Random Forest o XGBoost.

Datos necesarios: Variables categóricas/numéricas limpias, codificadas.

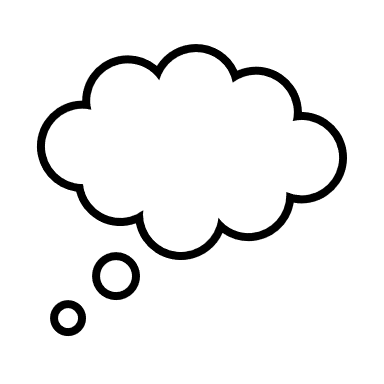
1. **Clustering (modelo no supervisado)**

Objetivo: Agrupar enfermedades raras con características similares (sin usar una variable objetivo).

Modelo: K-means, DBSCAN o PCA para visualización.

Identificar clusters de enfermedades poco estudiadas pero con rasgos comunes.

Explorar patrones ocultos que no se ven a simple vista.

ANÁLISIS FUTUROS

Analizar la interacción entre el fármaco cannabidiol (CBD) y la proteína (Ras/Rap GTPase-activating protein SynGAP) del gen SYNGAP1.

* ¿Hay datos que indiquen la existencia de una relación directa entre CBD y SYNGAP1?
* ¿Hay otras proteínas en el gen SYNGAP1 que puedan interaccionar con el CBD?
* ¿Qué efectos biológicos o rutas de señalización comparten el CBD y SYNGAP1?

¿Qué hacer?

1. Consultar bases de datos bioinformáticas para buscar interacciones existentes.
2. Analizar redes de proteínas para buscar unas similares a la de SYNGAP1.
3. Analizar la similitud química entre CBD y otros compuestos que interactúan con SYNGAP1.
4. Analizar enriquecimiento funcional para ver si están involucrados en rutas comunes.

Paso a paso:

1. STITCH (Search Tool for Interactions of Chemicals) y ChEMBL para verificar las interacciones CBD – SYNGAP1. [STITCH: chemical association networks](http://stitch.embl.de/)
2. STRING para encontrar proteínas similares a la de SYNGAP1.
3. Ver si hay otras proteínas similares a SYNGAP1. En ChEMBL o SwissTargetPrediction se puede buscar si existe interacción entre ellas y el CBD
4. Buscar fármacos que interactúen con SYNGAP1 y ver si son químicamente similares al CBD con RDKit.