MAPEO SISTEMATICO

"Estudio de técnicas de machine learning y puntuación de riesgo poligénico para predecir la enfermedad de Alzheimer"

Mapeo sistemático

El mapeo sistemático se llevó a cabo siguiendo el proceso descrito por [1] y [2] que proponen pautas y directrices para el mapeo y revisiones sistemáticas. El mapeo sistemático tiene el propósito identificar los vacíos de conocimiento y las necesidades de investigación sobre la predicción de la EA. Por lo tanto, presenta tres etapas principales, en la primera etapa la planificación, en el cual se definen los objetivos principales y las preguntas de investigación a responder. La segunda etapa, se genera la estrategia de búsqueda, la selección, evaluación y extracción de datos de los artículos seleccionados. La tercera etapa se presentan los resultados obtenidos que responden a las preguntas de investigación expuestas en la primera etapa.

1.1.1 Planificación

En esta primera etapa de planificación, se define los objetivos principales del mapeo y las preguntas de investigación que se debe responder con los resultados del mapeo.

A. Objetivos:

El propósito de este estudio secundario, es determinar el alcance de las investigaciones realizadas sobre la predicción temprana de la EA utilizando diversas técnicas de aprendizaje automático y la puntuación de riesgo poligénica. Este estudio se basa en el análisis de investigaciones previas para luego clasificar los trabajos recientes de investigación y ver si la comunidad científica planteó inquietudes similares.

B. Preguntas de Investigación:

Estas preguntas de investigación no son las preguntas de búsqueda que se realizan a las bases de datos, estas preguntas se basan en el objetivo del mapeo definido previamente. Si bien el objetivo general de este estudio se puede resumir en comprender el uso de técnicas de ML y PRS para predecir la EA, este objetivo se divide en cuatro preguntas de investigación concretas para obtener un conocimiento más detallado y una visión integra del tema.

1. ¿Qué técnicas y/o métodos se aplicaron en los estudios?

2. ¿Cuáles son los tipos de datos que se utilizaron en los estudios?

3. ¿Qué limitaciones presentaron los artículos?

4. ¿Qué base de datos utilizaron los artículos para obtener el conjunto de datos?

1.1.2 Estrategia de búsqueda

La estrategia de búsqueda se elabora teniendo en cuenta el problema de terminología existente respecto a la predicción de la EA. Con el objetivo definido para esta búsqueda se tiene la intersección de dos líneas de investigación, la primera la predicción de la EA por medio de un método tradicional PRS y la segunda a través de técnicas de ML.

C. Base de datos:

Los motores de búsqueda seleccionados (de acuerdo al contexto de investigación) para realizar la búsqueda de artículos son PubMed Central, misma tiene una amplia cobertura de publicaciones del área de salud y las bases de datos Scopus y Elsevier Science Direct que indexan el campo de la salud y tecnología.

Cadena de búsqueda: “machine learning and polygenic risk score and alzheimer”.

D. Selección de estudios - Criterios de inclusión y exclusión

Para la selección de estudios se diseñaron un conjunto de criterios de inclusión (IC) y un conjunto de los criterios de exclusión (CE).

CI 1: El artículo tenía al menos uno de los términos "aprendizaje automático", "puntaje de riesgo poligénico" and "Alzheimer" en el título o resumen.

CI 2: El artículo informó sobre la aplicación de técnicas de ML para la predicción de alzheimer

CI 3: Los artículos fueron escritos en inglés

CI 4: Los artículos fueron publicados en revistas.

CI 5: El texto completo del artículo estaba disponible

Los artículos se excluyeron si se cumplían los siguientes criterios:

CE 1: El artículo no está dirigido a la predicción de alzheimer

CE 2: El artículo no presenta la aplicación de alguna técnica de ML o PRS

CE 3: Los artículos no fueron escritos en inglés

CE 4: Los artículos no fueron publicados en revistas.

CE 5: El texto completo del artículo no está disponible

CE 6: El artículo no utiliza datos genéticos

Fases de revisión:

Una vez definido el objetivo del mapeo, las preguntas de investigación, la cadena de búsqueda en las diferentes bases de datos y los criterios de selección, se procede a realizar la revisión de los artículos. Para ello se elabora el diagrama de flujo PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [3] para representar las cuatro fases de la revisión (Figura 2).

Fases de la revisión:

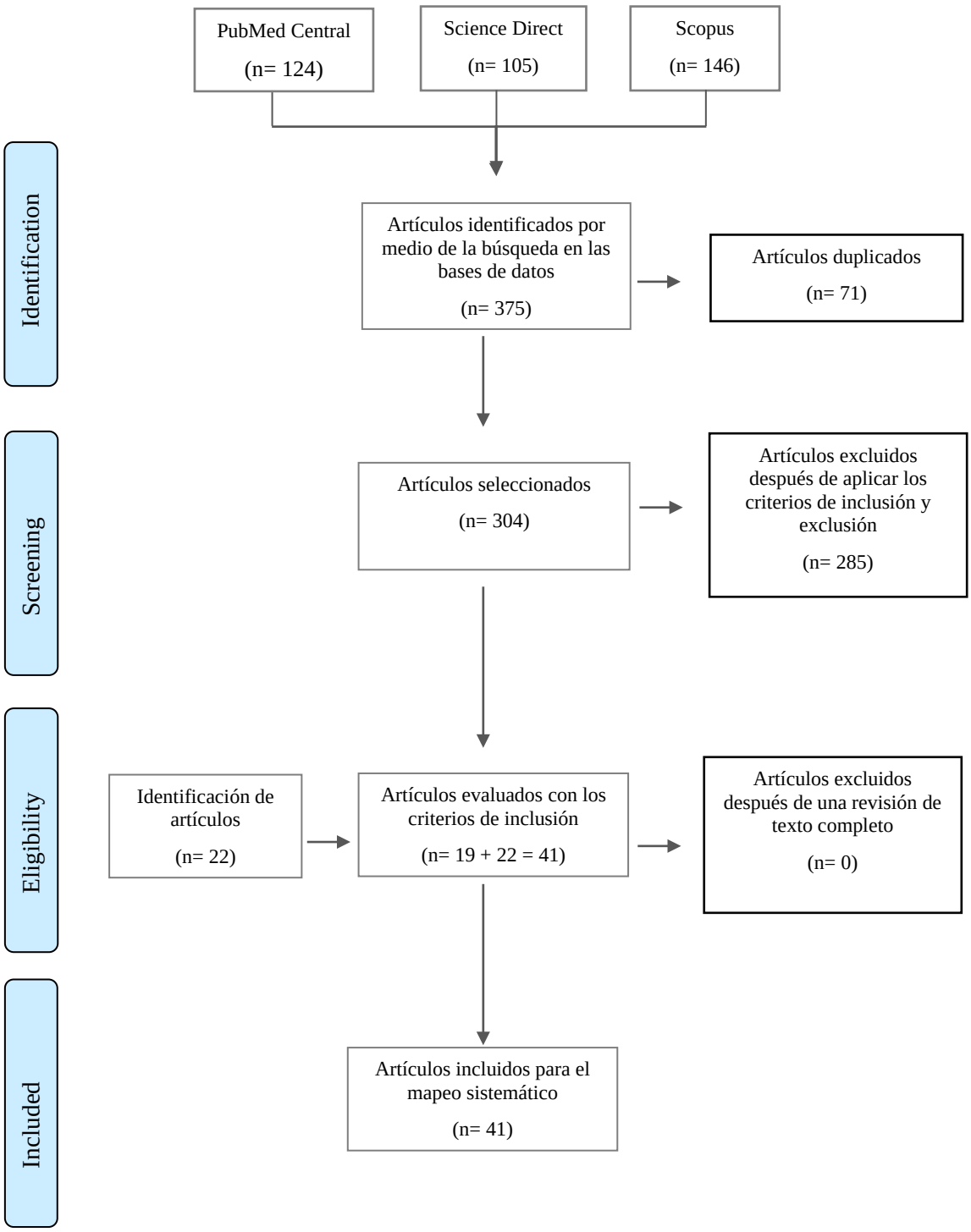

Figura 1: Diagrama de flujo PRISMA [66] representación de las fases de la revisión

E. Extracción de datos

Primeramente, para describir la extracción de datos se utilizó el diagrama de flujo de PRIMA [3] como se observa en la Figura 1. Para proceder con las fases de la revisión se utilizó la herramienta https://parsif.al/ donde se definió la cadena de búsqueda en las diferentes bases de datos, la revisión de duplicados, y la selección de artículos según los criterios de inclusión y exclusión. Posteriormente a la revisión completa de los artículos, se identificó 22 artículos relevantes a través del seguimiento referencias y citados. Toda esta información se encuentra almacenada en un repositorio de GitHub de acceso libre (https://tinyurl.com/y9vwnzhu).

Sintetizando el diagrama de flujo de PRISMA, se recogieron 375 artículos en total como resultado de la búsqueda en las 3 bases de datos, donde:

- Se eliminaron 71 artículos duplicados
- Se eliminaron 285 tras aplicar los criterios de inclusión y exclusión.
- Después de la lectura de texto completo y la profunda evaluación de los artículos, se identificaron 22 artículos a través de las referencias y citados que fueron agregados, teniendo en total 41 artículos para el mapeo sistemático.

1.1.3 Resultados

Luego de concluir con la búsqueda y selección de artículos, se procede a responder las preguntas de investigación planteadas en la subsección 3.5.1 con la información extraída de cada artículo.

I. ¿Qué técnicas y/o métodos se aplicaron en los estudios?

La pregunta tiene como objetivo analizar los métodos y/o técnicas utilizados para construir los modelos de predicción para la EA, en tal tabla 2 se presenta la lista de técnicas y/o métodos usados por los artículos. Se observa que los métodos más utilizados: Support vector Machine (SVM) y Random Forest.

Tabla 1: Artículos agrupado por técnicas

Técnica – Método	Artículos	Total
Algoritmos genéticos	[68], [87], [106]	3
LASSO	[68], [72], [73], [76], [107]	5
Step-wise (BSWiMS)	[68], [76]	2
Gradient boosted eXtrame, XG, Light,	[67], [105]	2
AdaBoost	[97]	1
Radial basis function (RBF)	[88]	1
Kernel-based extreme (KELM)	[89]	1
Support vector Machine (SVM)	[40], [67], [72], [76], [87], [88], [89], [91], [95], [96], [97], [98], [101], [102], [103], [105], [107]	17
Support vector classifier (SVC)	[74]	1
Support vector regressor (SVR)	[69], [74]	2
Multiple kernel learning (MKL)	[72]	1
Generalized linear models (GLM)	[72]	1
Naïve Bayesian	[69], [73]	2
k-Nearest Neighbors (kNN)	[76], [101], [103]	3
Elastic Net	[72], [101]	2
Bayesian network (BN)	[75]	1
Gaussian processes (GP)	[67], [86], [101]	3

Logistic Regression	[21], [40], [68], [70], [71], [72], [78], [79], [80], [81], [82], [85], [86], [93], [94], [104], [106]	17
Random Forest	[40], [67], [69], [76], [79], [90], [95], [99], [103], [104]	10
Linear Regression	[77], [80], [81], [82], [84], [83], [88], [96]	8
Neural Network Artificial, Convolutional, Recurrent	[40], [69], [72], [92], [100], [105]	6
Cross-Validation	[67], [70], [72], [74], [77], [78], [86], [89], [91], [98], [101], [106], [107]	13

II. ¿Cuáles son los tipos de datos que se utilizaron en los estudios?

Se tiene diferentes tipos de datos que utilizaron los autores para realizar sus respectivos estudios. En la tabla 3, se presenta la agrupación de los estudios por tipo de dato. Algunos estudios utilizaron los 3 tipos de datos debido a las técnicas que emplearon.

Tabla 2: Artículos agrupados por tipo de datos

Tipo de dato	Artículos	Total
Datos genéticos (GWAS, exoma, genotipo, fenotipo, etc.)	[21], [40], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107]	43
Datos demográficos	[67], [70], [74], [77],[78], [80], [81], [82], [83], [86], [88], [89], [90], [95], [96], [99],[100], [104],[107]	19
Imágenes de resonancia magnética (MRI)	[72], [74], [76], [77], [80], [83], [87], [88], [89],[90],[91], [95], [96], [97], [98], [99],[100], [103], [107]	20

III. ¿Qué limitaciones presentaron los artículos?

Durante la revisión de los artículos se identificaron 2 limitantes comunes que se muestran en la tabla 4, hubo algunos artículos no reportaron limitaciones.

Tabla 3: Limitaciones de los estudios del mapeo

Limitaciones	Artículos	Total
Métodos aplicados	[95], [101], [102]	3
Conjunto de Datos:	[21], [40],[68], [69], [70], [71], [72], [73], [74], [75], [76], [77],[78], [80], [81], [82], [83], [84], [85], [86], [87], [88], [90],[91],[94], [96], [97], [99], [100], [102], [103], [104], [106]	33

IV. ¿Qué base de datos utilizaron los artículos para obtener el conjunto de datos?

La disponibilidad de datos es un tema muy importante en este tipo de estudios, ya que la cantidad del conjunto de datos puede afectar directamente al resultado eficiente de un estudio. Por ello se vio necesario identificar que bases de datos más utilizadas por los estudios. En la tabla 5 se presenta la lista de bases de datos y los artículos clasificados, donde se puede observar que la base de datos comúnmente usada es la ADNI.

Tabla 4: Bases de datos para la obtención del conjunto de datos

Base de Datos	Artículos	Total
National Institute on Aging	[68]	1
GTEx data	[40]	1
National Center for Biotechnology Information (NCBI)	[68]	1
Alzheimer's Disease Neuroimaging Initiative (ADNI)	[67], [70], [72], [74], [76], [77], [82], [83], [85], [86], [87], [88], [89], [90], [91], [95], [96], [97], [98], [99], [100], [103], [107]	23
National Institute of Biomedical Imaging and Bioengineering (NIBIB)	[67], [86], [91]	3
Rush Alzheimer's Disease Center (RADC)	[75]	1
University of California, San Francisco Memory and Aging Center (UCSF MAC)	[79]	1
Cache County population-based study.	[93]	1
Avon Longitudinal Study of Parents and Children ALSPAC	[70], [84]	2
Alzheimer's Disease Genetics Consortium (ADGC)	[21]	1
AlzGene.org	[93]	1
International Genomics of Alzheimer's Project (IGAP)	[21], [70], [71], [77], [78], [80], [81], [82], [84], [85], [94]	10
Genetic and Environmental Risk for Alzheimer's disease (GERAD)	[21], [71], [78], [94]	4
Memory Disorders at Mount Sinai Medical Center, Alzheimer's Disease Research Center (ADRC)	[101]	1
Layton Aging and Alzheimer's Disease Center and the Oregon Center for Aging and Technology Research Repository	[102]	1
longitudinal clinical pathology cohort study of AD (ROS and MAP)	[104]	1
Ethics Committee of the Medical Faculty, University, Kiel.	[105]	1
Australian Imaging, Biomarkers and Lifestyle (AIBL)	[106]	1

Bibliografía

[1] García-Peñalvo, F. J. (2019). Revisiones y mapeos sistemáticos de literatura. Salamanca, España: Grupo GRIAL. Disponible en: https://goo.gl/yt7wKt.

[2] Petersen, K., Vakkalanka, S., Kuzniarz, L., (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information Software Technology* 64, 1-18

[3] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. In *PLoS Medicine* (Vol. 6, Issue 7). Public Library of Science. https://doi.org/10.1371/journal.pmed.1000097