Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Máster Universitario en Innovación Digital

Complex Data in Health

Studying Alzheimer's Disease

Authors:

Emanuele Alberti
Leandro Duarte
Ottavia Biagi

December 2025

Contents

Introduction

For this assignment, we were assigned to study **Alzheimer's Disease (AD)**. The following identifiers were provided: UMLS CUI `C0002395`, NCIt code `C2866`, recommended protein Apolipoprotein E (UniProt: `P02649`), transcriptome dataset `GSE300079`, and image dataset from OASIS (Kaggle).

1 Medical Terminologies and Semantic Datasets

1.1 Disease General Information

We explored biomedical databases to gather standardized information about Alzheimer's Disease.

NCBI MedGen (UID: 1853) provides the following definition: *"A degenerative disease of the brain that causes dementia, which is a gradual loss of memory, judgment, and ability to function. This disorder usually appears in people older than age 65, but less common forms of the disease appear earlier in adulthood."* The disease is classified as *Disease or Syndrome* with concept ID `C0002395`. The directly associated gene is `APP` (21q21.3), with related genes including `APOE`, `PSEN1`, `PSEN2`, `ABCA7`, `MPO`, and `PLAU`.

NCBI MeSH (ID: D000544) describes AD as a degenerative brain disease with insidious onset of dementia, impairment of memory, judgment, and attention span, followed by apraxias and global loss of cognitive abilities. Pathologically, it is marked by senile plaques, neurofibrillary tangles, and neuropil threads.

ICD-10 classifies AD under code `G30`, with subtypes: `G30.0` (early onset, before age 65), `G30.1` (late onset, after age 65), `G30.8` (other), and `G30.9` (unspecified).

Orphanet (ORPHA:1020) describes Early-onset autosomal dominant Alzheimer disease (EOAD), representing less than 1% of all AD cases, caused by mutations in `PSEN1` (69%), `APP` (13%), or `PSEN2` (2%).

Table 1 summarizes the disease codifications across vocabularies.

Table 1: Alzheimer's Disease codifications across medical vocabularies.

Vocabulary	Code
UMLS CUI	C0002395
MedGen UID	1853
MeSH	D000544
ICD-10	G30
SNOMED CT	26929004
NCIt	C2866
OMIM	104300, 516000
Orphanet	ORPHA:238616, ORPHA:1020
MONDO	MONDO:0004975
HPO	HP:0002511

1.2 SPARQL Query Results

We queried the NCIt SPARQL endpoint (`https://shared.semantics.cancer.gov/sparql`) to retrieve annotation properties for Alzheimer's Disease (code `C2866`). The query extracted the preferred label, synonyms, definition, semantic type, and UMLS CUI.

Summary of SPARQL Findings: The query retrieved annotation properties for NCIt code C2866. The disease is classified under semantic type "Mental or Behavioral Dysfunction" and maps to UMLS CUI C0002395. Seven synonyms were identified, including "Alzheimer's Dementia", "Alzheimer Disease", and "Alzheimer dementia". The definition describes AD as a progressive neurodegenerative disease characterized by nerve cell death leading to loss of cognitive function such as memory and language.

2 Bioinformatics

2.1 Disease Genes

We identified genes with established roles in Alzheimer's Disease using NCBI databases. These genes are categorized by their influence on Early-Onset (familial) versus Late-Onset (sporadic) AD.

APP (Amyloid Precursor Protein) – NCBI Gene ID: 351, chromosome 21q21.3. This gene encodes a transmembrane precursor protein cleaved by secretases. Cleavage by β- and γ-secretases produces amyloid-beta (Aβ) peptides. Aβ accumulation is recognized as a key factor in AD pathogenesis: these peptides aggregate to form amyloid plaques in AD brains. The protein also has antimicrobial properties with bacteriocidal and antifungal activities. Mutations in APP cause autosomal dominant Alzheimer disease and cerebroarterial amyloidosis.

PSEN1 (Presenilin 1) – NCBI Gene ID: 5663, chromosome 14q24.2. This gene encodes a catalytic subunit of the γ-secretase complex, the enzyme that cleaves APP. Mutations alter enzyme function, increasing the ratio of the more toxic Aβ42 peptide versus Aβ40. PSEN1 mutations are the most common cause of Early-Onset Familial AD (69% of EOAD cases per Orphanet). Presenilins are also involved in cleavage of the Notch receptor.

APOE (Apolipoprotein E) – NCBI Gene ID: 348, chromosome 19q13.32. This gene encodes a protein involved in lipid transport and cholesterol metabolism in the brain. It binds to liver and peripheral cell receptors and is essential for catabolism of triglyceride-rich lipoproteins. Three isoforms exist: ε2, ε3, and ε4. The ε4 allele is the strongest genetic risk factor for Late-Onset AD. The risk effect is estimated at 3-fold for heterozygotes (ε3/ε4) and 15-fold for homozygotes (ε4/ε4). A heterozygote has approximately 10%-20% chance of developing AD by age 75; a homozygote has 25%-35% risk. The ε2 allele is considered protective.

Gene Sequence: Following the assignment recommendation to focus on Apolipoprotein E, we retrieved its mRNA sequence from NCBI Nucleotide (RefSeq: `NM_000041.4`). The FASTA sequence is provided in Appendix A.

A APOE Gene FASTA Sequence

>NM_000041.4 Homo sapiens apolipoprotein E (APOE), transcript variant 2, mRNA
CTACTCAGCCCCAGCGGAGGTGAAGGACGTCCTTCCCCAGGAGCCGACTGGCCAATCACAGGCAGGAAGA
TGAAGGTTCTGTGGGCTGCGTTGCTGGTCACATTCCTGGCAGGATGCCAGGCCAAGGTGGAGCAAGCGGT
GGAGACAGAGCCGGAGCCCGAGCTGCGCCAGCAGACCGAGTGGCAGAGCGGCCAGCGCTGGGAACTGGCA
CTGGGTCGCTTTTGGGATTACCTGCGCTGGGTGCAGACACTGTCTGAGCAGGTGCAGGAGGAGCTGCTCA
GCTCCCAGGTCACCCAGGAACTGAGGGCGCTGATGGACGAGACCATGAAGGAGTTGAAGGCCTACAAATC
GGAACTGGAGGAACAACTGACCCCGGTGGCGGAGGAGACGCGGGCACGGCTGTCCAAGGAGCTGCAGGCG
GCGCAGGCCCGGCTGGGCGCGGACATGGAGGACGTGTGCGGCCGCCTGGTGCAGTACCGCGGCGAGGTGC
AGGCCATGCTCGGCCAGAGCACCGAGGAGCTGCGGGTGCGCCTCGCCTCCCACCTGCGCAAGCTGCGTAA
GCGGCTCCTCCGCGATGCCGATGACCTGCAGAAGCGCCTGGCAGTGTACCAGGCCGGGGCCCGCGAGGGC
GCCGAGCGCGGCCTCAGCGCCATCCGCGAGCGCCTGGGGCCCCTGGTGGAACAGGGCCGCGTGCGGGCCG
CCACTGTGGGCTCCCTGGCCGGCCAGCCGCTACAGGAGCGGGCCCAGGCCTGGGGCGAGCGGCTGCGCGC
GCGGATGGAGGAGATGGGCAGCCGGACCCGCGACCGCCTGGACGAGGTGAAGGAGCAGGTGGCGGAGGTG
CGCGCCAAGCTGGAGGAGCAGGCCCAGCAGATACGCCTGCAGGCCGAGGCCTTCCAGGCCCGCCTCAAGA
GCTGGTTCGAGCCCCTGGTGGAAGACATGCAGCGCCAGTGGGCCGGGCTGGTGGAGAAGGTGCAGGCTGC
CGTGGGCACCAGCGCCGCCCCTGTGCCCAGCGACAATCACTGAACGCCGAAGCCTGCAGCCATGCGACCC
CACGCCACCCCGTGCCTCCTGCCTCCGCGCAGCCTGCAGCGGGAGACCCTGTCCCCGCCCCAGCCGTCCT
CCTGGGGTGGACCCTAGTTTAATAAAGATTCACCAAGTTTCACGCA