UNIVERSITY OF TURKU

Recognizing early signs of dementia utilizing [18F]-fluorodeoxyglucose positron emission tomography brain imaging data and machine learning

Molecular Systems Biology/Department Life Technologies

Master's thesis

Author:
Veikka Savila

6.6.2025

Turku

Master's thesis

Subject: Molecular Systems Biology
Author: Veikka Savila
Title: Recognizing early signs of dementia utilizing [18F]-fluorodeoxyglucose positron emission tomography brain imaging data and machine learning
Supervisor(s): Doc. Marco Bucci, MSc Janne Isojärvi
Number of pages: 69 pages
Date: 6.6.2025

Artificial intelligence, powered by advanced machine learning models, has revolutionized scientific research. In parallel, available neuroimaging data has increased. However, the utilization of machine learning models to analyse the data has been limited to small datasets, producing only directional results. Furthermore, processing of the positron emission tomography (PET) imaging data has varied, generating results inapplicable for cross-comparison. This thesis aimed to provide a reproducible and generalizable analysis of [^{18}F]-Fluorodeoxyglucose (FDG) PET images with focus on revealing underlying changes in brain metabolism preceding dementia. Magia, a standardized pipeline for analysis of the PET scans, was utilized, as well as machine learning models on the data acquired from Magia, for feature extraction and classification between diseased and healthy subjects.

This thesis applied methods from previous studies in the fields of neuroimaging, neuroinformatics, and machine learning. The PET imaging data, with available memory problem diagnoses, was extracted from AIVO database of Turku PET Centre. The Magia pipeline for image analysis utilized three PET image quantification methods: 1) standardized uptake value, 2) fractional uptake rate, and 3) linear regression Patlak plot. The results of the three quantification methods in combination with the diagnostic data from AIVO were used in the training of three separate machine learning models. The machine learning models were used to classify between A) healthy controls, B) other memory problems, C) memory disorders, and D) other subjects.

The produced prediction models were able to outperform an uninformed guess but performed worse than models in prior studies. Achieving robust prediction models proved challenging for example due to misplaced data. Furthermore, due to underrepresentation of subjects within groups during the training of the models, the prediction models exhibited poor sensitivity. Multiple points of improvement were characterized during the study. Further research in the topic is required. A robust prediction model relying on PET image derived data rather than the images themselves will decrease the required resources in the study of memory problems. Over and above that, the features extracted will contribute to a better understanding of the pathophysiology of neurodegeneration.

Key words: Alzheimer’s disease, automated neuroimaging data analysis, brain scan, dementia, [^{18}F] FDG PET brain imaging, machine learning, neurodegeneration, pipeline

Contents of the thesis

Abbreviations

^{1}H-MRS - Proton Magnetic Resonance Spectroscopy

[^{18}F]FDG-PET - [^{18}F]-Fluorodeoxyglucose Positron Emission Tomography

AD - Alzheimer’s Disease

ADNI - Alzheimer's Disease Neuroimaging Initiative

AI - Artificial Intelligence

AIF - Arterial Input Function

APOE - Apolipoprotein E (gene)

APP - Amyloid Precursor Protein

Aβ - Amyloid Beta

BGU - Brain Glucose Uptake

BLS - Broad Learning System

BMI - Body Mass Index

CERAD - Consortium to Establish a Registry for Alzheimer's Disease

CNN - Convolutional Neural Network

CRP - C-Reactive Protein

CSF - Cerebrospinal Fluid

DL - Deep Learning

EMCI - Early MCI

FDG - Fluorodeoxyglucose

fMRI - functional MRI

FTD - Frontotemporal Dementia

FUR - Fractional Uptake Rate

GFAP - Glial Fibrillary Acidic Protein

Glx - Glutamate/Glutamine

HC - Healthy Control

IDIF - Image Derived Input Function

IF - Input Function

IL-6 - Interleukin 6

IQR - Interquartile Range

IR - Input Recovery

K_i - Net influx rate

LMCI - Late MCI

MCI - Mild Cognitive Impairment

MI - Myo-Inositol

ML - Machine Learning

MLM - Machine Learning Model

MMSE - Mini-Mental State Examination

MRI - Magnetic Resonance Imaging

MS - Multiple Sclerosis

MTL - Medial Temporal Lobe

NA - Not Applicable

NAA - N-Acetylaspartate

NIR - No Information Rate

PD - Parkinson's Disease

PI - Principal Investigator

PiB - [^{11}C] Pittsburgh compound B

PET - Positron Emission Tomography

PNFA - Progressive Non-Fluent Aphasia

ROI - Region of Interest

SD - Semantic Dementia

SUV - Standardized Uptake Value

SVM - Support Vector Machine

T2D - Type 2 Diabetes

t-SNE - t-Distributed Stochastic Neighbour Embedding

TDP-43 - TAR DNA-Binding Protein 43

TNF-α - Tumour Necrosis Factor Alpha

TSPO - 18-kDa Translocator Protein

VD - Vascular Dementia

1 Introduction

1.1 Thesis Overview

As the population continues to age due to rise in standards of living and to live older especially due to medical advancements, the prevalence of degenerative neuronal disorders increases. Furthermore, as the treatment of these disorders currently limits only to mitigate the symptoms rather than curing the disease altogether, new insight in the disease pathology is required. Scientists from different fields and their research all provide valuable information in defining the intricate systems underlying cognitive functions such as memory, language, and executive functions. Psychologists have defined tests to measure the cognitive skills, psychiatrists have studied the connection of medication and cognitive symptoms, and neuroscientists have drawn atlases of the brain, connecting the cognitive functions to their neural pathways. Imaging methods, like Magnetic Resonance Imaging (MRI), have been developed to study the anatomical changes of the brain, and methods, like [^{18}F]-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET), to study the metabolic activity in brain tissue. Through an interdisciplinary approach, this valuable research data can be compiled, and with the aid of Machine Learning Models (MLMs) this large amount of data can be analysed. The data analysis can be utilised to discover new patterns in the pathophysiology of the degenerative neuronal disorders.

1.2 Dementia and Memory Disorders

Dementia is an umbrella term for symptoms affecting cognitive functions such as memory, language, and executive functioning (Risacher and Saykin 2013). These symptoms correspond to neurodegeneration, neuronal loss, in certain areas of the brain associated with said tasks. Various diseases cause dementia, such like Alzheimer's disease (AD), which is characterized by the accumulation of amyloid plaques, which are aggregations of the amyloid β (Aβ) peptide within brain tissue (Risacher and Saykin 2013), which in turn precedes the neuronal loss. Other diseases causing dementia include vascular dementia (VD), frontotemporal dementia (FTD), and Parkinson's disease (PD). The pathology of these diseases differs significantly, even though they result in similar memory

symptoms. However other cognitive symptoms vary between these disorders greatly.

Memory disorders on the other hand, are common in the general population. They can be the result of aging, brain injuries, stress, depression and neurological conditions like AD. With conditions like stress or depression for example, the memory issues may be reversible, while AD and brain injuries cause more lasting changes in the brain, and therefore longer lasting memory issues.

1.3 Different Types of Dementia and Neuroimaging

1.3.1 Differentiation between Dementias

Neurodegenerative disorders are characterized by the death of neurons. This loss of neurons results in decline in cognitive functions. Although cognitive symptoms may overlap between different diseases, some of these disorders differ in their symptoms, and furthermore, neuronal imaging methods let us differentiate between the diseases, by localizing the changes in brain tissue. The location of the changes usually correlates with the symptoms, as different brain functions require certain regions. It is also possible, that the imaging findings may not match with the symptoms; cases are known, where cognitively healthy patients' brains have been severely affected by accumulation of protein aggregates (Jansen *et al*. 2022).

Changes in the anatomy of the brain can be detected with MRI; changes such as differences in size and composition of the brain regions. While with PET, changes in the metabolic activity of the brain, with the use of radiotracers like FDG, glucose metabolism, can be measured. In a review article by Risacher & Saykin (2013), the differences between memory disorders were examined in depth. They report of radiotracers utilized in the detection of the pathological changes, and of where these changes occur in different disorders.

A common cognitive symptom in AD is the loss of new memories. The physical changes in the brain include extracellular Aβ plaques throughout the brain, and neurofibrillary tau tangles. While the exact order of events remains partially unelucidated, these peptide aggregations and fibrillose tangles precede the cell

death and neuronal loss. The molecular components of these structures can be targeted with PET using specific radiotracers. The Aβ plaques can be targeted with [^{11}C] Pittsburgh compound B (PiB) or [^{18}F]-florbetapir radiotracers (Risacher & Saykin, 2013), while the tau tangles can be targeted with [^{18}F]-flortaucipir (Fleisher *et al*. 2020).

The first anatomical changes in AD occur in the Medial Temporal Lobe (MTL) as neuronal loss, atrophy (Braak and Braak 1997). This area of the brain includes multiple regions associated with memory, such as hippocampus as well as parts of the cortex.

On the other hand, VD is caused by obstructed blood flow to the brain proceeding a vascular event such as stroke. While Aβ plaques cannot be detected in this form of dementia, FDG-PET studies indicate a decrease in glucose metabolism in the temporoparietal lobe, and MRI reveals atrophy of both cerebral cortex and MTL (Risacher & Saykin, 2013).

In FTD different variants of the disorder can be differentiated based on both physical and clinical symptoms, and the different variants were further described in a review by Rohrer (2012). For example, behavioural variant of FTD features tau accumulation as well as intracellular aggregation of another biomarker, TAR DNA-binding protein 43 (TDP-43). On the other hand, semantic dementia (SD) is characterized by symptoms affecting language comprehension and production, and TDP-43 accumulation can be detected, while tau pathology is only rarely detected. Progressive non-fluent aphasia (PNFA) is also characterized by difficulties in language production but typically features tau pathology (Risacher and Saykin 2013; Rohrer 2012).

While PD is characterized by parkinsonism, the neuronal atrophy also affects other cognitive functions as it progresses. The distribution of inclusion deposits called Lewy bodies were described in a review article by Watson *et al*. (2009). Lewy bodies are found in the neuronal tissue, and symptoms of the disease include impairment in visual aspects as well as language and memory. Even with PD with no dementia, atrophy is detected in cortex and subcortex. FDG-PET also shows decrease in metabolism in the basal ganglia and cerebral cortex in both disorders (Risacher and Saykin 2013; Watson *et al*. 2009).

Multiple sclerosis (MS) is characterized by large lesions in the white matter. FDG reveals hypoactive glucose metabolism in thalamus, deep grey matter structures, and frontal lobe. Symptoms vary largely but include visual, motor, and sensory dysfunctions (Compston and Coles 2008; Risacher and Saykin 2013).

In addition, Risacher & Saykin (2013) reported on the activation of microglia in these disorders. Activation of microglia is recognized in AD, PD with no dementia, and MS although the location of the activated microglia varies. Hence a comprehensive approach is required for adequate diagnosis; imaging of the brain and clinical assessment of cognitive functions together can robustly differentiate between the disorders, as well as personal experience of the patient themself. A summary of the aforementioned neurodegenerative diseases and their molecular changes is presented in **Table 1**.

1.3.2 Different Stages of Alzheimer's Disease

Due to the nature of the development and progression of AD, the disease pathology can be divided into three phases: 1) cognitively normal, 2) mild cognitive impairment (MCI), and 3) AD. A difference in glucose metabolism between these phases can be detected with FDG-PET. Subjects with MCI can show glucose hypometabolism in posterior cingulate cortex and parietotemporal region with variable differences in the frontal lobe, while for subjects with AD the regions of hypometabolism can be the posterior cingulate, parietotemporal cortices, and frontal lobes (Duan *et al*. 2023).

In addition to imaging findings, the subjects must show clinical symptoms of cognitive decline, backed up by illness-affected decline in daily life. The clinical symptoms are evaluated with cognitive tests like the mini-mental state examination (MMSE). For example, in study of Duan *et al*. (2023) MCI subjects were required an MMSE score between 24 and 30, with no disturbance of daily activities or presence of dementia, while for the AD subjects the score was set between 20 and 26, and a standing diagnosis of AD.

While MMSE is a brief cognitive test with 30 questions, more comprehensive questionnaires like the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) are required to differentiate between the stages of AD.

Table 1. Summary of neuroimaging findings in selected neurodegenerative diseases and dementias.
Modified from the original table by Risacher and Saykin (2013).

Disease	Clinical Symptoms	Molecular Changes
AD	Dementia, Functional decline	FDG: Decreased glucose metabolism[1], Amyloid PET positive, Tau tangles, activated microglia?
VD	Cognitive decline (in domain coupled with vascular event)	FDG: Decreased glucose metabolism[2], Amyloid PET negative
FTD (Behavioural variant)	Personality and behaviour changes	FDG: Decreased glucose metabolism[3], Amyloid PET negative, Tau tangles, TDP-43 accumulation
FTD with motor neuron disease	Behavioural/language impairments with motor dysfunction	FDG: Decreased glucose metabolism[4], Amyloid PET negative, Tau tangles
SD	Impaired language/fluency	FDG: Decreased glucose metabolism[5], Amyloid PET negative, TDP-43 accumulation
PD with dementia/ dementia with Lewy bodies	Motor dysfunction (parkinsonism), cognitive impairment, visual hallucinations	FDG: Decreased glucose metabolism[6], Some Amyloid PET positive
PD no dementia	Motor dysfunction (parkinsonism), visual hallucinations	FDG: Decreased glucose metabolism[7], activated microglia[A]
MS	Heterogenous symptoms (autonomic, visual, motor and/or sensory dysfunction)	FDG: Decreased glucose metabolism[8], activated microglia[B]

[1]Tempoparietal regions, [2]Frontal and parietal lobes, [3]Frontal lobe, [4]Frontal and temporal lobes, basal ganglia, and thalamus, [5]Left anterior temporal lobe, [6]Basal ganglia, cerebellum, and cerebral cortex, [7]Basal ganglia, thalamus, and cerebral cortex, [8]Thalamus, deep grey matter structures, and frontal lobe, [A]Striatal and extrastriatal regions, [B]Normal and lesioned grey matter and white matter

The CERAD is a wide battery of questionnaires focusing on memory, language, visuospatial skills, and executive function. MMSE on the other hand focuses on skills related to arithmetic, memory, and orientation. However, the nature of the cognitive tests raises an issue; can the tests effectively differentiate between a decline in cognitive skills without being affected by other characteristics of the subjects, such as background, educational level, age, or socioeconomic status.

1.4 Neurodegeneration and Biomarkers

1.4.1 Inflammation

Neuronal loss has been linked to inflammation in multiple studies. One of these studies was conducted by Kipinoinen *et al*. (2022), as they studied the association of midlife inflammatory markers with cognitive performance over a ten-year period. In their study, they focused on inflammation biomarkers in the blood, interleukin 6 (IL-6), tumour necrosis factor alpha (TNF-α), and C-reactive protein (CRP). The results revealed that IL-6 and TNF-α levels correlated with cognitive skill test scores. Especially elevated IL-6 in midlife was correlated with decline in cognitive performance ten years later. Kipinoinen *et al*. (2022) also contemplated on the differences between the inflammatory markers, as IL-6 and TNF-α are secreted by adipose tissue, while CRP by liver. They proposed that the elevation of CRP levels occur later as the inflammation cascade proceeds, in the obese subjects. To conclude, IL-6 and TNF-α were evaluated as more sensitive indicators of inflammation preceding cognitive decline associated with the adipose tissue.

1.4.2 Genetic Background

In addition to inflammation biomarkers, the genes behind amyloid clearance have proven to be good predictors of memory disorders. In a study by Snellman *et al*. (2022) the connection of *Apolipoprotein E* gene (*APOE*) and its different alleles and their role as a risk factor of late onset sporadic AD was explored. The Apolipoprotein E, which is encoded by *APOE*, has a crucial role in maintenance of the cell membranes of neurons as it for example transports cholesterol and Aβ. *APOE* is a polymorphic gene with three distinct alleles: ε2, ε3, and ε4. Mammals have a diploid genome, and so each individual has two

copies of the *APOE* gene. In the research of Genin *et al*. (2011) it was found, that one copy of *APOE* ε4 allele increases the risk of AD three to four times, and two copies of *APOE* ε4 allele increase the risk five to nineteen times compared to the carriers of two ε3 allele carriers.

The *APOE* ε4 allele is connected to impaired amyloid clearance and thus accumulation of Aβ in the neuronal tissue. The accumulation results in the formation of Aβ plaques, which precedes neuronal loss. The study of Snellman *et al*. (2022) discussed the connection between extracellular Aβ plaques as well as intracellular tau aggregates and neuroinflammation.

This study was continued by Snellman *et al*. (2023) with a focus on the *APOE* ε4 gene dose effect. As the previous 2022 study established, changes in the brain can be detected with brain scans and blood tests, during the so-called silent period of AD, the period when symptoms don't yet affect cognitive abilities. One tell-tale sign in the brain scans is the Aβ plaques, the aggregations of Aβ peptides Aβ40 and Aβ42, which are the proteolytic products of amyloid precursor protein (APP). Both, the alleles of the *APOE* gene and the mutations in the APP affect the risk of Aβ plaque formation, but in the study by Snellman *et al*. (2023), the focus was alone on the *APOE* gene ε4 allele as a risk factor of AD.

The study further explored neuroinflammation, especially chronic inflammation in the brain, and how it contributes to the damage and cell loss of the neuronal tissue. This neurodegeneration is again linked to microglia and astrocytes, the immune cells of the neurons, and their activation. In addition, the different biomarkers for the activation of these immune cell types were discussed as hallmarks of the neuroinflammation. 18-kDa translocator protein (TSPO) PET imaging was utilised to reflect microglial density, and glial fibrillary acidic protein (GFAP) blood plasma level was used as a biomarker of neuroinflammation, indicating astrocytic activity.

The hypothesis of Snellman *et al*. (2023) was that early Aβ related neuroinflammatory changes should be present also in the cognitively unimpaired *APOE* ε4 carriers. However, they were unable to find associated increase of TSPO binding in ε4 double carriers even in secondary analyses.

They discussed the result and pointed out, that the ε4 genotype has been suggested to attenuate the response of microglia to amyloid accumulation. These findings further reinforce the complex nature between genotype, inflammation, and neurodegeneration.

Another finding which reinforces prior studies was, that even though patients were cognitively healthy, *APOE* ε4 genotype correlated with the amount of amyloid plaques in the brain. Furthermore, elevated GFAP in the blood, in the cognitively normal subjects, suggested the presence of reactive astrocytosis and its relation to Aβ pathology.

1.5 Brain and Glucose Metabolism

1.5.1 Glucose Metabolism as a Basis for [^{18}F]-FDG-PET

Fluorodeoxyglucose (FDG) (2-Deoxy-2-[^{18}F]fluoroglucose) is the most widely used radiotracer in PET scans, resulting in a substantial amount of scan data available for research. As a glucose analogue, FDG is taken up by tissues in proportion to their metabolic activity. This enables assessment of functional alterations in the brain on a regional level. Previous studies have demonstrated, that metabolic activity decreases prior to the onset of symptoms of dementia (Duan *et al*. 2023; Risacher and Saykin 2013). Therefore, this thesis aims to characterize alterations in brain functionality associated with dementia by utilizing FDG-PET, to better the understanding of dementia-related changes.

PET data analysis is based on the injection of a radiotracer to the subject. A portion of the tracer is taken up by tissue, while the rest remains in circulation. The radiotracer undergoes radioactive decay, and this decay emits a detectable signal. Based on this signal, the location of the decayed radiotracer as well as the intensity of the signal, the amount of the decayed radiotracer, can be quantified. In this thesis, three distinct quantification methods 1) standardized uptake value (SUV), 2) fractional uptake rate (FUR), and 3) Patlak were used for analysing the radiotracer signal, and through it, changes in glucose metabolism.

1.5.2 Standardized Uptake Value

SUV is a numeric value, which is used to reflect the concentration of radiotracers in tissues. It can be used in clinical and empirical applications to evaluate the metabolic activity of tumours and organs. SUV is calculated by normalizing the concentration of the radiotracer in the tissue, obtained from the scan data, to the injected dose and the patient's body weight. Herein lies the limitation of the SUV method; SUV is semiquantitative, several factors cause variation, such as subject preparation and imaging protocol, and the assumption of homogenic distribution of the radiotracer throughout the tissue (Chen and Nasrallah 2017).

1.5.3 Fractional Uptake Rate

To overcome the flaws of SUV's assumption, a time corrected method, FUR, can be used instead. FUR describes the rate at which a radiotracer is taken up by a specific tissue or organ relative to the administered dose over a given period. This method is often used during dynamic PET imaging. FUR requires the input function (IF), which represents the dose of active radiotracer in subject's system over the time of the scan (Thie 1995). In other words, the uptake of glucose, and FDG, is monitored throughout the study to provide more reliable data.

1.5.4 Patlak

Finally, to assess the metabolic activity of the brain tissue, Patlak-Gjedde method can be used. Patlak-Gjedde analysis is a graphical method used in PET to model the kinetics of radiotracers, with the assumption of the tracer being irreversibly trapped in tissues. The Patlak plot is a linear graphical representation used to estimate the net influx rate (K_i) of a tracer from plasma to tissue (Patlak *et al*. 1983). The assumption of radiotracer being trapped in the tissue is acceptable due to the nature of glucose metabolism. As FUR requires the IF, so does Patlak, to produce values of the activity in the tissues.

1.5.5 Glucose Metabolism

Brain FDG-PET examines glucose metabolism of the brain, i.e. the activity within neuronal tissue. However, due to metabolic differences between

subjects, FDG-PET scans should follow a protocol to minimize the variance between subjects, and the results should be normalised based on the subjects' metabolic profile. One way of accomplishing this is to perform the scans either 1) in fasting conditions or 2) under clamp. In fasting condition studies, the subject is required at least 6 hours of fasting. This ensures a euglycemic state in the metabolism of the subject. However, due to individual differences in metabolism, metabolism-affected changes in FDG uptake still occur. On the other hand, with clamp conditions, glucose and insulin infusions are used to control their levels with more consistency than fasting conditions.

In a study by Rebelos *et al*. (2021) the effect of insulin resistance on FDG-PET scans was examined in euglycemic hyperinsulinemia clamp, a state in which both glucose and insulin levels of the study subjects are maintained at a fixed level. In this study the cognitively normal subjects went through FDG-PET and MRI scans. The study yielded somewhat unexpected results. Insulin resistance is known to correlate with increased blood glucose levels in skeletal muscle tissue, but this seems to be the opposite with the brain; insulin resistance was negatively correlated with glucose uptake in the brain (Rebelos *et al*. 2021). The more insulin resistant a subject was, the more glucose the brain was utilizing. This finding raises the question of whether there is a difference in insulin signalling between the brain and muscle tissue. Further analysis of the results unveiled a connection between type 2 diabetes (T2D) and brain glucose uptake (BGU); BGU is increased significantly in subjects with T2D (Rebelos *et al*., 2021). This finding shapes the nature of dementia as a disease to be of metabolic in nature. This theory is backed up by overactive glucose metabolism being characterized as an early sign of dementia. On the other hand, gender was not proven to cause a significant difference in BGU (Rebelos *et al*. 2021).

Furthermore, increased BGU was not global when cross-compared with age. Limbic and temporal lobes were affected strongly, while frontal and parietal lobes showed a negative trend. The cognitive functions of these brain regions were studied with NeuroSynth database. By comparing the FDG-PET scans with NeuroSynth database functional MRI (fMRI) scans, Rebelos *et al*. (2021) were able to distinguish certain regions of the brain; 1) frontal, 2) parietal and, 3) temporal lobes associated with higher level cognitive functions such as

working memory, attention, and executive functions to strongly correlate with the insulin dependent BGU. However, they expressed this notion with caution as a correlation does not imply causality. Through this increased BGU in these regions a question remains; whether the increased BGU serves as an indication of the tissues working on overdrive. Are they burning out because of the increased glucose metabolism? It was also pointed out, that a natural decline with glucose usage in the brain occurs along aging. They concluded their main finding to be, that insulin sensitivity correlated negatively with BGU, especially with T2D. They stated that through their findings and in line with previous research, that during euglycemic hyperinsulinemia, BGU correlates inversely with insulin sensitivity. In other words, even in healthy brain with increase in age there is a decrease in BGU.

Nonetheless, both glucose and insulin levels have a direct effect on FDG-PET scan data, so the euglycemic hyperinsulinemia clamp method solidifies the data, and is considered “The Gold Standard” of FDG-PET scan protocol.

1.5.6 Other Metabolites Connected to Glucose Metabolism

The connection between metabolic profile and its effect on glucose metabolism was further examined by Rebelos *et al*. (2024), as they investigated the brains metabolic adaptations to obesity with proton magnetic resonance spectroscopy (^{1}H-MRS), an imaging method for quantifying metabolite concentrations, and FDG-PET under fasting, insulin clamp, and after weight loss conditions on neurologically healthy subjects. They also measured a selection of metabolites which are altered in obesity and insulin resistance; N-acetylaspartate (NAA), brain myo-inositol (MI), the pool of brain glutamine (glutamic acid, Gln, Q) and glutamate (Glu, E), (Glx), total choline, and total creatine levels.

NAA is synthesized mostly by neurons in the brain, and decreased NAA concentrations have been associated with neuronal damage in obesity (Rebelos *et al*. 2024).

MI has a complex role in glucose energy metabolism. It is vital for signalling cascades in the insulin signalling pathway. In addition, MI levels in the brain have been shown to increase with age and prior to cognitive impairment (Rebelos *et al*. 2024).

Glutamate is the major excitatory neurotransmitter. After it is cleared from the synapse by astrocytes it is converted into glutamine which can be reconverted into glutamate by neurons. Rebelos *et al*. (2024) stated, that previous studies had shown, that Glx increases with insulin stimulation and hyperglycaemia in type 1 diabetes, independently of neuronal activity, enhancing glutamate-glutamine cycling.

Cholines are a significant class of biomolecules in the synthesis and metabolism of membrane phospholipids and therefore cell membranes. ^{1}H-MRS studies have demonstrated increased concentrations of choline in different inflammatory states (Rebelos *et al*. 2024).

Finally, creatine and phosphocreatine are vital parts of energy metabolism, enabling rapid conversion of ADP to ATP by trafficking phosphate groups. As stated by Rebelos *et al*. (2024), brain creatine levels have previously been shown to be increased in metabolic syndrome during fasting, and in subjects with type 1 diabetes during hyperglycaemia.

The study of Rebelos *et al*. (2024) found statistically significant difference in brain metabolites only in NAA levels between the lean controls and severely obese subjects.

A portion of the obese subjects were restudied one year after a bariatric surgery. In addition to significant weight loss, they showed improved insulin sensitivity and all measured brain metabolites showed an increase, with creatine showing significantly higher increase.

In euglycemic hyperinsulinemic clamp studies, for both the obese subject group and the lean control group, there was no significant difference between the groups. However, during fasting in the obese group, plasma glucose was higher than under controlled euglycemia, but plasma glucose levels and their changes did not correlate with brain metabolite concentrations or their fluctuations (Rebelos *et al*. 2024).

They conclude that elevated insulin levels increase the metabolic activity in brain tissue, especially in obese subjects' brains. However, they were unable to

answer the question, what is the underlying mechanism of the altered glucose metabolism in the brain in obese individuals.

1.6 [^{18}F]-FDG-PET and Magia

1.6.1 Magia

As neurodegenerative disorders have their own distinct characteristics, research around the pathology of them has increased. The goal has shifted to catching the disease before it develops. This early intervention is becoming possible through methods like FDG-PET as it detects hypometabolism in the brain, which by the estimation of Duan *et al*. (2023) serves as a precursor of brain damage.

The limitations of earlier studies on analysing PET scan data have consistently stemmed from the insufficient dataset size, resulting in poor statistical power and reduced reliability of study findings. Searching for telltale signs of dementia before the occurrence of the disease itself requires large datasets for reliable and reproducible results. In addition, analysis of the scans requires definition of reference regions and regions of interest (ROIs). Reference region is a part of the tissue, where the activity of the radiotracer remains predictable, stable, and preferably minimal. It is used as the baseline of radiotracer activity, and the activity in the ROIs is compared to it, creating comparable results of the levels of ROI radiotracer activity. As described by Karjalainen *et al*. (2020) the analysis of brain imaging data has largely relied on experts manually selecting reference regions and ROIs in the scans leading to inter-operator variance in the data.

Especially with large datasets, automation significantly increases reliability and reproducibility as well as decreases the time and resources required for the data analysis. Karjalainen *et al*. (2020) introduced an automated image processing and kinetic modelling toolbox for PET scans, called Magia. The accuracy of Magia was tested with four different radiotracers: 1) [^{11}C]-carfentanil, 2) [^{11}C]-raclopride, 3) [^{11}C]-MADAM, and 4) [^{11}C]- PiB. These radiotracers bind to different molecules in the nervous system; mu-opioid receptors, D2- and D3-dopamine receptors, serotonin transporters, and Aβ aggregates respectively. In this study, for each of the four radiotracers 30 scans

of different subjects were used, and reference regions were manually curated by five operators, and automatically by Magia. For raclopride, MADAM, and PiB cerebral cortex was used as a reference region whereas occipital cortex for carfentanil (Karjalainen *et al*., 2020). The reference regions were defined, by both experts and Magia, using MRIs, as the scans depict the anatomical differences between subjects. The automated region generation of Magia used (a) FreeSurfer, a software designed for automatic segmentation of brain MRIs, to assign anatomic regions for each part of the PET scan voxels, (b) ROI extraction based on prespecified input, (c) anatomical correction to out-rule voxels affected by spillover radiation, and a final step of (d) tail exclusion of voxels with values from the tails of the radioactivity concentration distribution (Karjalainen *et al*., 2020).

Unsurprisingly, the inter-operator variance between the reference regions was significant, as maximum overlap of these regions was a modest 41% for raclopride and only 14 – 21 % for the other three tracers. The reference regions automatedly created by Magia were larger in volume for all tracers. This difference in reference region size resulted in a low overlap with manually drawn reference regions. The automatically curated occipital reference region produced a 14% overlap with the manually drawn. Cerebral regions overlapped with more satisfactory percentages of >50% (Karjalainen *et al*., 2020). It is remarked by Karjalainen *et al*. (2020) that Magia generated data has a higher signal-to-noise ratio.

In other words, automatically curated reference regions and ROIs could result in higher accuracy than the manually curated regions by expert operators. Nonetheless, Magia enabled cross-referencing of the data between different studies operated by different researchers with varying scan protocols.

1.6.2 Limitations of FDG

FDG-PET scans themselves provide the data of the activity of the radiotracer in the tissue of interest, which reflects glucose metabolism, glucose uptake in the tissue to be exact. However, interpretating this data to fully describe changes in metabolic activity, more than just the radio decay signal is required. Additional information is required for quantifying the activity, and this information includes

the dose of the radiotracer injected, the half-life of the radiotracer, dose injection time and scan start time as well as for more quantitative methods like FUR and Patlak, IF is also required. The IF can be derived from either blood sampling during the scan, that is arterial input function (AIF), or from the PET images themselves with image-derived input function (IDIF). IDIF derives the input function from PET images by ROIs defined around large blood vessels or the heart ventricles, to assess the amount of tracer in circulation of the subject's body.

When provided with the IF, the scan data can be interpreted to provide robust quantitative analysis. Based on the IF, an input curve, which describes the concentration of the radiotracer in the plasma, radiotracer activity in the tissue can be modelled. This curve can usually be divided into a peak following injection, and a tail following the peak. In the case of AIF failing, IDIF is the only method available for the interpretation of the scan data.

Some of the data produced by [^{18}F]-FDG-PET scans is not reliable as it is. This issue was addressed in a study by Bucci *et al*. (2024). The aim of the study was to rescue flawed image data of brain scans. Bucci *et al*. (2024) introduced input recovery (IR) method, to correct flawed input curves. Due to various reasons, like the invasive nature of AIF method, the blood sampling in the beginning of the scan may fail. This failure leads to flawed input curves, where the peak is disturbed, while the tail remains intact. With IR method, Bucci *et al*. (2024) were able to recover plasma inputs of poor quality.

To ensure the accuracy of the reconstructed peak, the IR model of Bucci *et al*. (2024) uses 1) a pre-established mathematical model by Feng *et al*. (1993) to refine the tails of the input curves, 2) exploits a Bayesian penalized-likelihood term to optimize the reconstruction of the input curve, and 3) utilizes biological reference values of input curves to validate the reconstructed input function.

The Feng model describes the shape of high-quality reference input curves. Poor-quality curves are fitted by the Feng model, corrected with Bayesian penalized likelihood. The Bayesian penalization is used to select the corrected peak formed from the input data, which best aligns with the reference data

derived boundary values. The boundary values represent biological limitations such as maximum uptake value of the radiotracer.

Bucci *et al*. (2024) conclude, that for scans with good-quality input curves, the IR model gave reliable results, and did not succumb in overfitting the values. Over and above that, for poor-quality input curves, the model improved the values, making previously unusable data suitable for analysis.

1.7 Statistical Methods in PET

1.7.1 Statistical Significance

In a review article by Xu & Kang (2024), the integration of PET imaging data, statistical analysis methods, and machine learning (ML) approaches were discussed in depth. Firstly, Xu & Kang (2024) state, that the non-invasive nature of PET imaging alone solidifies its status as a superior tool in medical diagnostics and research, compared to invasive techniques such as lumbar puncture for extraction of cerebrospinal fluid (CSF) to measure biomarkers. Furthermore, as Xu & Kang (2024) summarize, PET imaging can be used for examination of the functions of the brain, quantification of the blood flow and metabolism, and quantification of receptor binding in the brain. This enables comprehension of pathological processes across areas of cardiology, neurology, infection detection, and oncology.

Xu & Kang (2024) also describe the limitations of PET imaging. One key limitation in PET scan analysis is the noise in the imaging data, which obscures the true signal of the radiotracer. The noise in the detected signal occurs from multiple sources like electronics and photon statistics, and it might lead to false signal variation, which can be mistakenly estimated as true signal variation (Xu and Kang 2024). A second limitation is met with a phenomenon called partial volume effect. This effect is the result of brain regions overlapping within voxels, the three dimensional “pixels” in which radiotracer signal is detected. As a single voxel might hold multiple tissue types within itself, due to the limited resolution of PET imaging, inaccurate measurements can be made (Xu and Kang 2024). Lastly Xu & Kang (2024) mention the challenge of quantification of

PET scan data, which requires accounting for factors like attenuation correction, scanner sensitivity, and tracer kinetics.

Statistical methods have been used in the processing of the PET images, due to the complexity of PET scan data, resulting from the kinetics of radiotracers. As stated by Xu & Kang (2024), kinetic modelling plays an important role of dynamic modelling in PET. Mathematical models are required to estimate the various kinetic parameters of the radiotracers, and to quantify the uptake and distribution of the radiotracers. Above and beyond that, analysis of PET data harnesses statistical methods like Spearman's rank-order correlation test, Wilcoxon test, Student's t-test, and Kruskal-Wallis test. These various tests are used for example to compare the PET scan derived data and biological parameters, and to assess their correlation.

Spearman's rank-order correlation test assesses the relationship between two variables without the assumption that the data is normally distributed. In the framework of PET imaging, the test can be used to assess correlation between for example Aβ accumulation and performance in cognitive tests.

Wilcoxon test compares two dependent groups and the significance between them, also without the assumption of the data being normally distributed. It can be used to assess for example the efficacy of treatment with PET images of the subjects before and after the treatment.

On the other hand, Student's t-test assumes that the data is normally distributed, and it assesses statistical significance between two independent groups, specifically the means of the groups. In PET imaging it can be used to assess the significance for example in the difference in radiotracer uptake between two groups, 1) healthy controls and 2) subjects with dementia.

For comparison of multiple groups, for example HC, MCI, and AD, Kruskal-Wallis test can be performed. It is again a non-parametric test, it doesn't assume normally distributed nature of the data, and it evaluates the significance of the difference between the medians of the groups.

1.7.2 Statistical Estimation

In addition to these tests of correlation or significance, heavier statistical methods are used as well. Linear regression is used to model the relationship between a dependent variable and one or more independent variables. It assumes that the data is parametric, and that the relationship between the variables is linear, meaning that changes in the independent variable or variables resemble relative changes in the dependent variable. It is a method to predict how one variable is changed in relation to the other. In PET imaging, linear regression could be used to model the correlation between for example different methods used in quantifying the PET scan data, like Patlak vs. SUV and Patlak vs. FUR.

Another example of more complex statistical approaches are linear mixed models. Like linear regression, linear mixed models model complex relationships, but with more flexibility. They take into account fixed and random effects and are useful for the analysis of repeated measures. In PET imaging, linear mixed models could be used to assess for example longitudinal scan data, where the progression of Aβ accumulation is observed while considering the individual differences between the subjects' metabolisms.

A common statistical method in medicine is the Kaplan-Meier curve, which estimates the survival probabilities of subjects over time. It is a non-parametric method to compare survival rates between subject groups. In PET imaging, it is commonly used to analyse disease progression in patients undergoing treatment, for example patients with increased accumulation of PiB might show a steep decrease in survival probability, linking the tracer accumulation, which reflects Aβ plaque formation, with disease severity.

In relation to Kaplan-Meier and survival analysis, Cox curves can be derived from the Cox proportional hazards model. It is a statistical method in survival analysis to estimate the relationship between predictor variables and the occurrence of an event, such as death. It is used to analyse risk factors affecting survival. In PET imaging the model can analyse for example how radiotracer uptake correlates with the severity of the disease. A decrease in

FDG uptake might be associated with a higher risk of cognitive impairment in subjects with AD.

All in all, statistical tests and models provide valuable tools for the interpretation of raw PET scan data as well as the analysis of refined PET images, revealing correlations between variables of statistical significance.

1.8 Machine Learning Models

1.8.1 Boosting the Analysis with Algorithms

The previously described IR method by Bucci *et al*. (2024) used statistical algorithms to rescue flawed scan data. Algorithms have indeed become more and more prominent tools in data analysis. While an algorithm might be a simple set of instructions to execute a task, they may also be complex mathematical computations, like the previously described Bayesian penalized-likelihood optimisation method.

In recent years, developments in the field of artificial intelligence (AI) have been made thanks to increase in available data, increase in computing power, and improved tools, that is better algorithms to be precise, for training neural networks. AI systems, like MLMs, have become a valuable tool in data analysis, because of their ability to learn patterns from data, in comparison with rigid algorithms, which follow predefined rules. MLMs utilize algorithms in processing the data, to find patterns or features, to make predictions. A specialized type of ML called deep learning (DL), uses neural networks to learn complex patterns. Neural networks are inspired by the structure of the human brain. Neural networks are systems that are formed by layers which are connected to each other by artificial neurons. These systems process and learn from data. A convolutional neural network (CNN) is a certain type of DL model, which was developed for processing image-like data. CNNs are widely used in tasks like image classification. The key difference between ML and DL lies in feature extraction, for while the MLMs are manually trained with features of interest, DL models learn features automatically. On the other hand, a successful prediction model can be achieved with a smaller dataset with MLMs, while DL models require large datasets.

Xu & Kang (2024) also addressed ML approaches in PET imaging in their review. The undeniable effect of MLMs on PET imaging can be witnessed in applications like automated image analysis, quantitative assessments, and augmented image fidelity (Xu and Kang 2024). ML approaches can be utilized in several stages of PET imaging, such as image acquisition and reconstruction, preprocessing, quantification and comprehensive analysis (Xu and Kang 2024).

Xu & Kang (2024) report of correcting photon attenuation with CNN, generating MRI images from [^{18}F]-florbetapir-PET scans with deep generative networks, analysing PET images with a DL network, and comprehensive analysis of amyloid PET quantification with a DL model.

1.8.2 MLM Trained with a Medium-Sized Data Set

Selecting the characteristics, or features, to look for in FDG-PET scans when studying signs of dementia, has become more efficient with the method of radiomics. As methods of the time were limited to either FDG-PET-derived uptake values, or ML and DL methods, Li *et al*. (2019) introduced radiomics as a feature extraction method for finding high order features in PET scans. The features extracted from medical imaging data represented small and subtle changes in the structure and function of the brain. The PET data used was gathered from cohorts of Alzheimer's Disease Neuroimaging Initiative (ADNI) databank, with 422 subjects, and from cohorts of the PET Center, Huashan Hospital, Fudan University, Shanghai, China, with 44 subjects. The data purposefully included scans from different PET scanner types with different imaging properties to produce evidence of the robustness and generalization of the radiomics method. The scans included subjects with 1) AD, 2) the stage of heightened risk of developing AD, MCI, and 3) healthy controls (HC). ROIs were distributed in the temporal, occipital and frontal areas and due to natural differences in shape and size of these regions between subjects, the data was normalized as well as smoothed following a preprocessing process (Li *et al*., 2019).

With AD and MCI, the temporal mid region in temporal lobe has proven to be the most applicable in assessing the disorders. Using the radiomics method in these ROIs, 215 features that cannot be seen with a naked eye from the PET

scans were firstly discovered. Secondly the features were assessed, and 168 statistically robust features were deemed fit, out-ruling features caused by noise and random error. To assess the accuracy of the selected features in revealing cognitive decline, the features were matched with clinical cognitive tests. Using the selected features, a support vector machine (SVM) was used to classify between AD and HC, MCI and HC as well as AD and MCI. Throughout the three classifications the SVM was able to differentiate the subjects with >80% accuracy.

Li *et al*. (2019) acknowledged the question of the preprocessing of the data possibly causing loss of valuable information. In terms of physics the radiomics method characterizes energy and entropy in brain scans. Energy is exhibited by the uniformity of the brain tissue, with greater uniformity indicating structural integrity. The uniformity decreases when Aβ plaques and wear and tear in the tissue is detected. Entropy on the other hand, increases by the randomness and complexity of abnormalities detected in the tissue, reflecting greater variability in tissue composition (Li *et al*. 2019). It was further discussed that the detected changes in the brain with radiomics results cannot only aid in diagnosing patients with these disorders but also in showing the efficacy of treatments of these disorders.

1.8.3 MLM Trained with a Large Data Set

As proved by Li *et al*. (2019) in addition to MLMs being applicable for refining PET scan data, they can be trained to differentiate, or classify, between healthy subjects and subjects with AD based on the PET scans. Furthermore, MLMs are trained to analyse PET scans, better than specialists. One of these models is introduced by Duan *et al*. (2023) and it is called BLADNet. BLADNet was described by Duan *et al*. (2023) as a broad learning system (BLS), a neural network system without deep structure, in other words, an MLM of certain type. Earlier MLMs have encountered limitations due to insufficient size of datasets used to train them, as well as manually defined ROIs in each individual scans. BLADNet was trained using a massive amount of data and the ROIs were automatically defined, which insured cross-comparison of the scans. The model can detect even small changes in the scans, providing an analysis more accurate than some experts in the field.

The data required for the training of the model was acquired from ADNI database with the total of 2,298 FDG-PET imaging studies of 1,045 subjects (Duan *et al*. 2023). 80% of the samples were used for training the model, and 20% for testing. The model was trained for two cases, 1) prediction of HC, MCI, AD, and 2) prediction of HC, early MCI (EMCI), late MCI (LMCI), AD, and it was able to perform both predictions with high accuracy. The results of the first case demonstrate the sensitivity, specificity, and precision of the prediction reaching values around 90%. For the second case, the prediction sensitivity, specificity, and precision values reached 81.63%, 85.19%, and 83.33% respectively (Duan *et al*., 2023).

To evaluate the feature extraction of the model, Duan *et al*. (2023) visualized dimension reduction in a two-dimensional, t-distributed stochastic neighbour embedding (t-SNE) plot. The t-SNE plot showed clear clusters for both prediction cases, with only few cases mixed between the clusters. In addition to statistical analysis, the performance of the model was compared against evaluations of PET images by professional radiologists, and it out-performed the specialists with clear numbers.

Duan *et al*. (2023) concluded that the model could robustly predict AD by FDG-PET of the brain. Furthermore, it could aid with forming the diagnosis, if integrated into the tools of diagnosis. This integration is reasoned, because current AD treatments work best at the early stage of the disease, and BLADNet can detect the disease earlier on than the specialist, therefore improving the treatment efficacy and the lives of patients.

1.9 Summary

1.9.1 All We Know So Far

In this introduction we have delved into the study of neurodegenerative memory disorders utilizing brain imaging methods and machine learning. Moreover, in this introduction we have reviewed dementia and memory disorders in general, and how different types of dementia can be distinguished, not only with neuroimaging methods aimed at finding structural changes in the brain, but also with the aid of specific biomarkers (Braak and Braak 1997; Compston and

Coles 2008; Duan *et al*. 2023; Fleisher *et al*. 2020; Li *et al*. 2019; Risacher and Saykin 2013; Rohrer 2012; Watson *et al*. 2009; Xu and Kang 2024). Not only do these dementias differ in their symptoms, which can be measured with clinical cognitive tests, but also in the amount and location of different biomarkers, which can be detected with the applications of neuroimaging and through the binding of specific radiotracers (Braak and Braak 1997; Compston and Coles 2008; Duan *et al*. 2023; Fleisher *et al*. 2020; Li *et al*. 2019; Risacher and Saykin 2013; Rohrer 2012; Watson *et al*. 2009; Xu and Kang 2024).

In addition, with these methods, differentiation between the stages of AD, and in general, neurodegeneration can be identified as well as the disease progression or treatment efficacy to be followed (Duan *et al*. 2023; Li *et al*. 2019; Risacher and Saykin 2013).

We have surveyed the multifaceted interplay of inflammation, genetic background, glucose metabolism of the brain and its metabolites in relation to memory (Kipinoinen *et al*. 2022; Rebelos *et al*. 2021, 2024; Risacher and Saykin 2013; Snellman *et al*. 2022, 2023).

We also examined [^{18}F]-FDG-PET and its methodology, as well as its limitations (Bucci *et al*. 2024; Chen and Nasrallah 2017; Duan *et al*. 2023; Feng *et al*. 1993; Li *et al*. 2019; Patlak *et al*. 1983; Thie 1995; Xu and Kang 2024). We reviewed Magia, the cornerstone of this thesis, an automated PET image processing and kinetic modelling toolbox, as a powerful tool for PET scan data interpretation (Karjalainen *et al*. 2020). Furthermore, we briefly assessed the use of statistical methods in PET, in evaluating statistical significance as well as statistical estimation as a tool for predicting dependencies between variables (Xu and Kang 2024). Above and beyond, we described MLMs as a tool for boosting the scan data interpretation as well as image data analysis, with multiple examples (Bucci *et al*. 2024; Duan *et al*. 2023; Li *et al*. 2019; Xu and Kang 2024).

1.9.2 The Aims of the Thesis

This thesis aims to:

1) Automate the processing of large PET scan datasets, by utilizing existing pipelines such as Magia, and the analysis of the datasets, with MLMs,

2) Evaluate the accuracy and degree of generalizability of the models, and

3) Discuss the ethics of health data usage.

2 Materials and Methods

The PET imaging scan data was extracted from Turku PET Centre databank AIVO. Initially 926 subjects in total were chosen for the project. To be included in the project the subjects were required to have undergone FDG-PET studies in fasting conditions, the IF needed to be available, as well as glucose values and subject weight, and only adult subjects were included. Due to this study design, and incompatibilities with Magia, the pipeline used to process the PET images, the number of analysed subjects narrowed down to 320.

For further analysis of the FDG images, the necessary PET scan metadata was acquired from AIVO database through R studio. The data extracted from AIVO included 1) details about the subject: subject ID or image ID, age, weight, height, possible diagnosis/diagnoses (such as MCI, AD, Other memory problem) or healthy control status, and the research or study project name. Furthermore, it included 2) information about the study protocol: glucose levels, insulin levels, hyperinsulinemic clamp or fasting conditions during scan, scan start time, injection time, dose of radiotracer, the type of radiotracer used, frames captured during the scan, decay correction, and the IF.

The subjects were divided into four diagnostic groups based on diagnostic data gathered from AIVO. **Table 2** provides details of the grouping. Out of the 320 subjects 109 were HC (34.1%), 37 were diagnosed with a memory disorder (11.6%), 42 with other memory problem (13.1%) and the rest 132 subjects formed a miscellaneous diagnostic group “Other” (41.3%).

The study sample was comprised of 200 females (63.1%) and 117 males (36.9%), as well as 3 subjects with no gender defined in AIVO database. Age distribution ranged from 18 to 90 years, with a mean age of 55 years and a median of 56 years. All subjects had brain [^{18}F]-FDG PET scans performed during fasting conditions. The imaging studies were conducted between 1997 – 2022.

Table 2. Division of subjects into diagnostic groups based on diagnoses extracted from AIVO database.

HC	Memory Disorder	Other Memory Problem	Other
no diagnosis	Alzheimer	MCI	NON-AD PD
	dementia	memory problem	parkinson
			diabetic
			anorexia
			cadasil
			childhood epilepsy
			patient
			unknown

A quality control of the scans was applied based on the visual appearance of the FDG-PET image. The FDG-PET images were graded on a scale of 0 – 10, from worst to best, based on their visual appearance.

Magia, a MATLAB-based program, an integrated brain image processing pipeline at Turku PET Centre, was used to normalize and analyse the PET scan data as described by Karjalainen *et al*. (2020). The FDG uptake was characterized in 25 ROIs. List of the ROIs is presented in **Table 3**.

Magia fetched the information of the subjects from the AIVO database, including the PET scan and the metadata. Using this information, it normalized the PET scan by adjusting the orientation of the subject's brain, defining the different regions of the brain, and adjusting the size of the regions and their borders based on an [^{18}F]FDG-PET template image.

After Magia extracted the ROIs, the activity within the ROIs was quantified using the detected radioactivity in the FDG-PET scans, the dose of the radiotracer injected, corrected by the decay correction and IF.

Table 3. List of used ROIs in the Magia analysis of the FDG-PET scans.

Anatomical regions 1 – 12 and 20 – 25 as well as functional networks 13 – 19 where glucose metabolism was quantified with Magia.

1	Cerebellum Anterior
2	Cerebellum Posterior
3	Frontal Lobe
4	Frontotemporal Space
5	Limbic Lobe
6	Medulla
7	Mid Brain
8	Occipital Lobe
9	Parietal Lobe
10	Pons
11	Sub Lobar structures
12	Temporal Lobe
13	default (functional regions 14 -19)
14	Dorsal attention network
15	Frontoparietal network
16	Limbic system
17	Somatomotor network
18	Ventral attention network
19	Visual network
20	Bilateral Inferior Temporal Gyri
21	Posterior Cingulate and Precuneus
22	Bilateral Angular Gyri
23	Whole Brain
24	Grey Matter
25	White Matter

Three distinct ways of quantifying the activities of the brain in the ROIs were applied in this project: SUV, FUR, and Patlak.

The values obtained by the three quantification methods of Magia were processed independently. “Not applicable” (NA) values and outliers were handled. If a method produced a value of zero for a ROI, the value was set to NA. Each quantification method was analysed separately. Statistical parameters, such as quartiles, minimum, maximum, mean, median, and standard deviation, were calculated for all 25 ROIs. A histogram was graphed for each ROI based on these statistics to visualize the distribution of the values. Outlier boundaries were determined for each ROI, based on the first and third quartiles and the interquartile range (IQR, [Q3 − Q1]). Any value below the lower outlier threshold [Q1 − 1.5 * IQR] or above the upper outlier threshold [Q + 1.5 * IQR] was considered as an outlier and set to NA. Assuming that the data is normally distributed, approximately 99.3% of data points are expected to fall within these threshold values. Finally, all NA values were removed from the dataset.

Based on the information extracted from AIVO, body mass index (BMI) was calculated in R Studio.

For all subjects, and for each quantification method results separately, differences in gender, age, and BMI between the diagnostic groups were evaluated with statistical methods in R Studio. Base R included functions for Shapiro-Wilk test, Kruskal-Wallis test, and Pearson’s chi-squared test. Dunn’s test was performed with the package FSA: Simple Fisheries Stock Assessment Methods, *library(FSA)*.

For gender, Pearson’s chi-square test was conducted for revealing a statistical difference in gender distribution between different diagnostic groups.

For age and BMI, whether the data was normally distributed was tested with Shapiro-Wilk test. Based on the resulting p-value, Kruskal-Wallis test was applied. If statistical significance was found, Dunn's test with Bonferroni

correction was conducted to compare the groups pairwise while correcting for multiple comparisons.

Three MLMs were trained to classify between the four groups of subjects from the image data in R Studio. The three MLMs used in this project were: 1) Breiman and Cutlers **Random Forest**s for Classification and Regression, *library(randomForest)*, 2) Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, **Support Vector Machine**s (SVM), *library(e1071)*, and 3) Feed-Forward **Neural Network**s and Multinomial Log-Linear Models, *library(nnet)*.

In addition to Magia data, the models were trained with metadata, such as scanner type, gender, age, BMI, and diagnostic group. The Neural Network model, which only allowed for one factorial variable, was trained with Magia data, age, BMI, and the diagnostic group.

The dataset was divided into training set and testing set with Classification And REgression Training package, *library(caret)*. The three models were trained with 75% of the whole dataset and tested using the remaining 25%.

The R Studio MLM packages provided statistics of the trained MLMs. These statistics included sensitivity, specificity, balanced accuracy, positive predictive values, negative predictive values, and no information rates (NIRs) of each diagnose group prediction. An average of these statistics, except for NIR, across all group predictions was calculated. In addition, confusion matrices were used to visualize and assess the prediction data.

3 Results

3.1 All subjects

The distribution of all subjects (N=320) into diagnostic groups, with differences in subject numbers across quantification methods, is described in **Table 4**.

Based on the visual quality control of PET images, statistical measures were produced, including minimum (0), maximum (8), mean (1.83), median (1), and standard deviation (2.13) values, reflecting the distribution of observed data.

Throughout the quantification methods, it was seen from the produced histograms for each ROI, that the results are normally distributed.

Pearson's Chi-squared test result (p = 0.937) indicates that gender distribution does not significantly differ across diagnostic groups. Proportional distribution of gender is illustrated in **Figure 1**.

The Shapiro-Wilk Normality test for age (p = 9.34e-08) indicates, that age is not normally distributed across diagnostic groups. The Kruskal-Wallis test for age across diagnostic groups (p = 2.43e-12) indicates very strong evidence of difference. Further analysis with Dunn's test with Bonferroni correction provided adjusted p-values, which revealed A) Healthy Control vs. Memory Disorder (p = 0.05) moderate evidence of age difference, B) Memory Disorder vs. Other (p = 1.10e-3) strong evidence of difference in age distribution, C) Healthy Control vs. Other Memory Problem (p = 3.19e-08) very strong evidence of significant age difference, and D) Other vs. Other Memory Problem (p = 8.83e-12) very strong evidence of significant age difference between these groups. Age distribution and statistically significant differences illustrated in **Figure 2**.

For BMI, Shapiro-Wilk Normality test (p = 3.66e-12) indicates that BMI is not normally distributed across groups. The Kruskal-Wallis test for BMI across diagnostic groups (p = 3.58e-4) indicates a significant difference. Dunn's test with Bonferroni correction of BMI across different groups revealed A) Healthy Control vs. Other (p = 4.90e-4) very strong evidence of statistically significant difference, and B) Other vs. Other memory problem (p = 0.03) moderate

evidence of statistically significant difference in BMI between these groups. BMI distribution and statistically significant differences illustrated in **Figure 3**.

Statistics of the trained MLMs for each quantification method separately including sensitivity, specificity, balanced accuracy, positive predictive values, negative predictive values, and NIR of each diagnose group prediction are presented in the **Tables 5, 6, 7, 8, 9, and 10** respectively. An average of the statistics except for NIR are presented in **Table 11**.

Table 4. Size of the diagnostic group by subjects of each Magia quantification method and by all subjects.

	HC	Memory Disorder	Other Memory Problem	Other	Total
SUV	75 (27.3%)	23 (8.4%)	33 (12.0%)	77 (28.0%)	208
FUR	84 (30.5%)	32 (11.6%)	38 (13.8%)	112 (40.7%)	266
Patlak	87 (31.6%)	33 (12.0%)	36 (13.1%)	119 (43.3%)	275
All	109 (34.1%)	37 (11.6%)	42 (13.1%)	132 (41.3%)	320

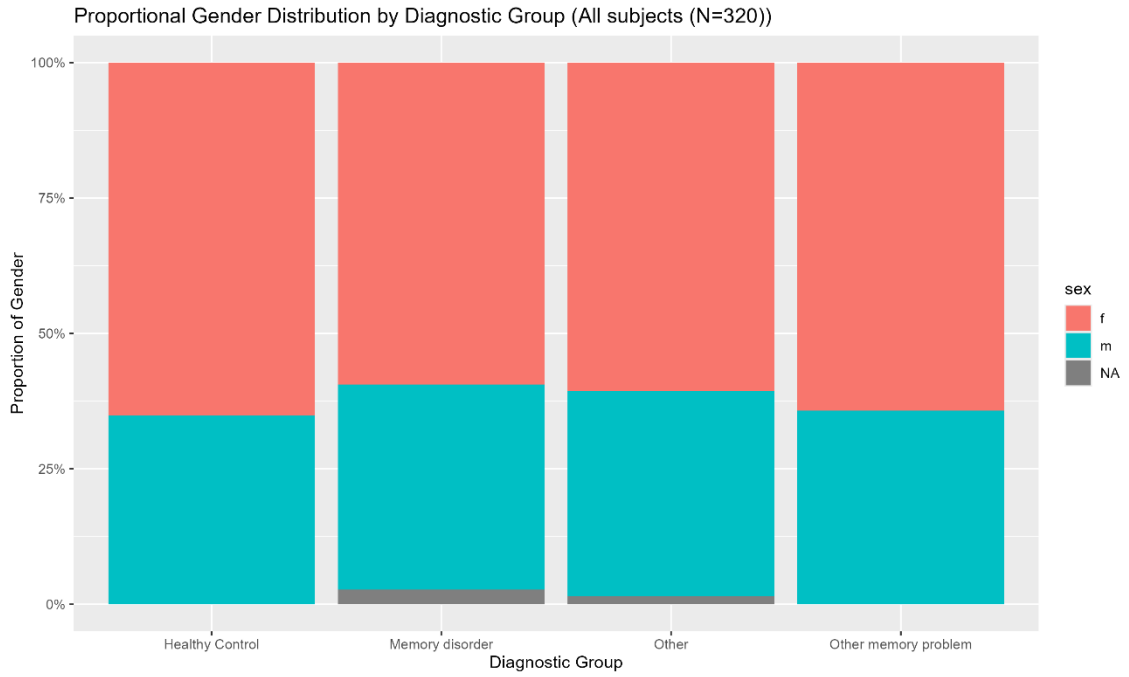

Figure 1. Proportional gender distribution by diagnostic group of all subjects. No statistically significant difference across groups.

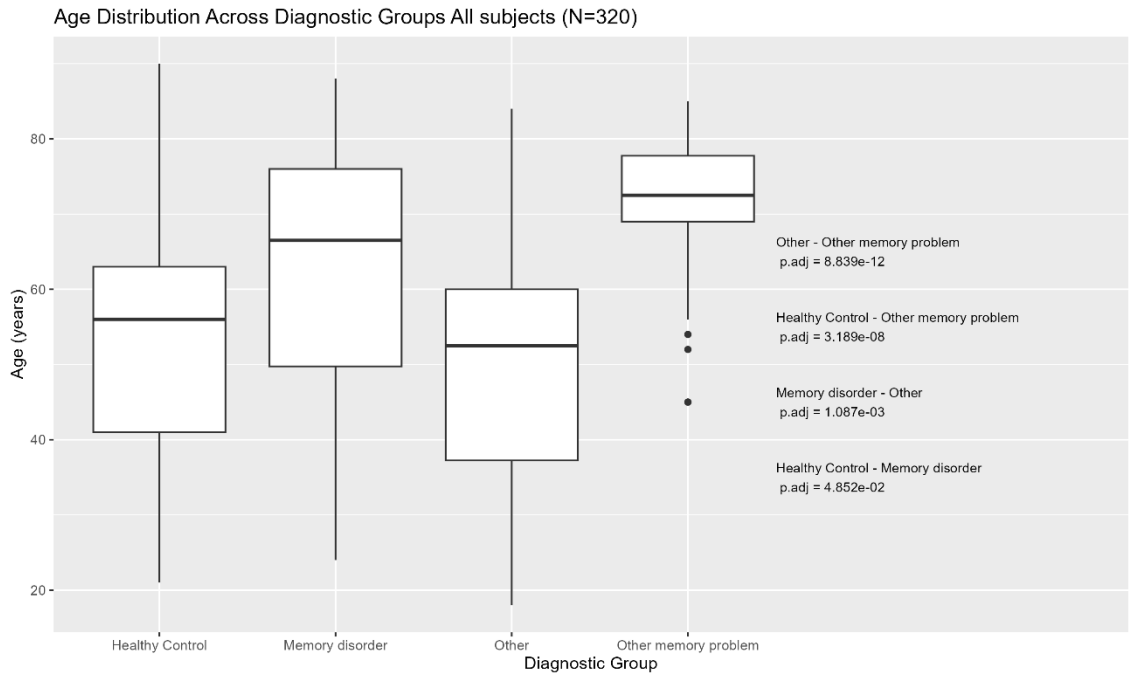

Figure 2. Age distribution of diagnostic groups. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

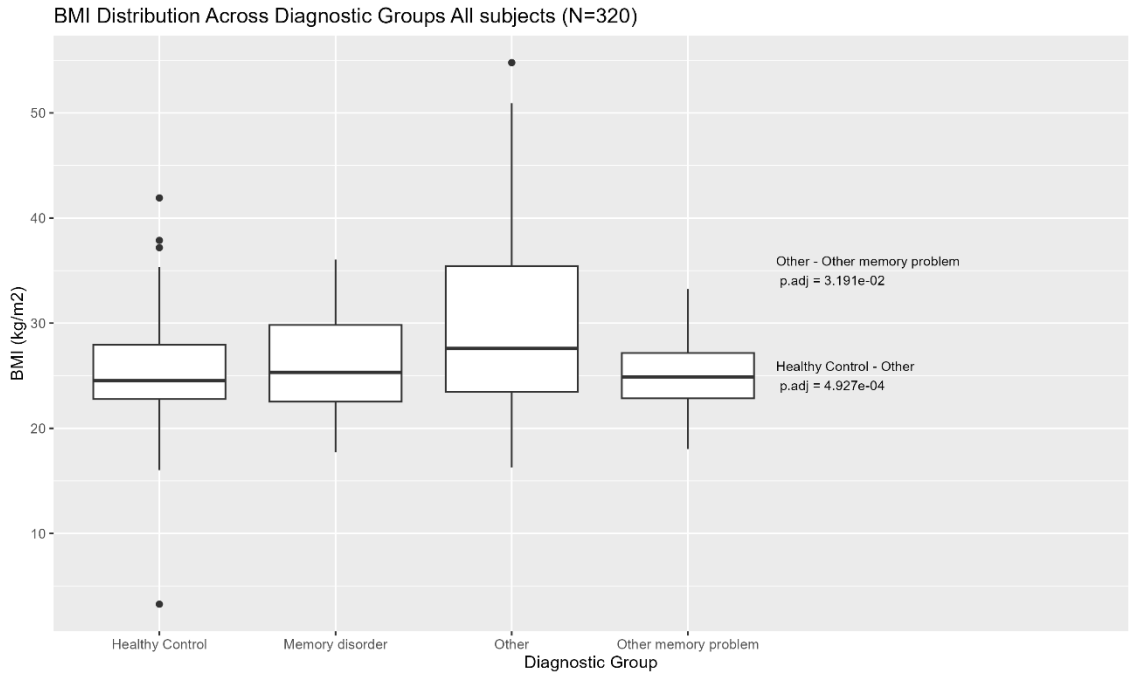

Figure 3. BMI distribution of diagnostic groups. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

Table 5. Sensitivity of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. Sensitivity, or true positive rate, is a value which indicates how many positive cases the model correctly predicted as positive.

Model and Method	Healthy Control	Memory disorder	Other	Other memory problem
Neural Network FUR	0.5000	0	0.5769	0.88890
Neural Network Patlak	0.5238	0	0.6429	0.37500
Neural Network SUV	0.4444	0	0.4737	0.87500
Random Forest FUR	0.5000	0.16667	0.5769	0.11111
Random Forest Patlak	0.5714	0	0.6071	0.12500
Random Forest SUV	0.2778	0	0.4211	0.62500
SVM FUR	0.4500	0	0.4615	0.44444
SVM Patlak	0.4286	0	0.6071	0.12500
SVM SUV	0.4444	0	0.3684	0.37500

Table 6. Specificity of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. Specificity, or true negative rate, indicates how many negative cases the model correctly predicted as negative.

Model and Method	Healthy Control	Memory disorder	Other	Other memory problem
Neural Network FUR	0.7317	1	0.7429	0.84620
Neural Network Patlak	0.6429	0.96491	0.7429	0.90909
Neural Network SUV	0.6562	1	0.6452	0.90480
Random Forest FUR	0.7317	0.98182	0.6286	0.82692
Random Forest Patlak	0.5952	0.96491	0.7143	0.92727
Random Forest SUV	0.6250	0.91110	0.6452	0.88100
SVM FUR	0.6098	1	0.6000	0.88462
SVM Patlak	0.5952	1	0.5429	0.94545
SVM SUV	0.4062	1	0.6129	0.97620

Table 7. Balanced Accuracy of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. Balanced accuracy considers the imbalance between group sizes. It is a metric which utilizes sensitivity and specificity combined.

Balanced Accuracy = (Sensitivity + Specificity)/2

Model and Method	Healthy Control	Memory disorder	Other	Other memory problem
Neural Network FUR	0.6159	0.50000	0.6599	0.86750
Neural Network Patlak	0.5833	0.48246	0.6929	0.64205
Neural Network SUV	0.5503	0.50000	0.5594	0.88990
Random Forest FUR	0.6159	0.57424	0.6027	0.46902
Random Forest Patlak	0.5833	0.48246	0.6607	0.52614
Random Forest SUV	0.4514	0.45560	0.5331	0.75300
SVM FUR	0.5299	0.50000	0.5308	0.66453
SVM Patlak	0.5119	0.50000	0.5750	0.53523
SVM SUV	0.4253	0.50000	0.4907	0.67560

Table 8. Positive Predictive Value of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. Positive predictive value, or precision, describes of all instances that the model predicted as positive, how many are correctly predicted as such.

Model and Method	Healthy Control	Memory disorder	Other	Other memory problem
Neural Network FUR	Not Available	Not Available	Not Available	Not Available
Neural Network Patlak	0.4231	0	0.6667	0.375
Neural Network SUV	Not Available	Not Available	Not Available	Not Available
Random Forest FUR	0.4762	0.5	0.5357	0.1
Random Forest Patlak	0.4138	0	0.6296	0.2
Random Forest SUV	0.2941	0	0.4211	0.5
SVM FUR	Not Available	Not Available	Not Available	Not Available
SVM Patlak	Not Available	Not Available	Not Available	Not Available
SVM SUV	Not Available	Not Available	Not Available	Not Available

Table 9. Negative Predictive Value of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. Negative predictive value indicates of all instances that the model predicted as negative, how many are correctly predicted as such.

Model and Method	Healthy Control	Memory disorder	Other	Other memory problem
Neural Network FUR	0.7500	0.90164	0.7027	0.97780
Neural Network Patlak	0.7297	0.90164	0.7222	0.90909
Neural Network SUV	0.6774	0.90000	0.6667	0.97440
Random Forest FUR	0.7500	0.91525	0.6667	0.84314
Random Forest Patlak	0.7353	0.90164	0.6944	0.87931
Random Forest SUV	0.6061	0.89130	0.6452	0.92500
SVM FUR	0.6944	0.90164	0.6000	0.90196
SVM Patlak	0.6757	0.90476	0.6333	0.88136
SVM SUV	0.5652	0.90000	0.6129	0.89130

Table 10. No Information Rate of the MLMs and their predictions by diagnostic group and FDG-PET quantification method combinations. The NIR can be used to evaluate between the robustness of the prediction of the model and an uninformed guess, as it takes into consideration the largest group within the data set.

Model and Method	NIR
Neural Network FUR	0.4262
Neural Network Patlak	0.4444
Neural Network SUV	0.3800
Random Forest FUR	0.4262
Random Forest Patlak	0.4444
Random Forest SUV	0.3800
SVM FUR	0.4262
SVM Patlak	0.4444
SVM SUV	0.3800

Table 11. Average statistical metrics of MLM predictions across all diagnostic groups by FDG-PET quantification method and MLM combinations.

	Sensitivity	Specificity	Balanced Accuracy	Positive Predictive Value	Negative Predictive Value
Neural Network FUR	0.49145	0.8302	0.660825	NA	0.833035
Neural Network Patlak	0.385425	0.81495	0.6001775	0.3662	0.8156575
Neural Network SUV	0.448275	0.80155	0.6249	NA	0.804625
Random Forest FUR	0.33867	0.79226	0.565465	0.402975	0.7937725
Random Forest Patlak	0.325875	0.80042	0.56315	0.31085	0.8026625
Random Forest SUV	0.330975	0.765575	0.548275	0.3038	0.7669
SVM FUR	0.338985	0.773605	0.5563075	NA	0.7745
SVM Patlak	0.290175	0.7708875	0.5305325	NA	0.77378
SVM SUV	0.29695	0.748825	0.5229	NA	0.74235

3.2 SUV

SUV method provided results for 208 subjects, of which 75 were HC (27.3%), 23 were diagnosed with memory disorder (8.4%), 33 with other memory problem (12.0%) and the rest 77 fell in the group other (28.0%).

The Chi-squared test result (p-value = 0.81) indicates that gender distribution does not significantly differ across diagnostic groups. Proportional distribution of gender for SUV subjects is illustrated in **Figure 4**.

The Shapiro-Wilk Normality test for age (p = 3.70e-08) indicates, that age is not normally distributed across diagnostic groups. The Kruskal-Wallis test for age across diagnostic groups (p-value = 2.15e-07) indicates strong evidence of significant difference.

Further analysis with Dunn's test with Bonferroni correction provided adjusted p-values, which revealed A) Healthy Control vs. Other memory problem (p = 5.68e-05) very strong evidence of significant age difference, B) Other vs. Other memory problem (p = 6.13e-08) very strong evidence of significant difference. Age distribution and statistically significant differences are illustrated in **Figure 5**.

For BMI, Shapiro-Wilk Normality test (p = 2.55e-4) indicates that BMI is not normally distributed across groups. The Kruskal-Wallis test for BMI across diagnostic groups (p-value = 0.05) indicates a significant difference. Dunn's test with Bonferroni correction of BMI across different groups revealed a p-value = 0.08 between Other and Other memory problem, which is not below the 0.05 threshold but is the closest to significance. BMI distribution and statistically significant differences are illustrated in **Figure 6**.

Confusion matrices for grouping based on SUV, predicted by the MLMs Random Forest, SVM, and Neural Network are presented in **Figures 7, 8, and 9** respectively.

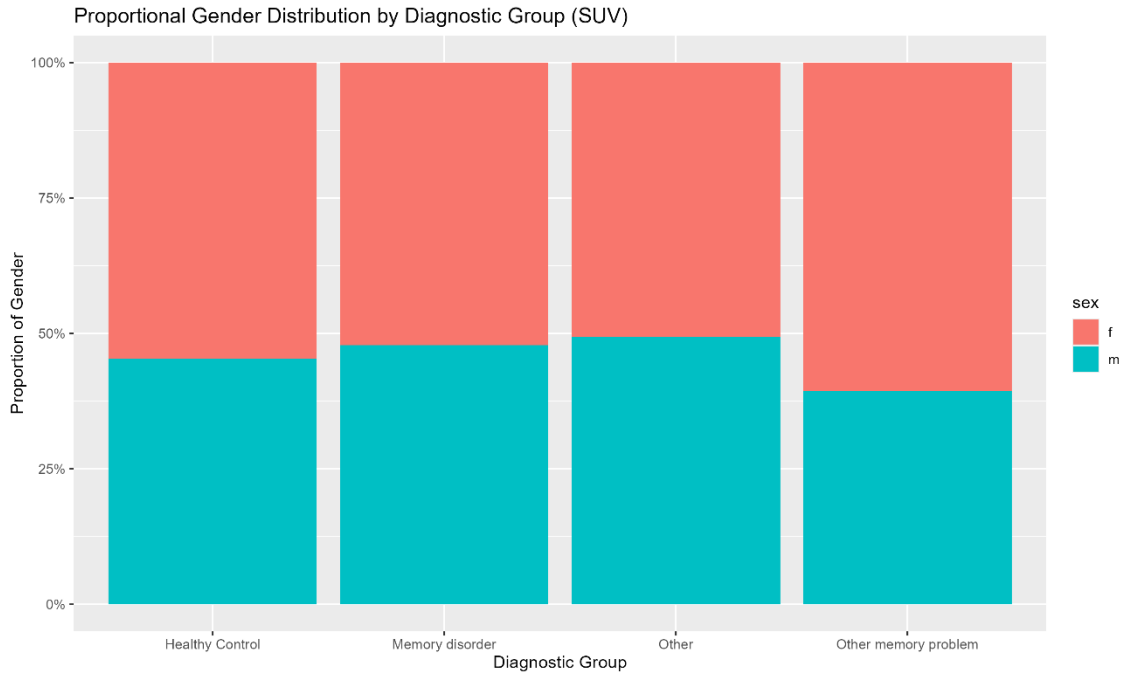

Figure 4. Proportional gender distribution by diagnostic group of subjects quantified by SUV method. No statistically significant difference across groups.

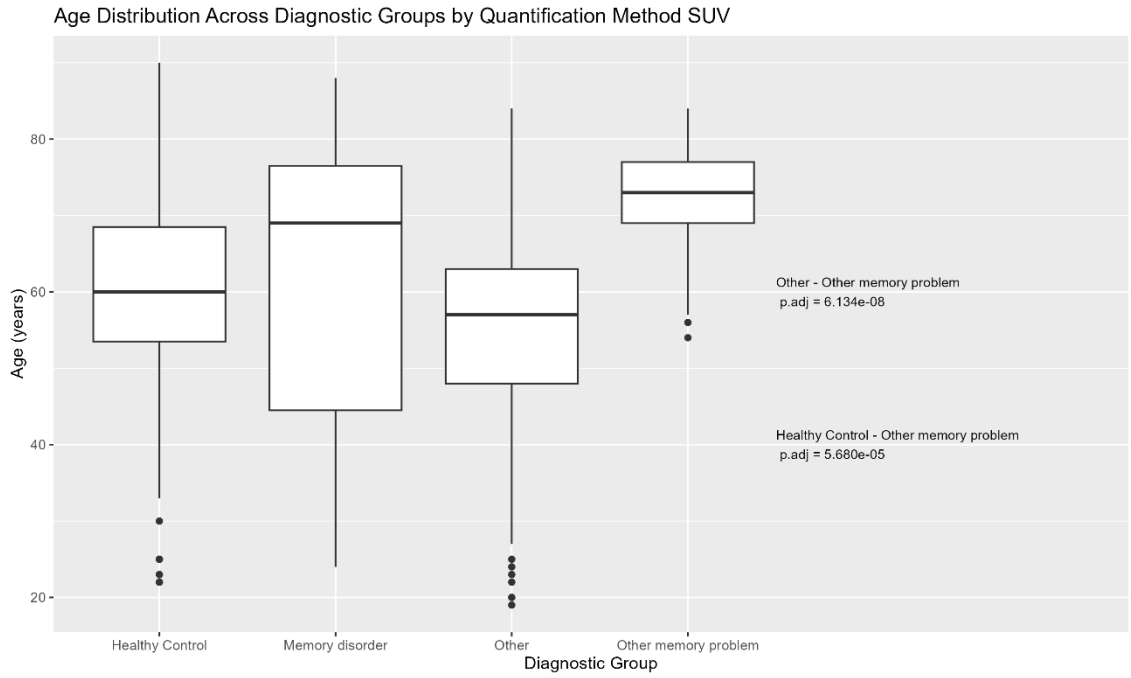

Figure 5. Age distribution of diagnostic groups by quantification method SUV. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

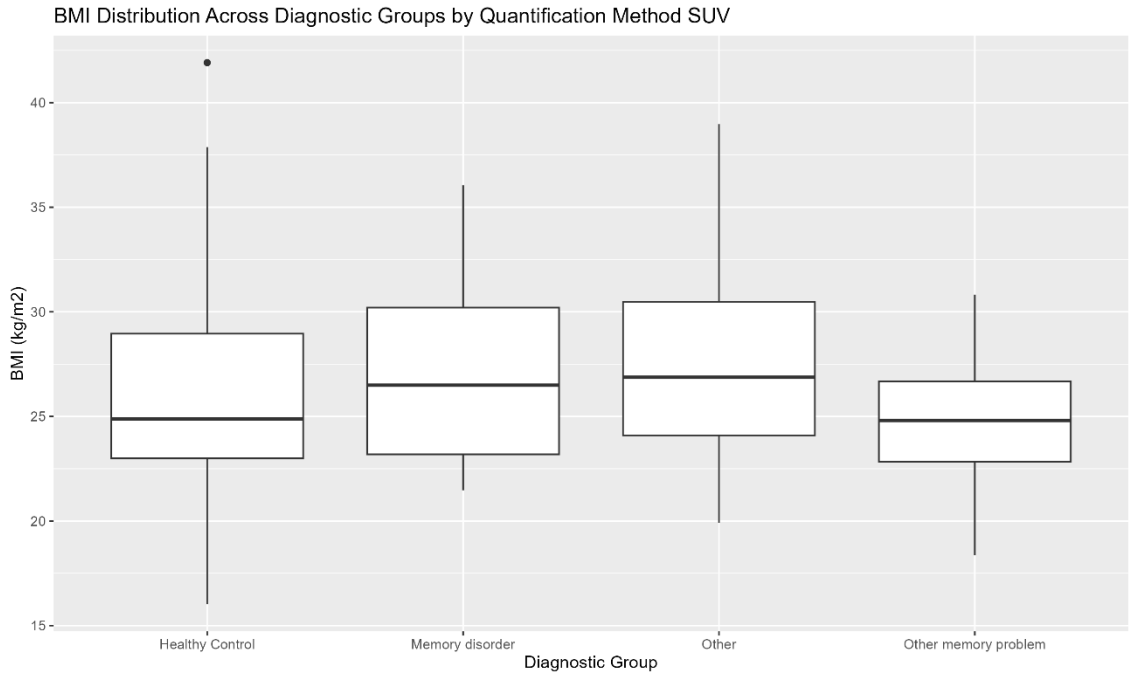

Figure 6. BMI distribution of diagnostic groups by quantification method SUV. No statistically significant differences detected between groups.

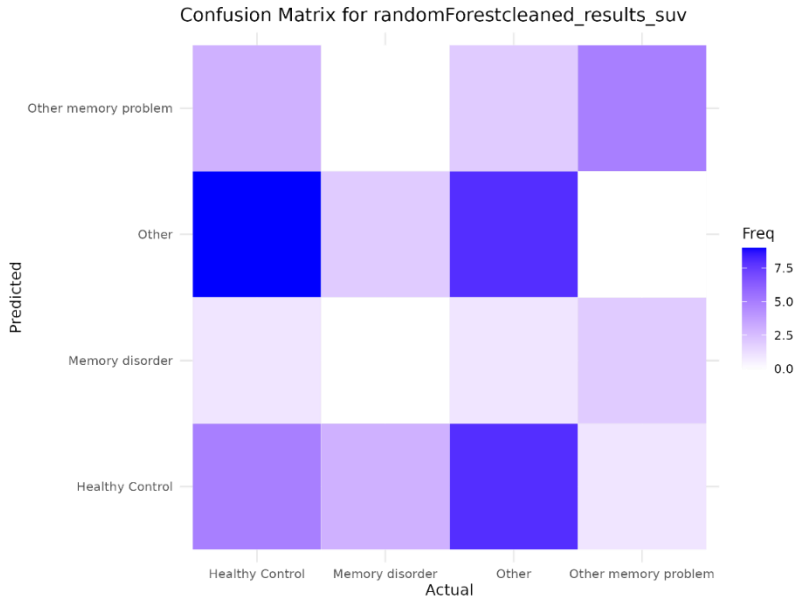

Figure 7. Confusion Matrix of group prediction by MLM Random Forest based on FDG-PET quantification method SUV results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

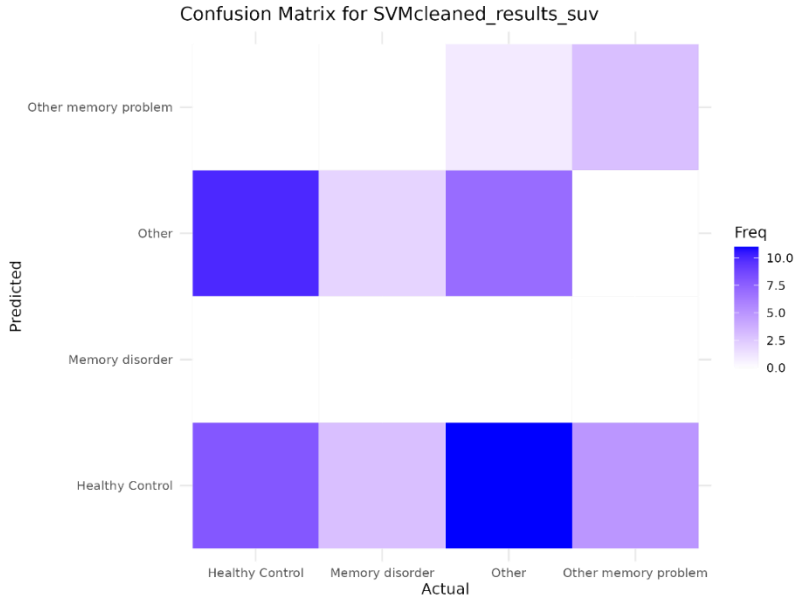

Figure 8. Confusion Matrix of group prediction by MLM SVM based on FDG-PET quantification method SUV results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

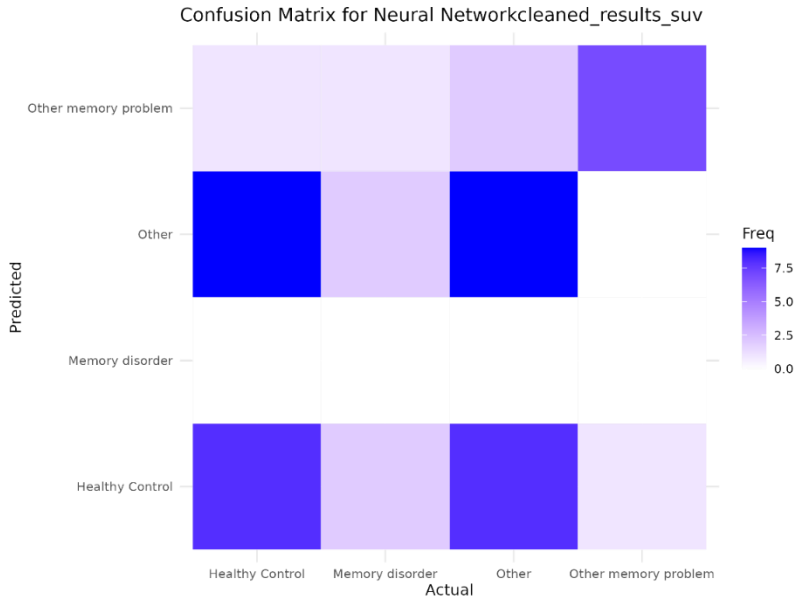

Figure 9. Confusion Matrix of group prediction by MLM Neural Network based on FDG-PET quantification method SUV results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

3.3 FUR

FUR provided results for 266 subjects, of which 84 were HC (30.5%), 32 were diagnosed with memory disorder (11.6%), 38 with other memory problem (13.8%) and the rest 112 fell in the group other (40.7%).

The Chi-squared test result (p-value = 0.94) indicates that gender distribution does not significantly differ across diagnostic groups. Proportional distribution of gender is illustrated in **Figure 10**.

The Shapiro-Wilk Normality test for age (p = 4.15e-4) indicates, that age is not normally distributed across diagnostic groups. The Kruskal-Wallis test for age across diagnostic groups (p = 3.83e-10) indicates very strong evidence of significant difference. Further analysis with Dunn's test with Bonferroni correction provided adjusted p-values, which revealed A) Memory Disorder vs. Other (p = 0.03) moderate evidence of significant difference, B) Healthy Control vs. Other Memory Problem (p = 2.05e-07) very strong evidence of significant difference in age between these groups, C) Other vs. Other Memory Problem (p = 4.61e-10) very strong evidence of significant discrepancy in age distribution, and D) Memory Disorder vs. Other Memory Problem (p = 0.05) moderate evidence of difference. Age distribution and statistically significant differences are illustrated in **Figure 11**.

For BMI, Shapiro-Wilk Normality test (p = 1.41e-11) indicates that BMI is not normally distributed across groups. The Kruskal-Wallis test for BMI across diagnostic groups (p = 8.09e-07) indicates a significant difference. Dunn's test with Bonferroni correction of BMI across different groups revealed A) Healthy Control vs. Other (p = 9.74e-06) very strong evidence of significant difference, and B) Memory Disorder vs. Other (p = 8.20e-3) strong evidence of statistically significant difference in BMI between these groups, and C) Other vs. Other Memory Problem (p = 8.90e-4) very strong evidence of significant difference in BMI. BMI distribution and statistically significant differences are illustrated in **Figure 12**.

Confusion matrices for grouping based on FUR, predicted by the MLMs Random Forest, SVM, and Neural Network are presented in the **Figures 13, 14, and 15** respectively.

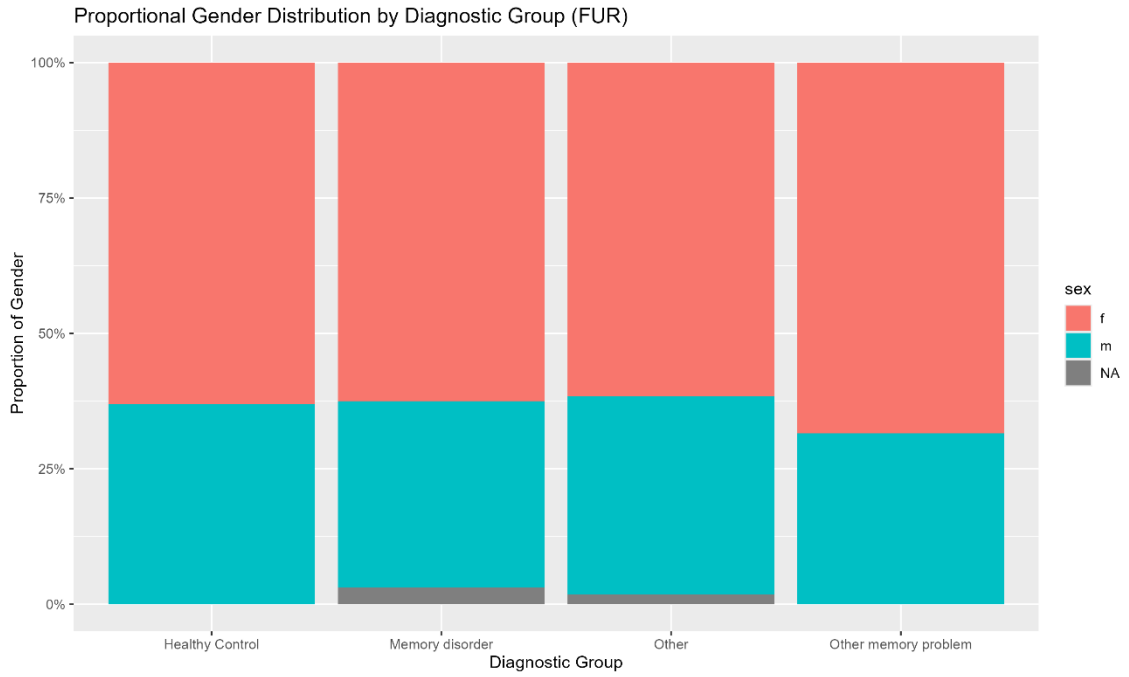

Figure 10. Proportional gender distribution by diagnostic group of subjects quantified by FUR method. No statistically significant difference across groups.

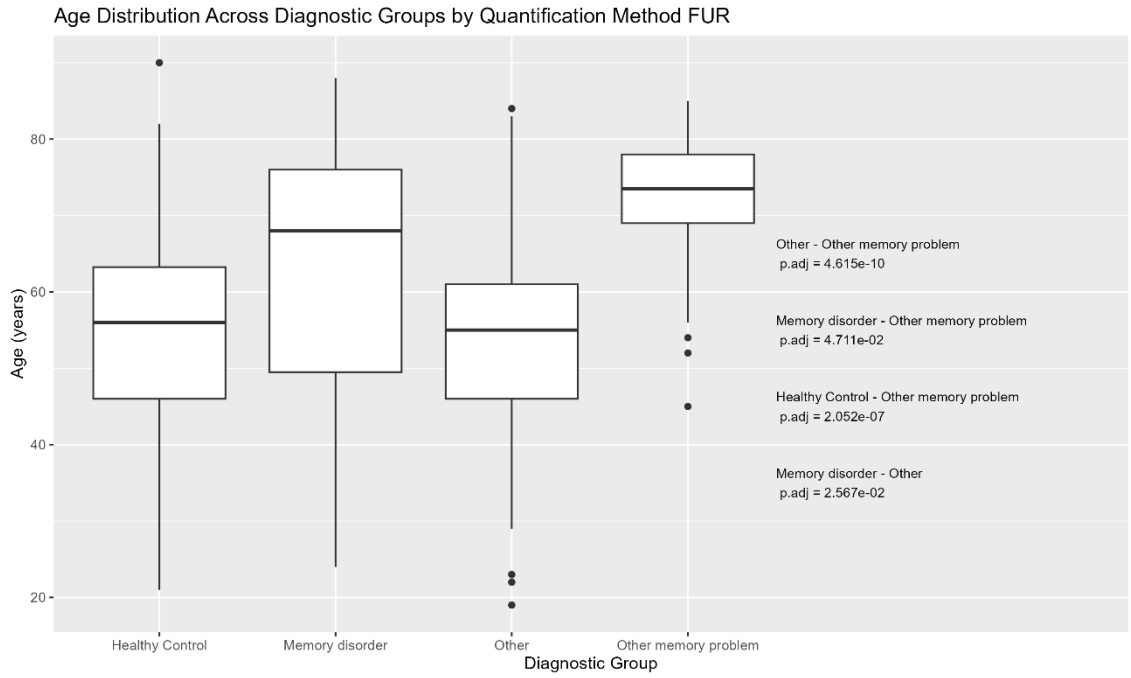

Figure 11. Age distribution of diagnostic groups by quantification method FUR. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

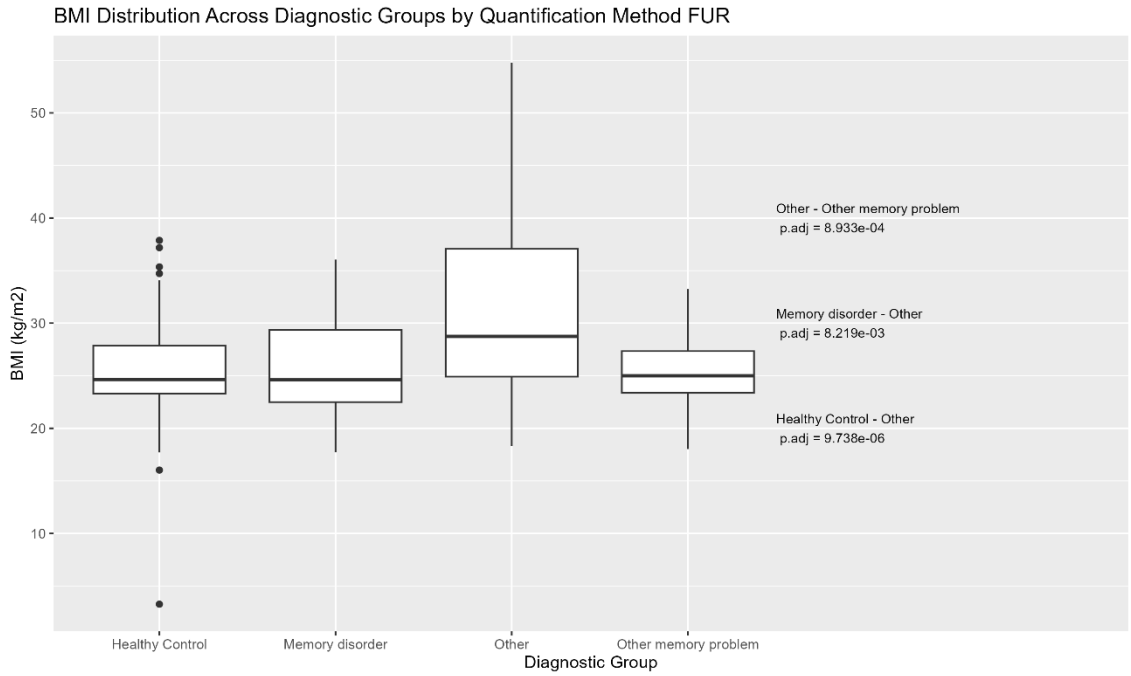

Figure 12. BMI distribution of diagnostic groups by quantification method FUR. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

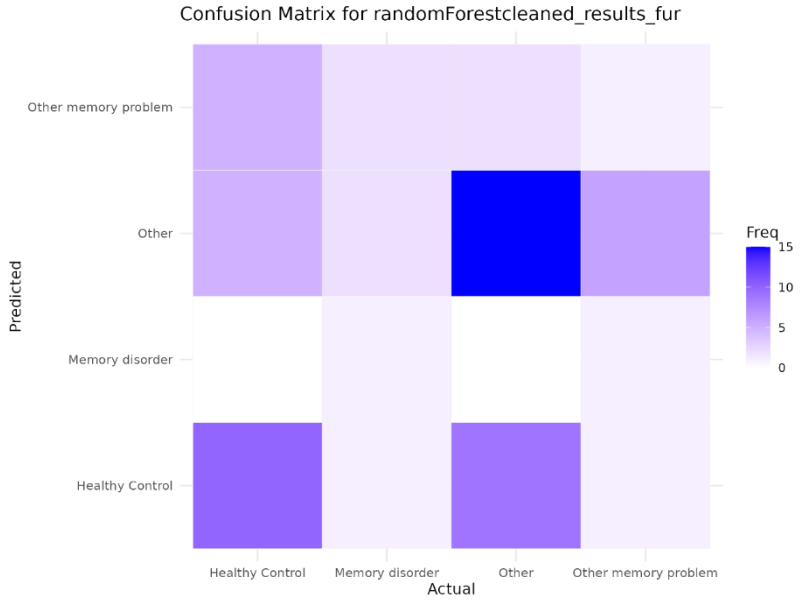

Figure 13. Confusion Matrix of group prediction by MLM Random Forest based on FDG-PET quantification method FUR results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

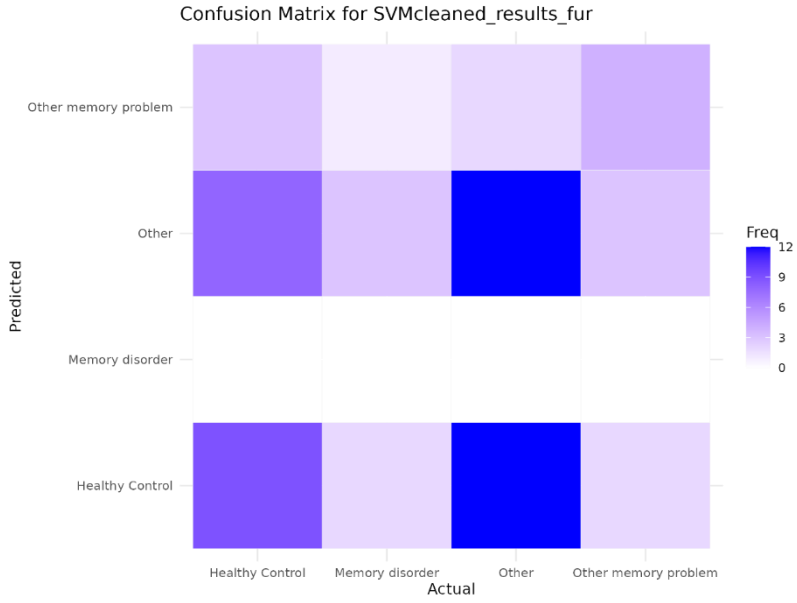

Figure 14. Confusion Matrix of group prediction by MLM SUV based on FDG-PET quantification method FUR results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

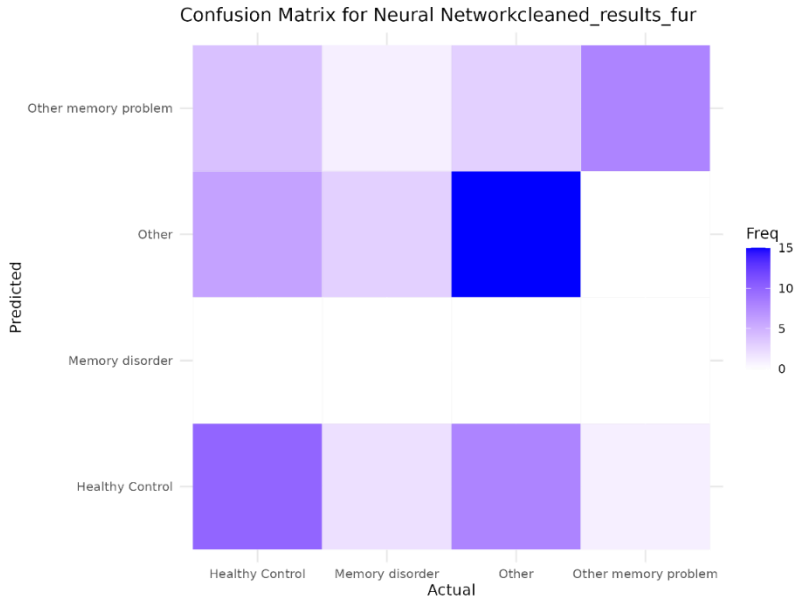

Figure 15. Confusion Matrix of group prediction by MLM Neural Network based on FDG-PET quantification method FUR results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

3.4 Patlak

Patlak provided results for 275 subjects, of which 87 were HC (31.6%), 33 were diagnosed with memory disorder (12.0%), 36 with other memory problem (13.1%) and the rest 119 fell in the group other (43.3%).

The Chi-squared test result (p-value = 0.98) indicates that gender distribution does not significantly differ across diagnostic groups. Proportional distribution of gender is illustrated in **Figure 16**.

The Shapiro-Wilk Normality test for age (p-value = 1.37e-06) indicates, that age is not normally distributed across diagnostic groups. The Kruskal-Wallis test for age across diagnostic groups (p-value = 6.01e-10) indicates significant difference. Further analysis with Dunn's test with Bonferroni correction provided adjusted p-values, which revealed A) Memory Disorder vs. Other (p = 4.20e-3) strong evidence of significant difference, B) Healthy Control vs. Other Memory Problem (p = 3.63e-06) very strong evidence of significant difference in age between these groups, and C) Other vs. Other Memory Problem (p = 9.32e-10) very strong evidence of significant discrepancy in age distribution. Age distribution and statistically significant differences are illustrated in **Figure 17**.

For BMI, Shapiro-Wilk Normality test (p-value = 1.05e-10) indicates that BMI is not normally distributed across groups. The Kruskal-Wallis test for BMI across diagnostic groups (p-value = 1.91e-4) indicates evidence of a significant difference. Dunn's test with Bonferroni correction of BMI across different groups revealed A) Healthy Control vs. Other (p = 9.50e-4) very strong evidence of significant difference, and B) Other vs. Other Memory Problem (p = 0.01) moderate evidence of statistically significant difference in BMI between these groups. BMI distribution and statistically significant differences are illustrated in **Figure 18**.

Confusion matrices for grouping based on Patlak, predicted by the MLMs Random Forest, SVM, and Neural Network are presented in the **Figures 19, 20, and 21** respectively.

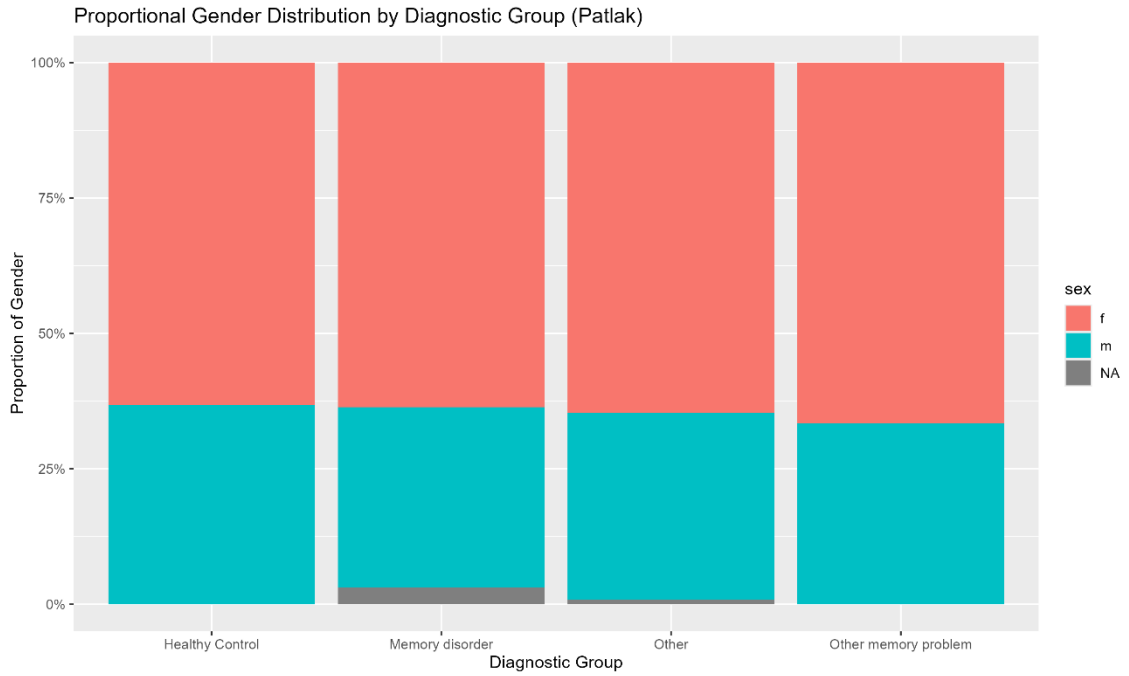

Figure 16. Proportional gender distribution by diagnostic group of subjects quantified by Patlak method. No statistically significant difference across groups.

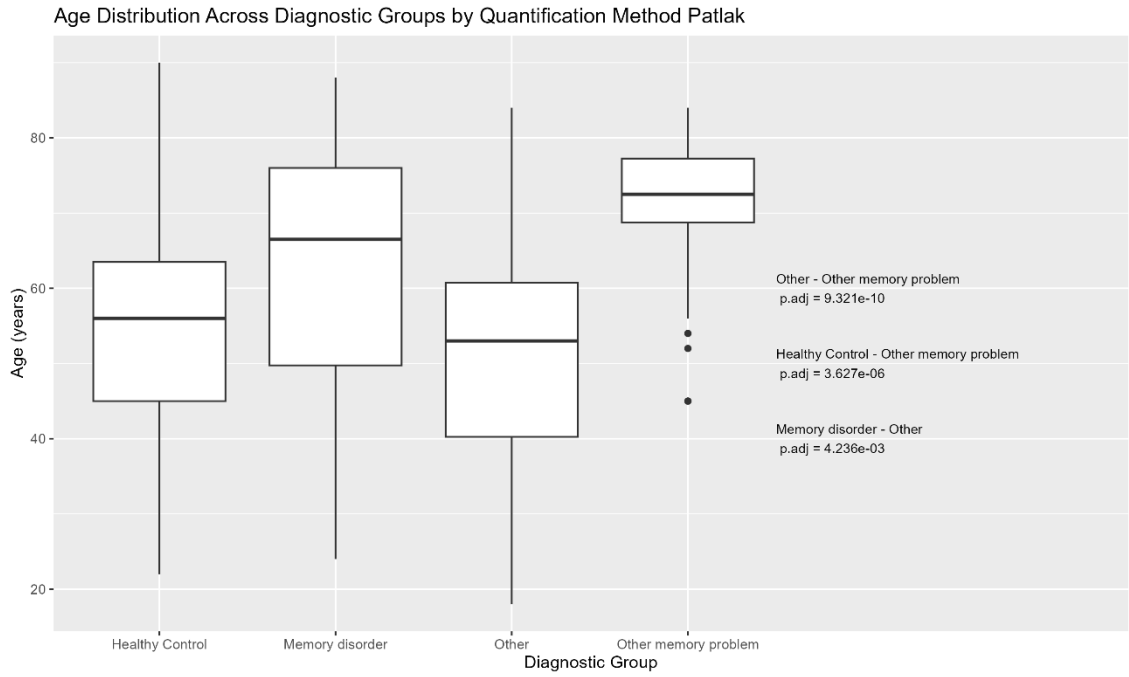

Figure 17. Age distribution of diagnostic groups by quantification method Patlak. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

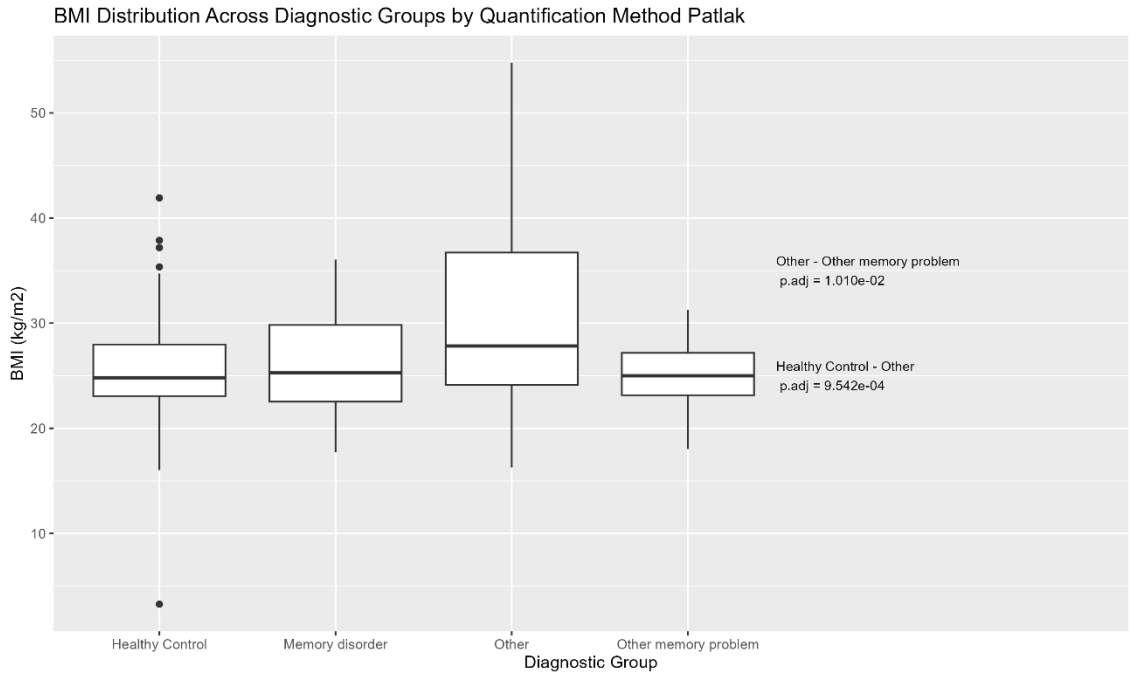

Figure 18. BMI distribution of diagnostic groups by quantification method Patlak. Statistically significant differences between groups are indicated on the right, based on Bonferroni-adjusted p-values from Dunn's test.

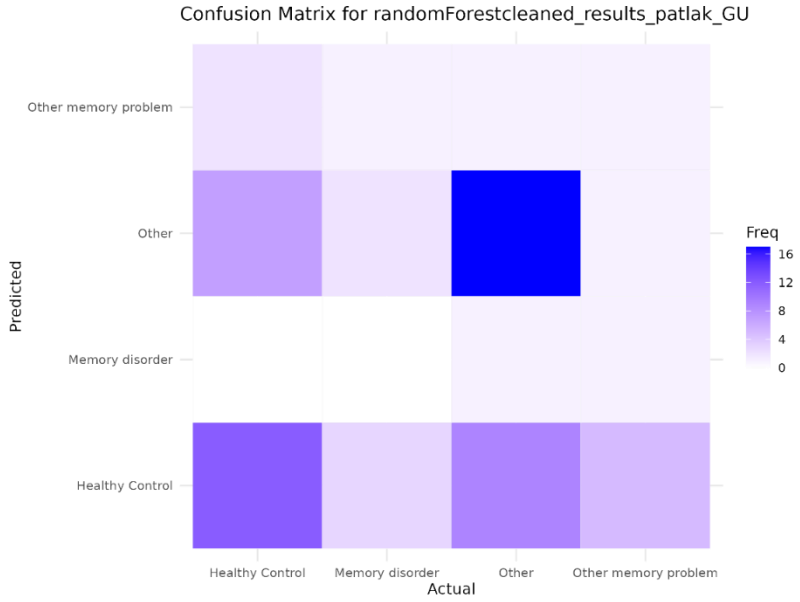

Figure 19. Confusion Matrix of group prediction by MLM Random Forest based on FDG-PET quantification method Patlak results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

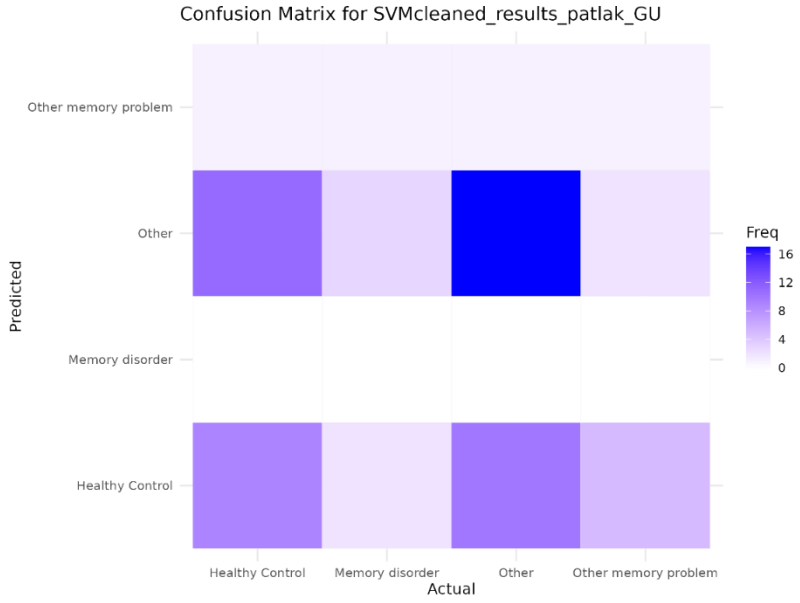

Figure 20. Confusion Matrix of group prediction by MLM SVM based on FDG-PET quantification method Patlak results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

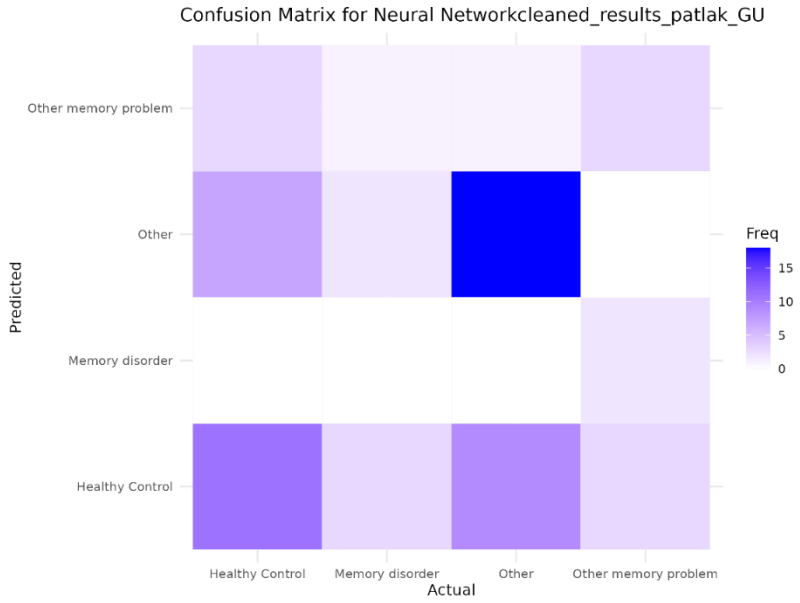

Figure 21. Confusion Matrix of group prediction by MLM Neural Network based on FDG-PET quantification method Patlak results. Rows represent actual class labels (actual diagnostic group), while columns indicate predicted labels (predicted diagnostic group). The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies.

4 Discussion and Conclusions

4.1 Data Acquisition and AIVO database

The objective during this study was to gather the data for the analysis from, and to store all the analysis data in the Turku PET Centre database, AIVO. The AIVO database is structured using PostgreSQL programming language. Information was fetched from the database using filters defined by the user. For data privacy reasons, the database is protected with usernames and passwords, and the user profiles are typically granted to read only access to specific data. In general, from the establishment of AIVO, AIVO has evolved and more fields have been added to its database making it so, that the originally imported projects did not have all the new metadata possible, so an update of some fields was needed and especially for the old projects it had to be manual.

A significant portion of the project was dedicated to gathering data from various research projects. However, obtaining data proved challenging due to the initial lack of information on the principal investigators (PIs) of these projects. With multiple studies, this crucial information was missing from AIVO and had to be sourced from other researchers, such as the senior researcher of our research group, professor Juha Rinne. After the PIs were identified, it became possible to obtain the previously missing information regarding the research subjects. This information included so called metadata, such as study date, data about possible treatment of subject and control groups, project descriptions etc.

Additionally, some missing information was gathered from the patient information systems of Varha, such as Uranus, or with a known project name, from the original publications. Accessing these systems required permissions to view confidential patient information which I do not possess. Therefore, these steps were conducted by other members of the research group, with the necessary connections, such as senior researcher, docent Marco Bucci, and permissions, doctoral researcher, Petra Parviainen, respectively.

The insufficiency of the data is explained through multiple reasons, such as the sources for the data are numerous as well as former breaches in protocol when storing data to AIVO from said sources. The main reason for the lack of data stemmed from the fact, that a project can be 20 – 30 years old and there were no protocols for storing data. An estimation of the loss of data is around 70 – 75 % when all tracer data is included.

From the perspective of data privacy, it has been reassuring to notice that, even though data is stored in multiple locations, the most sensitive information, such as social security numbers or patient/subject names, is stored in a specific location rather than being distributed across all data storage sites. To be granted credentials for these systems was a time-consuming process, and even with the credentials, not all permissions to access every piece of data were granted.

4.2 Quality Control and Data Preprocessing

The visual quality control was done manually, since no available pipelines were found. Grading the FDG-PET images from 0 – 10 by a non-expert yielded directional results at best, and none of the images were excluded from the data set based on this evaluation. However, Magia encountered some difficulties analysing a portion of the PET images, which resulted in variable data set sizes of Magia results based on the quantification method. It can be presumed that some low-grade PET images did not yield Magia results, that is, their quality was too low for Magia to handle them properly. The mean grade of 1.83 and median value of 1 of the whole data set describe the relevance of poor-quality images well.

For the Magia pipeline to run smoothly, the metadata necessary with the [^{18}F]-FDG scans was derived from the AIVO database. This metadata was incomplete at the beginning of this research project and one of the tasks was to partially update it, given the constraints of access permission and time. A big portion of the metadata has been rescued but more work has yet to be done to decrease the missing values even further.

4.3 Magia Results

4.3.1 Standardized Uptake Value

In the sense that SUV only produced results for 208 subjects, it can be evaluated to have performed poorly compared to the other quantification methods. This is surprising, since the SUV method is semiquantitative at best, and widely used in both clinical and empirical applications. One possible explanation for the low number of produced results could be the insufficiency of patient body weight data stored in AIVO for the [^{18}F]-FDG scans considered. Simply the formatting of the weight might be wrong, resulting in an error in the calculation of the SUVs.

4.3.2 Fractional Uptake Rate

FUR on the other hand requires IF to produce results. FUR produced results for 266 subjects, which is the second highest number of results among the three used quantification methods. Magia was able to report missing IF during the analysis, and this problem was handled by locating some of the misplaced IFs, while some studies did not have IF acquired by design.

4.3.3 Patlak

Out of the three methods, Patlak produced the highest number of results, 275 subjects. As with FUR, Patlak also required IF. The difference in the number of results produced by these two methods might be due to the quantification analysis requiring a fixed time window during the scan. K_i is calculated between set time point of the scan and a cut-off point. The PET images used in this study varied by their length, and therefore it is possible, that some scans were too short for either of these methods to analyse the images correctly.

4.4 Data Preprocessing

After the extraction of Magia results with the three quantification methods, the data was cleaned, as some subjects and their variables produced obviously corrupted values. The corruption of values can be a result of multiple reasons such as poor quality of the scan, variation between scanner types, variation in study protocol, insufficient or incorrectly stored metadata etc. The corrupted

values were at worst negative, as if the tissue would have pushed out the radiotracer, or so high, that it was biologically impossible to achieve such levels. The data cleanup using the IQR assumed normally distributed data, which might have handled some values as outliers wrongly. After data cleanup the different quantification methods and their results were handled separately.

After the data clean of Magia results, the subjects were divided into four groups, based on the diagnostic data. This division was necessary, as the groups HC, Memory disorder, and Other memory problem should reflect different stages of neurodegeneration as described by Duan *et al*. (2023), Li *et al*. (2019), and Risacher & Saykin (2013). Based on the research of Duan *et al*. (2023), Li *et al*. (2019), and Xu & Kang (2024) MLMs are able to differentiate between these stages.

The different stages of AD are characterised by anatomical changes occurring in different regions of the brain (Braak and Braak 1997; Risacher and Saykin 2013). As the aim of this thesis was to reveal uncharacterised changes in the brain glucose metabolism in relation to AD, the MLMs were trained with all 25 ROIs produced by Magia. The FDG-PET quantification results in specific ROIs could have been compared between the diagnostic groups, with statistical methods, to compare the differences in them and connection with prior study results. In addition, an analysis could have been performed on the ROIs known to be affected by AD, which would have demonstrated that these changes are detectable with FDG-PET image-derived data.

The subjects, which produced Magia results were reviewed by their gender, age, and BMI distribution. Out of all the subjects with at least one quantification method results (N=320), over 50% were female across all diagnostic groups. This detail might be explained through many reasons, such as women are more likely to participate to scientific studies as a subject, they live longer and therefore undergo at least some form of treatment in the medical services, and are more likely to seek medical care, compared with men. However, Pearson's Chi-square test provided no evidence for statistically significant difference. As seen in the **Figure 1**, some subjects had missing values for gender. This again is due to the inconsistency of data in AIVO for FDG scans. Every subject's

gender should be known, but it could have been improperly stored in AIVO. After the establishing of AIVO, with a second improved version of the database, an automatic fetching of the gender from the subject personal data has been implemented and these few cases with missing information will need a manual handling in the future.

For age and BMI Shapiro-Wilk test of normality and Kruskal-Wallis test for difference between groups was conducted. Across all quantification methods, neither age or BMI was normally distributed, and evidence for statistically significant difference between groups was established. With Dunn's test, the differences were examined more closely, and the resulting p-values were adjusted with Bonferroni correction. This test provided evidence of statistically significant differences in age and BMI across all three quantification methods, with varying differences between groups. The only non-significant difference between the subjects of different diagnostic groups was revealed of the BMI distribution by SUV method.

On the level of the full data set and by each quantification method, most often the group Healthy control differed in both age and BMI in relation to another group. For age, this can be explained through reasons like younger people are more likely to participate in studies, and are in general healthier, therefore shifting the age distribution in the HC group. On the other hand, people with memory problems tend to be older, which results in the shift of age distribution in the Memory disorder and Other memory problem groups. For BMI, the group with largest variance was Other, which included diagnoses like diabetes. This group was always present if evidence was found for statistical difference between two groups, except for the subjects quantified with the method SUV.

4.5 Data Analysis

Training the MLMs in R was simple. However, as can be seen from the statistics of the MLMs in **Tables 5, 6, 7, 8, 9, and 10**, as well as averages of the statistics except for NIR in **Table 11**, none of the models worked as robust predictors. The division of the data set into training set and testing set could have been stratified resulting in balanced classes across the predictions. This

could have prevented the poor accuracy of the prediction models and yielded better results.

Sensitivity, or true positive rate, is a value which indicates how many positive cases the model predicted as positive correctly. For example, how many of the subjects from the diagnostic group memory disorder correctly classified as a member of the group. In the context of medical applications, a rate above 90% of sensitivity is often required, however none of the evaluated model-quantification method combinations reached this level for the prediction of any group. The closest ones were Neural Network with FUR for predicting Other memory problem (88.9%) and Neural Network with SUV for predicting Other memory problem (87.5%). For the other sensitivity values, another threshold could be set at 70 – 80 %. This is commonly used for sensitivity in general classification tasks, but none of these values reached this level. Based on the average of sensitivity of all group predictions, the Neural Network model provided highest three results, that is the combinations Neural Network with FUR, SUV, and Patlak, from highest to lowest respectively (49.2%, 44.9%, and 38.5%).

Specificity, or true negative rate, indicates how many negative cases the model correctly predicted as negative. As for sensitivity, for specificity the same thresholds can be applied. This statistic yielded more promising results as only one was below 50%; SVM with SUV for predicting Healthy control with 40.6%. For predicting the group Memory disorder, all the combinations achieved values greater than 90%. Other memory problem achieved levels above 80% across all the combinations. Based on an average of specificity of all groups, all combinations achieved levels above 70%. The Neural Network model provided highest three average results with all three quantification methods, FUR, Patlak, and SUV from highest to lowest respectively (83%, 81.5%, and 80.2%).

Balanced accuracy considers the imbalance between group sizes. It is a metric which utilizes sensitivity and specificity combined, to assess the robustness of a model. For balanced accuracy, Neural Network model achieved the highest values in combination with FUR and SUV, 86.6% and 89% respectively, for predicting the group Other memory problem. Other than that, only Random

Forest with SUV for predicting the group Other memory problem reached the lower threshold with 75.3%. With the average of all groups, Neural Network FUR, SUV, and Patlak performed the best from highest to lowest respectively (66.1%, 62.5%, and 60%).

Positive predictive value, or precision, describes of all instances that the model predicted as positive, how many are correctly predicted as such. For example, of all the subjects the model predicted to be from the diagnostic group memory disorder, how many are actually members of the group, that is correctly classified as a member of the group. Positive predictive value could not be applied for all combinations, as Neural Network with FUR and SUV, or SVM with FUR, Patlak, and SUV was not calculated. This was likely because no false positives were predicted, and therefore precision could not be calculated. The highest average precision across all groups was achieved with Random Forest FUR combination (40.3%).

Negative predictive value indicates of all instances that the model predicted as negative, how many are correctly predicted as such. Negative predictive value on the other hand yielded statistics for all combinations and group predictions. The groups Memory disorder and Other memory problem all set close or above the threshold of 90%. Based on the average across all groups, the Neural Network model performed best with FUR (83.3%), Patlak (81.6%), and SUV (80.5%) in comparison to the other models.

The NIR can be used to evaluate between the robustness of the prediction of the model and an uninformed guess. For example, in our data sets the largest group is always Other. An uninformed guess could yield good predictions if the prediction is always set on the group Other; many of the guess predictions would be correct. The NIR was below 45% in all the combinations, meaning that an accuracy above 45% should be achieved for indicating, that the prediction performs better than a simple guess. All combinations reached a balanced accuracy above this level. Based on these statistics, the Neural Network model in combination with FUR produced the best predictions.

Duan *et al*. (2023) reported of a BL model achieving sensitivity, specificity, and precision of 92.16%, 97.56%, and 94.00%, respectively for their model on

predicting AD. On predicting MCI, the model reached levels of 89.34%, 95.58%, and 91.60% respectively. And thirdly, on predicting cognitively normal, HC that is, the model reached values of 95.16%, 95.09%, and 91.47% respectively.

Li *et al*. (2019) conducted three separate classification tests for 1) AD vs. HC, 2) MCI vs. HC, and 3) AD vs. MCI made with SVM MLM. They conducted the classification on two separate cohorts, one from ADNI and another from Huashan Hospital. For ADNI cohorts, the classification achieved average accuracies of 91.2% (AD vs. HC), 85.5% (AD vs. MCI), and 82.3% (MCI vs. HC), with Huashan cohort ROIs. For Huashan cohorts, they reported of AD vs. HC with the average accuracy of 92.1%, with Huashan cohort ROIs.

When compared with the prior research of Duan *et al*. (2023) and Li *et al*. (2019), no model reached as high levels in accuracy, sensitivity, specificity, or precision.

It is worth noting that in prior studies, PET images were used as is, instead of image-derived values, in training the models for prediction. Also worth noting is that prior studies compared only HC, MCI, and AD, or EMCI and LMCI, without the Other group. Predictions could be made focusing on specific groups, or between two groups, to yield better predictions.

The confusion matrices of predictions by quantification method and MLM provide a quick glance in the robustness of the trained models. The x-axis represents the actual labels, while the y-axis represents the predicted labels. The colour intensity indicates the frequency of predictions, with darker colours representing higher frequencies. Throughout the matrices, it can be seen that none of the trained models were robust predictors.

4.6 Final Thoughts

Even though the models were unable to produce robust predictions of the diagnostic group, based on the NIR, they were able to perform better than a simple guess. It had been established that MLMs can predict these groups with full FDG-PET images, but it came as a surprise, that image-derived data, produced by the quantification methods of Magia, did not result in strong prediction models.

Multiple points of improvement were characterized throughout the study, and when applied, those will better the models. After all, processing the full scan data requires significantly larger amount of processing time and power. To save resources, utilizing PET image-derived data could prove useful.

The main constraints were met with the insufficiency of the data stored in AIVO. Fixing the entries of the database would result in more available, as well as more reliable quantification data. With an increase in the quantification data, more robust MLMs can be trained.

As mentioned, multiple improvements can also be made for the training of the MLMs. A more rigorous stratification approach to ensure training and testing set to be balanced in the proportions of classes to be classified, for example, would greatly improve the modelling.

The next steps involve assessing the quantification methods, to determine which of them require more additional information, like age, BMI, or gender, to provide the data necessary for training MLMs for producing robust predictions. For example, in addition to FDG-PET data, genomic data of *APOE* or of the cognitive tests could prove useful in training the models.

5 Acknowledgements

This study was conducted at the Turku PET Centre (University of Turku, Åbo Akademi University, and Turku University Hospital, Turku, Finland).

Special thanks are due for the crew of Bucci Brain Imaging Lab, with a special mention of the thesis advisors dos Marco Bucci and MSc Janne Isojärvi.

5.1 Funding

Bucci Brain Imaging Lab and I-PREDICT-DEM is funded by

Sigrid Jusélius Foundation

Varha (The wellbeing services county of Southwest Finland): VTR, Finnish Governmental Research Funding (Valtion tutkimusrahoitus)

Finnish Brain Foundation (Aivosäätiö)

6 References

Braak, H. & Braak, E. (1997) Staging of Alzheimer-related cortical destruction. *Int Psychogeriatr* **9**:257–261.

Bucci, M., Rebelos, E., Oikonen, V., Rinne, J., Nummenmaa, L., Iozzo, P. & Nuutila, P. (2024) Kinetic modeling of brain [18-F]FDG positron emission tomography time activity curves with input function recovery (IR) method. *Metabolites* **14**:114.

Chen, Y. J. & Nasrallah, I. M. (2017) Brain amyloid PET interpretation approaches: from visual assessment in the clinic to quantitative pharmacokinetic modeling. *Clin Transl Imaging* **5**:561–573.

Compston, A. & Coles, A. (2008) Multiple sclerosis. *Lancet Br Ed* **372**:1502–1517.

Duan, J., Liu, Y., Wu, H., Wang, J., Chen, L. & Chen, C. L. P. (2023) Broad learning for early diagnosis of Alzheimer's disease using FDG-PET of the brain. *Front Neurosci* **17**:1137567.

Feng, D., Huang, S.-C. & Wang, X. (1993) Models for computer simulation studies of input functions for tracer kinetic modeling with positron emission tomography. *Int J Biomed Comput* **32**:95–110.

Fleisher, A. S., Pontecorvo, M. J., Devous, M. D., Lu, M., Arora, A. K., Truocchio, S. P., … A16 Study Investigators (2020) Positron emission tomography imaging with [^{18} F]flortaucipir and postmortem assessment of Alzheimer disease neuropathologic changes. *JAMA Neurol* **77**:829–839.

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., … Campion, D. (2011) APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Mol Psychiatry* **16**:903–907.

Jansen, W. J., Janssen, O., Tijms, B. M., Vos, S. J. B., Ossenkoppele, R., Visser, P. J., … Zetterberg, H. (2022) Prevalence estimates of amyloid abnormality across the alzheimer disease clinical spectrum. *JAMA Neurol* **79**:228–243.

Karjalainen, T., Tuisku, J., Santavirta, S., Kantonen, T., Bucci, M., Tuominen, L., … Nummenmaa, L. (2020) Magia: robust automated image processing and kinetic modeling toolbox for PET neuroinformatics. *Front Neuroinformatics* **14**:3.

Kipinoinen, T., Toppala, S., Rinne, J. O., Viitanen, M. H., Jula, A. M. & Ekblad, L. L. (2022) Association of midlife inflammatory markers with cognitive performance at 10-year follow-up. *Neurology* **99**:2294–2302.

Li, Y., Jiang, J., Lu, J., Jiang, J., Zhang, H. & Zuo, C. (2019) Radiomics: A novel feature extraction method for brain neuron degeneration disease using 18F-FDG PET imaging and its implementation for Alzheimer's disease and mild cognitive impairment. *Ther Adv Neurol Disord* **12**:1–12.

Patlak, C. S., Blasberg, R. G. & Fenstermacher, J. D. (1983) Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. *J Cereb Blood Flow Metab* **3**:1–7.

Rebelos, E., Bucci, M., Karjalainen, T., Oikonen, V., Bertoldo, A., Hannukainen, J. C., … Nuutila, P. (2021) Insulin resistance is associated with enhanced brain glucose uptake during euglycemic hyperinsulinemia: a large-scale PET cohort. *Diabetes Care* **44**:788–794.

Rebelos, E., Latva-Rasku, A., Koskensalo, K., Pekkarinen, L., Saukko, E., Ihalainen, J., … Nuutila, P. (2024) Insulin-stimulated brain glucose uptake correlates with brain metabolites in severe obesity: A combined neuroimaging study. *J Cereb Blood Flow Metab* **44**:407–418.

Risacher, S. & Saykin, A. (2013) Neuroimaging biomarkers of neurodegenerative diseases and dementia. *Semin Neurol* **33**:386–416.

Rohrer, J. D. (2012) Structural brain imaging in frontotemporal dementia. *Biochim Biophys Acta BBA - Mol Basis Dis* **1822**:325–332.

Snellman, A., Ekblad, L. L., Koivumäki, M., Lindgrén, N., Tuisku, J., Perälä, M., … Rinne, J. O. (2022) ASIC-E4: Interplay of beta-amyloid, synaptic density and neuroinflammation in cognitively normal volunteers with three levels of genetic risk for late-onset Alzheimer's disease – study protocol and baseline characteristics. *Front Neurol* **13**:826423.

Snellman, A., Ekblad, L. L., Tuisku, J., Koivumäki, M., Ashton, N. J., Lantero-Rodriguez, J., … Rinne, J. O. (2023) APOE ε4 gene dose effect on imaging and blood biomarkers of neuroinflammation and beta-amyloid in cognitively unimpaired elderly. *Alzheimers Res Ther* **15**:71.

Thie, J. A. (1995) Clarification of a fractional uptake concept. *J Nucl Med* **36**:711–712.

Watson, R., Blamire, A. M. & O'Brien, J. T. (2009) Magnetic resonance imaging in Lewy body dementias. *Dement Geriatr Cogn Disord* **28**:493–506.

Xu, K. & Kang, H. (2024) A review of machine learning approaches for brain positron emission tomography data analysis. *Nucl Med Mol Imaging* **58**:203–212.