NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

Unravel Heterogeneity in Human Brain Aging with Neuroimaging and Artificial Intelligence: Clinical, Lifestyle, Cognitive, and Genetic Associations.

A thesis submitted for the degree of
Doctor of Philosophy

by

Ioanna Skampardoni

Athens, December 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Μελέτη των Μοτίβων της Γήρανσης στον Ανθρώπινο Εγκέφαλο με χρήση Νευροαπεικονιστικών Δεδομένων και Μηχανικής Μάθησης – Συσχέτιση με Κλινικούς Βιοδείκτες, Τρόπο Ζωής, Γνωστικές Δοκιμασίες και Γενετικά Δεδομένα.

Διδακτορική Διατριβή

της

Ιωάννας Σκαμπαρδώνη

Αθήνα, Δεκέμβριος 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

Unravel Heterogeneity in Human Brain Aging with Neuroimaging and Artificial Intelligence: Clinical, Lifestyle, Cognitive, and Genetic Associations.

PhD Thesis

Ioanna Skampardoni

Advisory Committee: Prof. Konstantina Nikita (Supervisor)
Prof. Christos Davatzikos
Prof. Georgios Stamou

Approved by the exam committee on December 16, 2024

…….	…….	…….
Konstantina Nikita	Christos Davatzikos	Georgios Stamou
Professor NTUA	Professor UPenn	Professor NTUA

…….	…….	…….	…….
Athanasios Voulodimos	Konstantia Zarkogianni	Panayiotis Tsanakas	Alexis Kelekis
Assistant Professor NTUA	Associate Professor MU	Professor NTUA	Professor NKUA

Athens, December 2024

.....
Ιωάννα Σκαμπαρδώνη Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Ιωάννα Σκαμπαρδώνη, 2024

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

It is forbidden to copy, store, and distribute this work, in whole or in part, for commercial purposes. Reproduction, storage, and distribution are permitted for non-profit, educational, or research purposes, provided that the source is referenced and this message is retained. Questions concerning the use of this work for profit should be addressed to the writer.

The views and conclusions contained in this document express the author and should not be interpreted as representing the official positions of the National Technical University of Athens.

Περίληψη

Η παρούσα διδακτορική διατριβή εξετάζει την ετερογένεια των αλλαγών που συμβαίνουν στον ανθρώπινο εγκέφαλο με τη γήρανση και την εμφάνιση νευροεκφυλιστικών παθήσεων. Για την επίτευξη αυτού του επιστημονικού στόχου, αξιοποιούνται σύγχρονες μέθοδοι μηχανικής μάθησης, οι οποίες εφαρμόζονται σε δεδομένα μαγνητικής τομογραφίας προερχόμενα από μεγάλους και ποικιλόμορφους πληθυσμούς. Συγκεκριμένα, μελετώνται οι νευροανατομικές μεταβολές που συμβαίνουν στον εγκέφαλο σε όλο το φάσμα της γήρανσης, από τα πρώιμα στάδια πριν την εκδήλωση γνωστικής εξασθένησης έως τα προχωρημένα στάδια της νόσου Αλτσχάιμερ (Alzheimer's disease). Στην ανάλυση αυτή λαμβάνονται υπόψη παράγοντες όπως η συννοσηρότητα (comorbidity), ο τρόπος ζωής, καθώς και περιβαλλοντικοί και γενετικοί παράγοντες, οι οποίοι ενδέχεται να επηρεάσουν τις εγκεφαλικές αλλαγές. Επιπλέον, η διατριβή επιδιώκει να αξιοποιήσει τα αναδυόμενα μοτίβα δομικών αλλαγών για την πρόβλεψη της πιθανότητας μελλοντικής εμφάνισης γνωστικής εξασθένησης και επιδείνωσης της νόσου με στόχο την βελτίωση των κλινικών παρεμβάσεων και την καλύτερη διαχείριση των ασθενών.

Στο πρώτο μέρος της διατριβής, χρησιμοποιείται μία μέθοδος βαθιάς μάθησης (deep learning), το Smile-GAN, που βασίζεται σε παραγωγικά αντιπαραθετικά δίκτυα (generative adversarial networks, GAN), με σκοπό την ανίχνευση μοτίβων δομικών αλλαγών σε γνωσιακά υγιή άτομα. Η μέθοδος εφαρμόζεται σε συγχρονικά (cross-sectional) δεδομένα που περιλαμβάνουν όγκους ανατομικών περιοχών και αλλοιώσεων της λευκής ουσίας (white matter lesions/hyperintensities) του εγκεφάλου, οι οποίοι προέρχονται από εικόνες T1- και T2- μαγνητικής τομογραφίας. Τα δεδομένα αντλούνται από ένα εκτενές δείγμα 27.402 ατόμων μέσης και προχωρημένης ηλικίας, τα οποία συμμετέχουν στην κοινοπραξία μελετών iSTAGING (imaging-based coordinate SysTem for AGIng and NeurodeGenerative diseases). Το δείγμα χωρίζεται σε ηλικιακές δεκαετίες, συγκεκριμένα 45-55 ετών μέχρι και 75-85 ετών, προκειμένου να μελετηθεί η ετερογένεια των δομικών αλλαγών ξεχωριστά για κάθε ηλικιακή ομάδα. Στη συνέχεια, μελετώνται συσχετίσεις των εξαγόμενων υποομάδων με παράγοντες όπως η αμυλοειδής β (amyloid β) πρωτεΐνη, που έχει συνδεθεί με τη νόσο Αλτσχάιμερ, καρδιαγγειακοί δείκτες, γνωστικές επιδόσεις, ο τρόπος ζωής και η γενετική προδιάθεση. Τέλος, διερευνάται η συσχέτιση αυτών των υποομάδων πρώιμων δομικών εγκεφαλικών αλλαγών με την πιθανότητα μελλοντικής εμφάνισης ήπιας γνωστικής διαταραχής (mild cognitive impairment).

Η μελέτη αυτή εντοπίζει υποομάδες που εμφανίζουν κοινά χαρακτηριστικά μεταξύ των διαφόρων ηλικιακών ομάδων. Συγκεκριμένα, παρατηρούνται μια τυπική υποομάδα γήρανσης με χαμηλά επίπεδα ατροφίας και αλλοιώσεων λευκής ουσίας και γενετικό προφίλ που προστατεύει από καρδιαγγειακές παθήσεις, και δύο υποομάδες προχωρημένης γήρανσης: η πρώτη χαρακτηρίζεται από αυξημένους παράγοντες κινδύνου για καρδιαγγειακές παθήσεις, διαταραχές της ακεραιότητας της λευκής ουσίας και αυξημένη εναπόθεση αμυλοειδούς β, ενώ η δεύτερη παρουσιάζει διάχυτη και υψηλής έντασης εγκεφαλική ατροφία, πιθανώς λόγω περιβαλλοντικών παραγόντων και παραγόντων που σχετίζονται με τον τρόπο ζωής. Αυτές οι υποομάδες φαίνεται να αντικατοπτρίζουν τη διαφορετική ευαισθησία των ατόμων των υποομάδων στη γνωστική έκπτωση και τον ρυθμό επιδείνωσής της, καθώς και τη μελλοντική εμφάνιση της νόσου Αλτσχάιμερ.

Στο δεύτερο μέρος της διατριβής, αναπτύσσεται μία νέα μεθοδολογία για τη μελέτη της ετερογένειας. Αυτή η προσέγγιση επιδιώκει να ξεπεράσει τους περιορισμούς που επιβάλλουν οι υπάρχουσες μεθοδολογίες, οι οποίες βασίζονται αποκλειστικά σε συγχρονικά δεδομένα για την εύρεση μοτίβων εγκεφαλικής γήρανσης, παραβλέποντας τη δυναμική εξέλιξη της γήρανσης και των σχετικών παθολογικών καταστάσεων σε βάθος χρόνου. Η προτεινόμενη μεθοδολογία, που ονομάζεται από κοινού μη-αρνητική παραγοντοποίηση συγχρονικών και διαχρονικών πινάκων (Coupled Cross-sectional and Longitudinal – Non-negative matrix factorization, CCL-NMF), βασίζεται στη μη-αρνητική παραγοντοποίηση πίνακα και εξάγει τα μοτίβα εκμεταλλευόμενη δύο είδη δεδομένων που προσφέρουν διαφορετικές και δυνητικά συμπληρωματικές πληροφορίες: χάρτες/πίνακες που αποτυπώνουν συγχρονικές και διαχρονικές (longitudinal) αλλαγές στον εγκέφαλο με τη γήρανση. Συγκεκριμένα, η παραγοντοποίηση διεξάγεται από κοινού στους δύο πίνακες: ο συγχρονικός πίνακας αναπαριστά τα μακροχρόνια και σωρευτικά αποτελέσματα της γήρανσης ενός ηλικιωμένου πληθυσμού (πληθυσμός στόχος) σε σχέση με έναν υγιή και νεότερης ηλικίας πληθυσμό (πληθυσμός αναφοράς), ενώ ο διαχρονικός πίνακας αποτυπώνει τις δυναμικές αλλαγές του εγκεφάλου με τη γήρανση σε ατομικό επίπεδο. Η μεθοδολογία CCL-NMF περιλαμβάνει την παραγοντοποίηση κάθε πίνακα σε δύο νέους πίνακες μικρότερων μεγεθών: ο πρώτος πίνακας απεικονίζει τα μοτίβα εγκεφαλικών αλλαγών (πίνακας λεξικό/βάση) και είναι κοινός για τους δύο τύπους δεδομένων και ο δεύτερος πίνακας (πίνακας συντελεστών/βαρών) προσδιορίζει τον βαθμό έκφρασης κάθε μοτίβου σε ατομικό επίπεδο και διαφοροποιείται για τα συγχρονικά και τα διαχρονικά δεδομένα. Επιπλέον, σε αντίθεση με προηγούμενα μοντέλα, η μέθοδος CCL-NMF δεν προσεγγίζει την ετερογένεια ως ένα πρόβλημα συσταδοποίησης, καθώς επιτρέπει στο κάθε άτομο την ταυτόχρονη έκφραση πολλαπλών μοτίβων με διαφορετικούς συντελεστές.

Αυτή η μεθοδολογία μπορεί να εφαρμοστεί για τη μελέτη διαφόρων χαρακτηριστικών που σχετίζονται με την εγκεφαλική γήρανση όπως η σταδιακή συρρίκνωση/ατροφία της φαιάς ουσίας, η προοδευτική εναπόθεση αμυλοειδούς β και tau πρωτεϊνών, η αυξανόμενη συσσώρευση αλλοιώσεων της λευκής ουσίας στον εγκέφαλο, κ.λπ. Η παρούσα διατριβή επικεντρώνεται στην εφαρμογή του μοντέλου για τη μελέτη της εγκεφαλικής ατροφίας, χρησιμοποιώντας όγκους ανατομικών περιοχών προερχόμενους από εικόνες Τ1-μαγνητικής τομογραφίας. Δεδομένου του χαρακτήρα του προβλήματος, το οποίο αφορά μη-επιβλεπόμενη μάθηση χωρίς σαφώς καθορισμένη λύση, η επικύρωση της μεθόδου πραγματοποιείται με τη χρήση ημι-συνθετικών δεδομένων, στα οποία έχουν προσομοιωθεί συγκεκριμένα μοτίβα ατροφίας. Με αυτό τον τρόπο, διερευνάται η ικανότητα του μοντέλου να ανιχνεύει τα προσομοιωμένα μοτίβα.

Στη συνέχεια, το μοντέλο χρησιμοποιείται για την εύρεση μοτίβων εγκεφαλικής ατροφίας σε έναν πληθυσμό προχωρημένης ηλικίας (N=48.949) έχοντας ως αναφορά έναν υγιή πληθυσμό μέσης ηλικίας (N=977), και οι δύο προερχόμενοι από το iSTAGING. Η ανάλυση των συσχετίσεων με παράγοντες κινδύνου καρδιαγγειακής νόσου, βιοδείκτες της νόσου Αλτσχάιμερ, γνωστική έκπτωση και πιθανότητα μελλοντικής επιδείνωσης αυτής, φανερώνει ότι τα αναδυόμενα μοτίβα ατροφίας ευθυγραμμίζονται με κλινικούς φαινότυπους. Επιπλέον, η εξαγωγή εξατομικευμένων επιπέδων έκφρασης αυτών των μοτίβων μέσω των συντελεστών προάγει την κατεύθυνση ενός πιο εξατομικευμένου προγράμματος διαχείρισης ασθενών και σχεδιασμού θεραπευτικών παρεμβάσεων.

Επιπλέον, η σύγκριση με προηγμένα μοντέλα βαθιάς μάθησης που εφαρμόζονται στα ίδια δεδομένα φανερώνει ότι οι συντελεστές έκφρασης των μοτίβων CCL-NMF προσφέρουν βελτιωμένη προβλεπτική ικανότητα για διάφορα κλινικά χαρακτηριστικά. Έτσι, συνεισφέρουν στην καλύτερη κατανόηση των μηχανισμών γήρανσης και νευροεκφυλισμού. Τέλος, αναπτύσσονται μοντέλα παλινδρόμησης για την γρήγορη και εύκολη εκτίμηση των συντελεστών αυτών των μοτίβων σε νέα σύνολα δεδομένων, χωρίς να απαιτείται η εφαρμογή της μεθόδου από την αρχή, διευρύνοντας έτσι την χρήση αυτής σε διαφορετικά ερευνητικά και κλινικά περιβάλλοντα.

Λέξεις κλειδιά
Νευροαπεικόνιση, Μηχανική Μάθηση, Συσταδοποίηση, Δεδομένα Μεγάλης Κλίμακας, Μη-αρνητική Παραγοντοποίηση Πίνακα, Εγκεφαλική Γήρανση, Νόσος Αλτσχάιμερ, Καρδιαγγειακή Νόσος, Ατροφία, Αλλοιώσεις Λευκής Ουσίας, Ετερογένεια, Εξατομικευμένη Ιατρική.

Abstract

The present thesis investigates the complex and multifaceted brain changes associated with aging, which lead to cognitive decline and the development of Alzheimer's disease (AD). Utilizing and advancing state-of-the-art machine learning techniques and harnessing large-scale datasets, distinct and homogeneous imaging patterns linked to various brain aging trajectories are identified. The overarching objective is to disentangle the neuroanatomical heterogeneity across the brain aging spectrum, examining the variability driven by AD-related degeneration and the influence of co-existing pathologies, lifestyle, environmental, and genetic risk factors. Additionally, this work seeks to leverage the identified dimensions of brain changes to predict future cognitive decline and clinical progression, providing insights that may ultimately improve early diagnosis, risk stratification, and intervention strategies in aging and neurodegenerative diseases.

First, the heterogeneity of neuroanatomical brain changes in aging at early asymptomatic phases is investigated by leveraging recent advancements in deep learning and big data analytics. Collectively, there has been an increasing understanding of the neurobiological processes related to various neuropathologies that affect the human brain, including AD and cerebrovascular disease. However, little is known about how people, at the individual level, transition from normal aging to pathologic manifestation. This knowledge gap is partly due to the lack of sufficiently large-scale neuroimaging datasets and the tools to model and validate such complex processes. Unraveling the neuroanatomical heterogeneity in aging at early stages before the emergence of clinical symptoms may provide prognostic information about susceptibility to or presence of neurodegenerative disease and influence patient management and clinical trial recruitment.

To address this challenge, a novel semi-supervised clustering method based on generative adversarial networks (GAN), termed Smile-GAN, is applied to cross-sectional anatomic regions of interest (ROI) volumetrics and white matter hyperintensities (WMH) derived from T1- and T2-weighted magnetic resonance imaging (MRI) data, respectively, consolidated by the iSTAGING (imaging-based coordinate SysTem for AGIng and NeurodeGenerative diseases) consortium for a large-scale and diverse harmonized multi-cohort sample of middle-to-late age cognitively unimpaired individuals (N=27,402). Neuroanatomical subgroups are independently examined in four decade-long age intervals spanning 45 to 85 years, with the use of decade intervals helping to mitigate age-related effects during clustering. The derived subgroups are

then correlated with genetic and lifestyle risk factors, biomedical measures, and cognitive decline trajectories.

Three subgroups, consistent across decades, are identified within the cognitively unimpaired population. Briefly, a typical aging subgroup characterized by low atrophy and white matter lesions and a genetic profile protective against vascular disease and two accelerated aging subgroups are found: one characterized by elevated cardiovascular disease risk factors, disruption of white matter integrity, and increased cerebral amyloid β deposition, while the other displays diffuse and severe atrophy, likely driven by lifestyle and exposure factors. These subgroups may reflect differential susceptibility to AD and other neurodegenerative conditions, cognitive decline, and clinical progression.

Next, a novel methodology for the study of heterogeneity is introduced. Unlike current approaches relying solely on cross-sectional data, thus neglecting dynamic observations of pathological changes, the proposed approach, termed Coupled Cross-sectional and Longitudinal Non-negative Matrix Factorization (CCL-NMF), develops a mutually constrained NMF framework to delineate components that encapsulate distinct patterns of brain alterations derived jointly from cross-sectional and longitudinal data. A cross-sectional map (C-map) captures the cumulative brain changes due to aging or disease over long periods inferred from broader population-level comparisons, while a longitudinal map (L-map) captures the dynamic patterns of brain change on an individual basis. CCL-NMF identifies components shared by C- and L-maps based on the assumption that an aging or disease effect estimated cross-sectionally at a population level should be compatible with dynamic brain changes captured by longitudinal data. It also estimates corresponding coefficients (loadings), representing the degree of expression of each component from each individual by optimizing the reconstruction of both data types, thereby capturing the complex interplay between static and dynamic aspects of brain alterations. Notably, CCL-NMF avoids rigid classification into mutually exclusive categorical subtypes, allowing individuals to exhibit varying degrees of co-expression across multiple patterns, which is important for capturing co-existing pathologies.

The proposed methodology is formulated in a general framework, enabling its application to analyses of heterogeneity of any disease characterized by monotonic brain alterations (e.g., gradual gray matter atrophy and cerebrospinal fluid expansion, progressive white matter lesion accumulation, or increasing deposition of neuropathologies such as amyloid and tau). This thesis applies CCL-NMF to parse the heterogeneity of aging-related atrophy using anatomic ROI volumetrics derived from T1-weighted MRI data. The method is

first validated using semi-synthetic data with predefined atrophy patterns and severity levels. Then, it is applied to delineate the heterogeneity in an aging population (N=48,949) with a healthy middle-aged cohort (N=977) as a reference. Both populations are drawn from the iSTAGING consortium. The identified components are correlated with AD biomarkers, cognition, cardiovascular disease risk factors, and disease progression, revealing meaningful patterns closely aligned with clinical phenotypes, highlighting the method's ability to offer deeper insights into the biological processes underlying aging. Importantly, by deriving individualized expression levels across these components, the approach facilitates personalized therapeutic interventions tailored to individual patient profiles, paving the way for more targeted and effective treatment strategies.

Moreover, comparisons with state-of-the-art deep learning models applied to the same dataset demonstrate that the CCL-NMF components provide improved predictive power for biomarkers and clinical variables, refining our understanding of brain aging pathways. Finally, the model facilitates out-of-sample application through regression-based loading estimation, broadening its utility in research and clinical contexts.

Keywords
Neuroimaging, Machine Learning, Clustering, Big Data, Non-negative Matrix Factorization, Brain Aging, Alzheimer's Disease, Cardiovascular Disease, Atrophy, White Matter Hyperintensities, Heterogeneity, Personalized Medicine.

Ευχαριστίες

Η παρούσα διδακτορική διατριβή εκπονήθηκε κατά το χρονικό διάστημα 2018-2024 στο Εργαστήριο Βιοϊατρικών Προσομοιώσεων και Απεικονιστικής Τεχνολογίας της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου και σε συνεργασία με το Center for Biomedical Image Computing and Analytics (CBICA) του University of Pennsylvania. Με την ολοκλήρωσή της, κλείνει ένας κύκλος έξι χρόνων, γεμάτος με νέες γνώσεις, πλούσιες εμπειρίες και αξέχαστες αναμνήσεις.

Θα ήθελα να εκφράσω τις πιο θερμές μου ευχαριστίες στην Καθηγήτρια κα. Κωνσταντίνα Νικήτα και στον Καθηγητή Κ. Χρήστο Νταβατζίκο για την καθοδήγησή τους, την αμέριστη υποστήριξη τους και τον ενθουσιασμό τους για το αντικείμενο της ερευνητικής μου εργασίας. Επίσης, θέλω να ευχαριστήσω θερμά τα υπόλοιπα μέλη της επταμελούς επιτροπής για τη συνεργασία και την υποστήριξη που μου παρείχαν καθ' όλη τη διάρκεια της ερευνητικής μου διαδρομής. Είμαι ιδιαίτερα ευγνώμων στον Δρ. Guray Erus για τις συμβουλές και τον χρόνο που πάντα με προθυμία αφιέρωνε για να με ενθαρρύνει και να με υποστηρίξει σε ερευνητικό και προσωπικό επίπεδο. Ακόμα, θα ήθελα να ευχαριστήσω τους Δρ. Ilya Nasrallah και Δρ. Haochang Shou για την υποστήριξή τους και την εξαιρετική συνεργασία που έχουμε αναπτύξει όλα αυτά τα χρόνια.

Θα ήθελα ακόμη να ευχαριστήσω τα μέλη των δύο εργαστηρίων για την δημιουργική συνεργασία μας, καθώς και για την αλληλεγγύη που οικοδομήσαμε όλα αυτά τα χρόνια. Η ανταλλαγή σκέψεων, ανησυχιών και γέλιων υπήρξε αναπόσπαστο κομμάτι αυτής της εμπειρίας.

Η ενθάρρυνση και η υποστήριξη των αγαπημένων μου ανθρώπων υπήρξαν η κινητήριος δύναμη σε αυτή την πορεία. Ευχαριστώ τους φίλους μου – ανάμεσά τους την Ελένη και τον Στάθη – για τις ευχάριστες στιγμές που μου προσέφεραν και για την κατανόηση που έδειξαν στις ανησυχίες μου, καθώς και όλους εκείνους που βρέθηκαν δίπλα μου και με στήριξαν σε αυτό το ταξίδι, ο καθένας με τον δικό του μοναδικό τρόπο.

Η διδακτορική μου διατριβή είναι αφιερωμένη στους γονείς μου, Σταμάτη και Κατερίνα. Τους ευχαριστώ από καρδιάς για την αγάπη, τη συνεχή φροντίδα και την πολύπλευρη στήριξή τους. Η συμβολή τους υπήρξε καθοριστική για την εκπόνηση και ολοκλήρωση αυτής της διατριβής.

Στους γονείς μου,
Σταμάτη και Κατερίνα

Εκτεταμένη περίληψη

Η παγκόσμια αύξηση του προσδόκιμου ζωής οδηγεί σε δραματική αύξηση του ηλικιωμένου πληθυσμού, ο οποίος αναμένεται να φτάσει το 1,5 δισεκατομμύριο μέχρι το 2050. Αυτή η δημογραφική εξέλιξη συνοδεύεται από αύξηση του κινδύνου εμφάνισης νευροεκφυλιστικών ασθενειών, όπως η άνοια (dementia) και η νόσος του Αλτσχάιμερ (Alzheimer's disease, AD). Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, η άνοια πλήττει περίπου 50 εκατομμύρια ανθρώπους παγκοσμίως, με το AD να αποτελεί την κύρια αιτία, αντιπροσωπεύοντας το 60-70% των περιπτώσεων. Ο αριθμός αυτός προβλέπεται να τριπλασιαστεί έως το 2050, προκαλώντας σοβαρές επιπτώσεις στα συστήματα υγειονομικής περίθαλψης και στις οικογένειες λόγω της ανάγκης για μακροχρόνια φροντίδα, ιατρικές παρεμβάσεις και υποστήριξη των φροντιστών. Αυτή η δημογραφική αλλαγή υπογραμμίζει την επείγουσα ανάγκη για ανάπτυξη στρατηγικών με στόχο την προώθηση της υγιούς γήρανσης και την έγκαιρη διάγνωση νευροεκφυλιστικών ασθενειών.

Ο εγκέφαλος αποτελεί ένα ιδιαιτέρως ευαίσθητο όργανο σε ό,τι αφορά τη γήρανση. Η εγκεφαλική γήρανση είναι μια εξαιρετικά πολύπλοκη διαδικασία που επηρεάζεται από την αλληλεπίδραση γενετικών, περιβαλλοντικών, και παθολογικών παραγόντων. Αυτοί οι παράγοντες επιφέρουν ετερογενείς αλλαγές στη δομή και τη λειτουργικότητα του εγκεφάλου, οι οποίες, στη συνέχεια, επηρεάζουν καθοριστικά τις γνωστικές ικανότητες, τη μνήμη, την επεξεργασία πληροφοριών και την ικανότητα μάθησης του ατόμου.

Παράγοντες του τρόπου ζωής, όπως η διατροφή, η σωματική δραστηριότητα, η κοινωνική συναναστροφή και η γνωστική διέγερση, καθώς και η εκπαίδευση και η κοινωνικοοικονομική κατάσταση, επηρεάζουν σημαντικά την υγεία του εγκεφάλου καθ' όλη τη διάρκεια της ζωής του ατόμου. Σύμφωνα με σχετικές μελέτες, η τακτική φυσική άσκηση και η ισορροπημένη διατροφή έχουν αποδειχθεί ιδιαίτερα προστατευτικές έναντι της εγκεφαλικής ατροφίας. Αντίθετα, το χρόνιο στρες και η έλλειψη νοητικής διέγερσης μπορεί να συμβάλλουν στην επιτάχυνση της γνωστικής έκπτωσης. Επιπλέον, οι περιβαλλοντικοί παράγοντες, όπως η έκθεση σε ρύπους και τοξίνες επηρεάζουν την υγεία του εγκεφάλου και την πορεία της γήρανσης.

Παράλληλα, οι γενετικοί παράγοντες διαδραματίζουν καθοριστικό ρόλο στη διαδικασία αυτή. Για παράδειγμα, το αλληλόμορφο ε4 της Απολιποπρωτεΐνης Ε (Apolipoprotein E, APOE) έχει συνδεθεί με αυξημένο κίνδυνο εμφάνισης AD. Αυτή η γενετική ποικιλομορφία υποδηλώνει ότι ορισμένα άτομα διαθέτουν μεγαλύτερη ανθεκτικότητα στις επιπτώσεις της γήρανσης, ενώ άλλα

είναι περισσότερο επιρρεπή στη γνωστική παρακμή. Αυτό προσθέτει ένα επιπλέον επίπεδο πολυπλοκότητας στην κατανόηση της διαδικασίας γήρανσης.

Τέλος, η συννοσηρότητα κατά τη διάρκεια της γήρανσης, η οποία περιλαμβάνει μεταξύ άλλων παθήσεις όπως οι καρδιαγγειακές, οι εγκεφαλοαγγειακές και ο διαβήτης μπορούν να επιταχύνουν τη γήρανση του εγκεφάλου. Η παρουσία αμυλοειδών πλακών, νευροϊνιδιακών δεσμίδων και βλαβών της λευκής ουσίας περιπλέκει περισσότερο τη διαδικασία της γήρανσης, καθώς αυτές οι παθολογίες μπορεί να επικαλύπτονται με τις τυπικές διαδικασίες γήρανσης, καθιστώντας δύσκολη τη διάκριση μεταξύ της φυσιολογικής/τυπικής γήρανσης και των πρώιμων σταδίων νευροεκφυλιστικών νοσημάτων.

Λόγω της μοναδικής αλληλεπίδρασης αυτών των παραγόντων σε κάθε άτομο, η γήρανση του εγκεφάλου παρουσιάζει σημαντική ετερογένεια μέσα στον πληθυσμό, γεγονός που υπογραμμίζει την ανάγκη για εξατομικευμένες προσεγγίσεις στη μελέτη και στο σχεδιασμό της θεραπείας. Η παρούσα διατριβή επιχειρεί να ανταποκριθεί σε αυτή την πρόκληση διερευνώντας τις σύνθετες αλλαγές που εμφανίζονται στον εγκέφαλο κατά τη διαδικασία της γήρανσης, οι οποίες συμβάλλουν στη γνωστική έκπτωση και στην ανάπτυξη AD. Μέσω της χρήσης και της εξέλιξης σύγχρονων τεχνικών μηχανικής μάθησης (machine learning, ML) που εφαρμόζονται σε δεδομένα μαγνητικής τομογραφίας (magnetic resonance imaging, MRI) από μεγάλα σύνολα δεδομένων, η διατριβή εντοπίζει διακριτά μοτίβα γήρανσης με συγκεκριμένα χαρακτηριστικά και πορείες εξέλιξης στο χρόνο. Ειδικότερα, εξετάζεται η νευροανατομική ετερογένεια σε όλο το φάσμα της εγκεφαλικής γήρανσης, από ασυμπτωματικά στάδια μέχρι τα τελικά στάδια του AD, διερευνώντας την επίδραση γενετικών, περιβαλλοντικών και παθολογικών παραγόντων. Επιπλέον, τα αναδυόμενα μοτίβα νευροανατομικών αλλοιώσεων χρησιμοποιούνται για την πρόβλεψη της πιθανότητας μελλοντικής εμφάνισης ή επιδείνωσης της γνωστικής εξασθένησης.

Η συμβολή της παρούσας διατριβής διαρθρώνεται σε δύο θεμελιώδεις άξονες:

- Διερευνά την ετερογένεια των νευροανατομικών μεταβολών του εγκεφάλου σε ασυμπτωματικά στάδια, παρέχοντας πολύτιμες πληροφορίες σχετικά με πρώιμους δείκτες γνωστικής έκπτωσης. Αυτή η προσέγγιση φωτίζει τις ποικιλόμορφες εγκεφαλικές αλλαγές, οι οποίες μπορεί να προηγούνται ή να υποδεικνύουν την εμφάνιση γνωστικών διαταραχών.
- Εξελίσσει τις υφιστάμενες μεθόδους ανάλυσης της ετερογένειας μέσω της ανάπτυξης ενός καινοτόμου μοντέλου. Το νέο αυτό μοντέλο ενσωματώνει συγχρονικά (cross-sectional) και διαχρονικά (longitudinal) δεδομένα, προσφέροντας τη δυνατότητα ενός πιο ακριβούς χαρακτηρισμού των εγκεφαλικών αλλαγών που σχετίζονται με τη γήρανση.

1. Ανάλυση της ετερογένειας των νευροανατομικών αλλαγών του εγκεφάλου κατά τη γήρανση σε πρώιμα ασυμπτωματικά στάδια μέσω μηχανικής μάθησης και εκτενών συνόλων δεδομένων.

Η διαδικασία της γήρανσης, όπως έχει προαναφερθεί, είναι εξαιρετικά σύνθετη και διαφέρει μεταξύ ατόμων, καθώς προκύπτει από την αλληλεπίδραση γενετικών, περιβαλλοντικών και παθολογικών παραγόντων. Αυτοί οι παράγοντες άλλοτε δρουν ανεξάρτητα, άλλοτε συνεργιστικά και άλλοτε ανταγωνιστικά προκαλώντας περίπλοκες δομικές και λειτουργικές μεταβολές στον εγκέφαλο, οι οποίες εκδηλώνονται με ποικιλία κλινικών συμπτωμάτων. Οι κοινές νευροπαθολογίες που σχετίζονται με την ηλικία, όπως η νόσος του Αλτσχάιμερ και οι αγγειακές παθήσεις, συχνά παρουσιάζουν παρατεταμένες προκλινικές φάσεις κατά τις οποίες οι δομικές αλλαγές στον εγκέφαλο μπορούν να ανιχνευθούν μέσω της μαγνητικής τομογραφίας. Η έγκαιρη ανίχνευση αυτών των μεταβολών στα πρώιμα στάδια είναι κρίσιμη για την κατανόηση της ευαισθησίας των ατόμων στη διαδικασία της νευροεκφύλισης. Αυτή η διαδικασία δεν είναι σημαντική μόνο για την έγκαιρη παρέμβαση αλλά και για τη σωστή κατηγοριοποίηση των ασθενών, επιτρέποντας τη βέλτιστη οργάνωση και εκτέλεση κλινικών δοκιμών.

Στο παρελθόν, το περιορισμένο μέγεθος των δειγμάτων, καθώς και ο συνδυασμός συνόλων δεδομένων που έχουν ληφθεί με διαφορετικές μεθόδους και πρωτόκολλα απεικόνισης εμποδίζαν την ικανότητά ανίχνευσης των λεπτών διαφορών στη γήρανση του εγκεφάλου. Όμως, τα τελευταία χρόνια, οι μέθοδοι εναρμόνισης μεγάλων συνόλων δεδομένων, σε συνδυασμό με την ανάπτυξη εξελιγμένων τεχνικών μηχανικής μάθησης, προσφέρουν νέες δυνατότητες για την αναγνώριση των λεπτών και πολυσύνθετων νευροανατομικών μεταβολών στον εγκέφαλο με τη γήρανση.

Η παρούσα διατριβή αξιοποιεί μία ημι-επιβλεπόμενη (semi-supervised) τεχνική βαθιάς μάθησης (deep learning, DL), το Smile-GAN, για τον εντοπισμό υποομάδων δομικών εγκεφαλικών αλλαγών σε πληθυσμούς μέσης και προχωρημένης ηλικίας που δεν έχουν εκδηλώσει συμπτώματα γνωστικής εξασθένησης. Μέσω της υλοποίησης αυτής της προσέγγισης σε ένα εκτενές και εναρμονισμένο σύνολο δεδομένων που περιλαμβάνει περισσότερους από 27.000 συμμετέχοντες, στόχος είναι η ανίχνευση νευροανατομικών υποομάδων με διακριτά γενετικά, κλινικά και γνωστικά προφίλ. Αυτές οι υποομάδες παρέχουν σημαντική βάση για την κατανόηση των προκλινικών διαδικασιών νευροεκφύλισης και μπορούν να αξιοποιηθούν για την πρόωρη διάγνωση, τη διαστρωμάτωση των ασθενών και την εξατομίκευση της θεραπείας. Συγκεκριμένα, η ανάλυση αυτή αποσκοπεί 1) στον εντοπισμό υποομάδων

ατόμων με διακριτά χαρακτηριστικά δομικών αλλοιώσεων του εγκεφάλου σε ένα μεγάλο πληθυσμό μέσης και προχωρημένης ηλικίας χωρίς διαγνωσμένα συμπτώματα γνωστικής εξασθένησης, 2) στη συσχέτιση αυτών των υποομάδων με γενετικούς, καρδιαγγειακούς και περιβαλλοντικούς παράγοντες κινδύνου, αμυλοειδές β (amyloid β, Aβ) και επιδόσεις σε γνωστικές δοκιμασίες, και 3) στην αξιολόγηση της διαχρονικής σταθερότητας αυτών των υποομάδων, καθώς και της σημασίας τους για την πρόβλεψη εμφάνισης γνωστικής εξασθένησης.

Η παρούσα ανάλυση αξιοποίησε δεδομένα από την κοινοπραξία iSTAGING, η οποία αποτελεί μια σύμπραξη νευροαπεικονιστικών, κλινικών και γνωστικών δεδομένων από περισσότερους από 39.000 συμμετέχοντες παγκοσμίως. Συγκεκριμένα, επιλέχθηκαν 27.402 άτομα ηλικίας 45 έως 85 ετών χωρίς διαγνωσμένη γνωστική εξασθένηση (cognitively unimpaired, CU) κατά την έναρξη της μελέτης, με περισσότερα από 58.000 χρονικά σημεία δεδομένων προερχόμενα από τις παρακάτω μελέτες: Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging, Biomarker, and Lifestyle (AIBL) Study, Biomarkers of Cognitive Decline Among Normal Individuals (BIOCARD), Baltimore Longitudinal Study of Aging (BLSA), Coronary Artery Risk Development in Young Adults (CARDIA) study, Open Access Series of Imaging Studies (OASIS), University of Pennsylvania Memory Center cohort (Penn-PMC), Study of Health in Pomerania (SHIP), UK Biobank, Women's Health Initiative Memory Study (WHIMS), και Wisconsin Registry for Alzheimer's Prevention (WRAP).

Η προεπεξεργασία των εικόνων περιλάμβανε τη διόρθωση της ανομοιογένειας της έντασης της μαγνητικού πεδίου (correction of magnetic field intensity inhomogeneity) καθώς και την αφαίρεση του κρανίου από τις εικόνες μαγνητικής τομογραφίας (skull-striping). Στη συνέχεια, πραγματοποιήθηκε κατάτμηση των T1-εικόνων (T1-weighted) μαγνητικής τομογραφίας για την εξαγωγή των ανατομικών περιοχών του εγκεφάλου (regions of interest, ROIs) και των T2-εικόνων (T2-weighted) μαγνητικής τομογραφίας για την εξαγωγή των περιοχών βλαβών λευκής ουσίας (white matter hyperintensities, WMHs).

Στην ανάλυση χρησιμοποιήθηκαν 145 ROI όγκοι (volumes) όπου καλύπτουν ολόκληρο τον εγκέφαλο και 8 όγκοι WMH από τους τέσσερις διμερείς (bilateral) εγκεφαλικούς λοβούς. Η εναρμόνιση των ROIs μεταξύ των διαφόρων μελετών πραγματοποιήθηκε με τη χρήση της μεθόδου ComBat-GAM (General Additive Model) η οποία χρησιμοποιεί τον εμπειρικό κανόνα Bayes για τη προσαρμογή της μέσης τιμής και της διακύμανσης των δεδομένων που απορρέουν από τη χρήση διαφορετικών πρωτοκόλλων και μεθόδων απεικόνισης στις διάφορες μελέτες. Παράλληλα διατηρεί τη μεταβλητότητα των τιμών λόγω βιολογικών χαρακτηριστικών (πχ, ηλικία, φύλο, ενδοκρανιακός όγκος (intracranial volume,

ICV)) και συγκεκριμένα επιτρέπει τη μη-γραμμική μοντελοποίησή τους. Οι υποομάδες των δομικών εγκεφαλικών αλλαγών εξετάστηκαν ανεξάρτητα σε 4 ηλικιακά διαστήματα διάρκειας δεκαετίας, συγκεκριμένα 45-55 ετών και μέχρι 75-85 ετών, προκειμένου να ελαχιστοποιηθούν οι επιδράσεις που σχετίζονται με την ηλικία κατά την συσταδοποίηση των ατόμων.

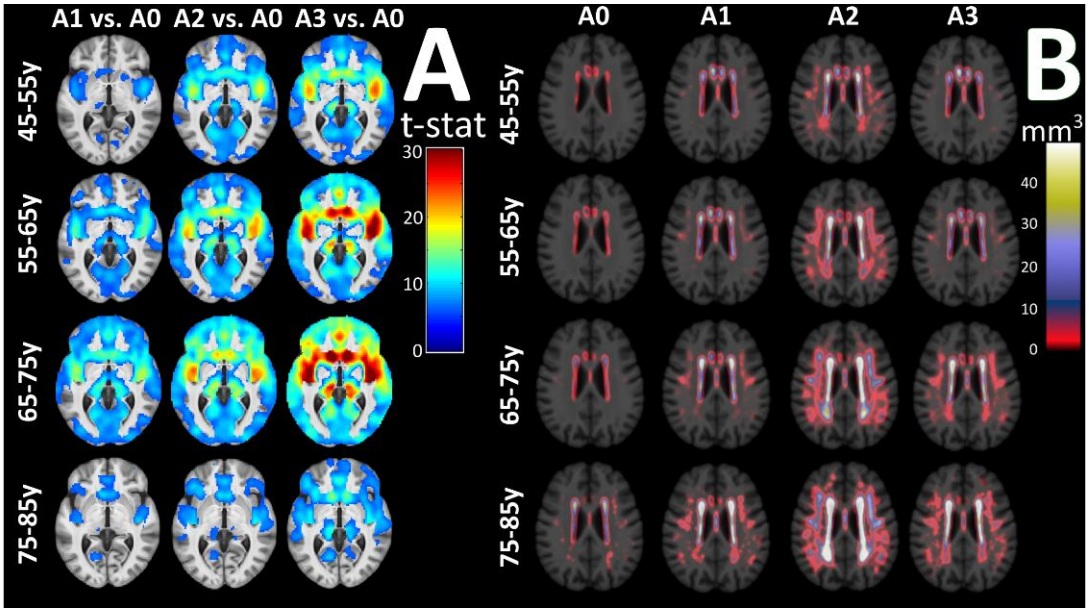
Σε κάθε ηλικιακή ομάδα, αρχικά εφαρμόστηκε ανάλυση κύριων συνιστωσών (principal component analysis, PCA) στους όγκους των 145 ROIs και των 8 WMHs, ξεχωριστά, με σκοπό τη μείωση της διαστατικότητας και την ανίχνευση μιας υποομάδας A0 (resilient brain aging) με χαμηλή ατροφία και περιορισμένους όγκους WMH. Χρησιμοποιώντας την υποομάδα A0 ως ομάδα αναφοράς, διερευνήθηκε η ετερογένεια εντός του υπόλοιπου πληθυσμού της ηλικιακής ομάδας με χρήση του μοντέλου Smile-GAN. Το Smile-GAN εφαρμόστηκε από κοινού στα 145 ROIs και 8 WMHs. Η έξοδος του Smile-GAN είναι μία πιθανότητα για κάθε υποομάδα με το άθροισμα των πιθανοτήτων να ισούται με 1. Η Smile-GAN ετικέτα (label) αποδίδεται βάσει της υποομάδας με τη μέγιστη πιθανότητα. Η επιλογή του αριθμού των υποομάδων καθορίστηκε μέσω του προσαρμοσμένου δείκτη Rand (adjusted rand index, ARI), ο οποίος αξιολογεί τη συμφωνία μεταξύ διαφόρων κατανεμημένων συνόλων και προσαρμόζεται για την τυχαιότητα. Τα μοντέλα των PCA και Smile-GAN εκπαιδεύτηκαν στα δεδομένα των πρώτων σαρώσεων (baseline scans) των ατόμων και στη συνεχεία εφαρμόστηκαν στα δεδομένα όλων των διαθέσιμων διαχρονικών σαρώσεων (longitudinal scans) εντός κάθε ηλικιακής ομάδας.

Το Smile-GAN είναι μία μέθοδος ημι-επιβλεπόμενης συσταδοποίησης (clustering) που βασίζεται σε παραγωγικά αντιπαραθετικά δίκτυα (generative adversarial networks, GAN). Σε αντίθεση με τις παραδοσιακές μεθόδους συσταδοποίησης, οι οποίες αναζητούν υποομάδες απευθείας μέσα στον πληθυσμό στόχο (target group) - συχνά επηρεαζόμενες από την μεταβλητότητα που σχετίζεται με χαρακτηριστικά ανεξάρτητα της μελετώμενης κατάστασης ή ασθένειας - το Smile-GAN εστιάζει στην εκμάθηση χαρτογραφήσεων μεταξύ του πληθυσμού αναφοράς (reference group), όπου εδώ είναι το A0, και του πληθυσμού στόχου, ο οποίος εδώ περιλαμβάνει τον υπόλοιπο πληθυσμό εκτός του A0. Αυτή η ημι-επιβλεπόμενη προσέγγιση διευκολύνει την αποτελεσματική μοντελοποίηση της ετερογένειας του πληθυσμού στόχου, ενώ δεν επηρεάζεται από τη φυσιολογική μεταβλητότητα που υπάρχει και στους δύο πληθυσμούς. Συγκεκριμένα, το Smile-GAN μαθαίνει διακριτές αντιστοιχίσεις από τον χώρο των υγιών (X) στο χώρο των ασθενών (Y) κατασκευάζοντας συνθετικά δεδομένα ασθενών (Y') που, χάρη στη διαδικασία της Παραγωγικής Αντιπαράθεσης, προσεγγίζουν όλο και περισσότερο τους πραγματικούς ασθενείς. Αυτή η διαδικασία υλοποιείται μέσω ενός παραγωγικού δικτύου (generator), το οποίο μαθαίνει μια συνάρτηση απεικόνισης (f) που συνδέει τον

χώρο των υγιών (X) και τον λανθάνοντα χώρο των υποτύπων (Z) με τον χώρο των ασθενών. Ως αποτέλεσμα, το παραγωγικό δίκτυο μετασχηματίζει δεδομένα υγιών (**x**) σε συνθετικά δεδομένα ασθενών (**y**′), σύμφωνα με μια μεταβλητή **z**, τα οποία δεν μπορούν να διακριθούν από τα πραγματικά δεδομένα ασθενών (**y**) από το διαχωριστικό δίκτυο (discriminator). Επιπλέον, επιβάλλεται στις συναρτήσεις f να παράγουν διακριτά μοτίβα ψευδοασθενών για διαφορετικές εισόδους **z**, επιτρέποντας στην αντίστροφη συνάρτηση g: Y → Z να ανιχνεύει αποτελεσματικά τη σωστή λανθάνουσα μεταβλητή ή υποτύπο **z**.

Σε σχέση με το A0, το Smile-GAN έδειξε μέγιστο ARI για 3 υποομάδες: A1, A2, και A3. Αν και η εξαγωγή των υποομάδων πραγματοποιήθηκε ανεξάρτητα σε κάθε ηλικιακή ομάδα, οι υποομάδες A1, A2 και A3, παρουσίασαν κοινές διαφορές σε σύγκριση με το αντίστοιχο A0 της ηλικιακής τους ομάδας (**Εικόνα 1**), αναφορικά με την ατροφία και το φορτίο των βλαβών στη λευκή ουσία στις τέσσερις ηλικιακές ομάδες.

Η παρούσα μελέτη χρησιμοποίησε την τεχνική της μορφομετρίας ογκοστοιχείων (voxel based morphometry, VBM), όπως υλοποιήθηκε μέσω του λογισμικού SPM (statistical parametric mapping) σε περιβάλλον MATLAB, έκδοσης R2017b (Mathworks Inc), για τη σύγκριση των μοτίβων της φαιάς ουσίας (gray matter) χρησιμοποιώντας χάρτες πυκνότητας ιστού (regional analysis of volumes examined in normalized space, RAVENS), και λαμβάνοντας υπόψη παραμέτρους όπως η ηλικία, το φύλο, και το ICV. Τα αποτελέσματα της ανάλυσης έδειξαν δομικές διαφορές μεταξύ των Smile-GAN υποομάδων σε σύγκριση με την υποομάδα A0 σε κάθε ηλικιακή ομάδα. Ειδικότερα, στην υποομάδα A1 παρατηρήθηκε ήπια ατροφία επικεντρωμένη στις περισυλβιακές (peri-Sylvian) περιοχές. Αντίθετα, η υποομάδα A2 παρουσίασε μέτρια ατροφία στις περισυλβιακές περιοχές, καθώς και στον κογχομετωπιαίο (orbitofrontal) φλοιό και σε διάφορες περιοχές του προμετωπιαίου (prefrontal) φλοιού. Η υποομάδα A3 παρουσίασε σοβαρή και διάχυτη ατροφία κυριότερα στις μεσαίες μετωπιαίες (frontal) περιοχές και στον θάλαμο (thalamus). Όσον αφορά τις βλάβες της λευκής ουσίας, η υποομάδα A2 παρουσίασε τις εντονότερες βλάβες. Μεταξύ των τριών Smile-GAN υποομάδων, η A1 είχε τη μικρότερη ατροφία και αποτέλεσε την πολυπληθέστερη υποομάδα, γεγονός που δηλώνει ότι μπορεί να θεωρηθεί ως η εκπρόσωπος της τυπικής ή συνήθους γήρανσης (typical brain aging). Συγκριτικά, η A2 (με υψηλότερο επίπεδο WMHs) και η A3 (με πιο σοβαρή ατροφία) μπορούν να χαρακτηριστούν ως υποομάδες προχωρημένης γήρανσης.



Εικόνα 1: Δομικά χαρακτηριστικά των υποομάδων γήρανσης του εγκεφάλου ανά ηλικιακή ομάδα. *Α) Ατροφία της φαιάς ουσίας για τις Smile-GAN υποομάδες σε σύγκριση με την υποομάδα Α0 σε κάθε ηλικιακή ομάδα, υπολογιζόμενη με χρήση τεχνικής μορφομετρίας ογκοστοιχείων (voxel based morphometry, VBM). Τα θερμότερα (ψυχρότερα) χρώματα αντιστοιχούν σε περιοχές με εντονότερη (χαμηλότερη) ατροφία. Έχει γίνει διόρθωση οικογενειακού ποσοστού σφάλματος (family-wise error rate, FWER) για πολλαπλές συγκρίσεις με όριο τιμής σημαντικότητας 0,001. Β) Χάρτες που απεικονίζουν τον όγκο των βλαβών λευκής ουσίας. Οι εν λόγω χάρτες έχουν δημιουργηθεί από το μέσο όρο των χαρτών βλαβών λευκής ουσίας των ατόμων κάθε υποομάδας. Ροζ (λευκά) χρώματα υποδεικνύουν περιοχές με χαμηλότερα (υψηλότερα) επίπεδα βλαβών λευκής ουσίας.*

Η εφαρμογή του Smile-GAN στα δεδομένα των διαχρονικών σαρώσεων αποκάλυψε ότι υπάρχει συνέπεια στην συσταδοποίηση των ατόμων κατά τη διάρκεια διαδοχικών ηλικιακών διαστημάτων. Ειδικότερα, οι Smile-GAN ετικέτες για τα δεδομένα των διαχρονικών σαρώσεων εντός ηλικιακής ομάδας έδειξαν συνέπεια της τάξεως του 85% (**Πίνακας 1**). Παρόμοια συνέπεια, ύψους 80%, παρατηρήθηκε και για τις Smile-GAN ετικέτες των δεδομένων των διαχρονικών σαρώσεων μεταξύ διαδοχικών ηλικιακών ομάδων (**Πίνακας 2**). Συνολικά, τα αποτελέσματα αυτής της μελέτης φανερώνουν μια ισχυρή διαχρονική σταθερότητα στις αναθέσεις των συμμετεχόντων σε υποομάδες, τόσο εντός των ηλικιακών διαστημάτων όσο και κατά τη μετάβαση μεταξύ διαδοχικών ηλικιακών ομάδων.

Πίνακας 1: Μέση μεταβολή της Smile-GAN πιθανότητας για κάθε υποομάδα μεταξύ δύο διαδοχικών σαρώσεων εντός της <u>ίδιας</u> ηλικιακής ομάδας. *2.775 άτομα έχουν τουλάχιστον δύο σαρώσεις εντός της ίδιας ηλικιακής ομάδας.*

Smile-GAN subgroup	**Mean Smile-GAN probability change**		
	A1(i+1)-A1(i)	**A2(i+1)-A2(i)**	**A3(i+1)-A3(i)**
	Age group [45,55)		
A1	0.02±0.25	-0.001±0.17	-0.02±0.16
A2	-0.09±0.22	0.18±0.30	-0.09±0.20
A3	-0.06±0.16	-0.01±0.18	0.07±0.19

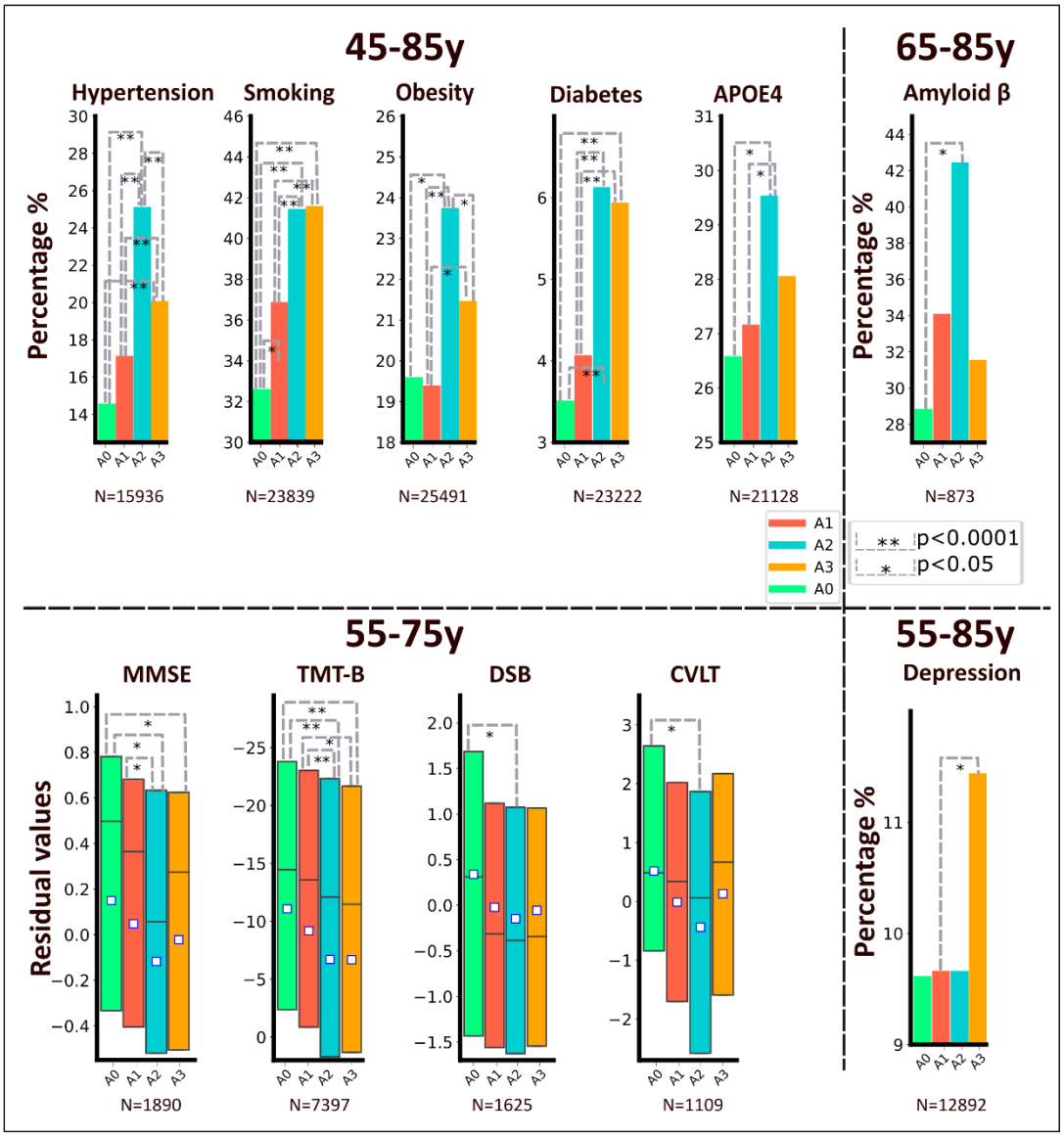
	Age group [55,65)		
A1	0.04±0.20	-0.03±0.15	-0.01±0.13
A2	-0.05±0.18	0.06±0.22	-0.002±0.14
A3	-0.08±0.18	-0.04±0.15	0.12±0.21
	Age group [65,75)		
A1	0.06±0.19	-0.06±0.14	0.01±0.13
A2	0.01±0.13	0.04±0.17	-0.05±0.12
A3	-0.07±0.17	-0.002±0.12	0.07±0.20
	Age group [75,85)		
A1	0.07±0.17	-0.03±0.13	-0.03±0.12
A2	-0.01±0.12	0.03±0.15	-0.02±0.09
A3	0.004±0.11	-0.01±0.11	0.005±0.15

Πίνακας 2: Μέση μεταβολή της Smile-GAN πιθανότητας για κάθε υποομάδα μεταξύ δύο διαδοχικών σαρώσεων σε διαφορετικές ηλικιακές ομάδες. *1.201 άτομα έχουν τουλάχιστον δύο σαρώσεις σε διαφορετικές ηλικιακές ομάδες.*

Smile-GAN subgroup	**Mean Smile-GAN probability change**		
	A1(i+1)-A1(i)	**A2(i+1)-A2(i)**	**A3(i+1)-A3(i)**
	Age group [45,55)		
A1	-0.01±0.25	-0.03±0.19	0.04±0.18
A2	-0.09±0.24	0.18±0.31	-0.09±0.24
A3	-0.09±0.23	0.06±0.21	0.03±0.28
	Age group [55,65)		
A1	0.10±0.26	-0.11±0.19	0.002±0.19
A2	-0.02±0.18	0.12±0.24	-0.10±0.19
A3	-0.09±0.23	0.01±0.17	0.08±0.26
	Age group [65,75)		
A1	0.07±0.26	-0.03±0.19	-0.03±0.19
A2	-0.03±0.15	0.02±0.22	0.003±0.16
A3	0.002±0.16	-0.08±0.22	0.08±0.26

Στη συνέχεια, εξετάστηκαν οι συσχετίσεις των υποομάδων με κλινικά χαρακτηριστικά, γνωστικές επιδόσεις, βιοδείκτες και την Απολιποπρωτεΐνη Ε με χρήση γραμμικών (linear regression) και λογιστικών μοντέλων παλινδρόμησης (logistic regression) (**Εικόνα 2**). Σημειώθηκαν διορθώσεις για παράγοντες όπως η ηλικία, το φύλο, η μελέτη προέλευσης και το επίπεδο εκπαίδευσης. Η υποομάδα A2 με το υψηλότερο όγκο βλαβών λευκής ουσίας είχε επίσης το υψηλότερο ποσοστό συμμετεχόντων με παράγοντες καρδιαγγειακού κινδύνου, όπως υπέρταση και παχυσαρκία. Επίσης, η A2 παρουσίασε και το υψηλότερο ποσοστό συμμετεχόντων με το αλληλόμορφο ε4 του γονιδίου της Απολιποπρωτεΐνης. Επιπλέον, η υποομάδα A2 έχει το υψηλότερο ποσοστό συμμετεχόντων με επίπεδα Αβ ανώτερα από το κλινικό κατώφλι, καθιστώντας

τους θετικούς σε σύγκριση με τους συμμετέχοντες με χαμηλά επίπεδα Αβ, ιδίως μετά την ηλικία των 65 ετών.

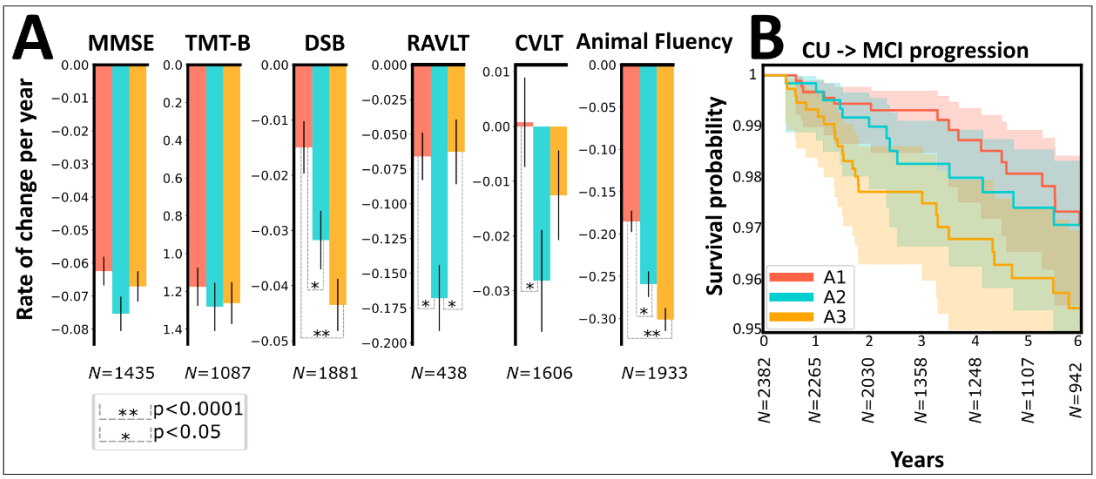


Εικόνα 2: Κλινικές μετρήσεις, γνωστικές δοκιμασίες, αμυλοειδής β πρωτεΐνη και αλληλόμορφο ε4 της Απολιποπρωτεΐνης για τις υποομάδες εγκεφαλικής γήρανσης. *Τα εικονιζόμενα χαρακτηριστικά αναφέρονται σε μη απεικονιστικά χαρακτηριστικά που εμφάνισαν συνεπείς τάσεις σε περισσότερες από μία ηλικιακές ομάδες, και παρουσιάζονται ως σύνοψη μετά την ανάλυση των δεδομένων από όλες τις αντίστοιχες ηλικιακές ομάδες. Τα ηλικιακά εύρη που αναγράφονται άνω των διαγραμμάτων υποδεικνύουν τις ευρύτερες ηλικιακές κατηγορίες που εξετάστηκαν. Φορείς του ε4 αλληλόμορφου της Απολιποπρωτεΐνης (APOE-ε4 carriers or APOE4) θεωρούνται τα άτομα που διαθέτουν ένα ή δύο αλληλόμορφα ε4. Τα διαγράμματα κουτιού απεικονίζουν τις υπολειμματικές τιμές, μετά από προσαρμογή για την ηλικία, το φύλο και την μελέτη προέλευσης, καθώς και την εκπαίδευση για τις επιδόσεις των γνωστικών δοκιμασιών, για κάθε υποομάδα. Υψηλότερες τιμές στις γνωστικές δοκιμασίες MMSE, DSB και CVLT υποδηλώνουν ανώτερες γνωστικές επιδόσεις, ενώ χαμηλότερες τιμές στην TMT-B υποδηλώνουν επίσης ανώτερες γνωστικές επιδόσεις – οι βαθμολογίες της TMT παρατίθενται με ανεστραμμένη κλίμακα, προκειμένου όλες οι παρατηρούμενες επιδόσεις να παρουσιάζουν την ίδια κατεύθυνση στα τέσσερα γραφήματα. Το N δηλώνει το μέγεθος του δείγματος που χρησιμοποιήθηκε για το κάθε γράφημα. Τέλος, εφαρμόστηκε διόρθωση του ποσοστού ψευδών ανακαλύψεων (False Discovery Rate, FDR) για τις πολλαπλές συγκρίσεις, με κατώφλι τιμής σημαντικότητας στο 0,05.*

Όσον αφορά τις διαφορές επιδόσεων σε γνωστικές δοκιμασίες, παρά το ότι οι συμμετέχοντες επελέγησαν ως άτομα χωρίς γνωστική εξασθένηση, οι υποομάδες A2 και A3 παρουσίασαν στατιστικά σημαντικά χειρότερη επίδοση σε σχέση με τις υπόλοιπες υποομάδες σε ποικίλα γνωστικά τεστ (**Εικόνα 2**). Αυτή η διαπίστωση φανερώνει τις προσθετικές επιδράσεις της ατροφίας και των βλαβών λευκής ουσίας στη γνωστική εξασθένηση. Τέλος, η υποομάδα A3 είχε το μεγαλύτερο ποσοστό συμμετεχόντων με κατάθλιψη μετά την ηλικία των 55 ετών.

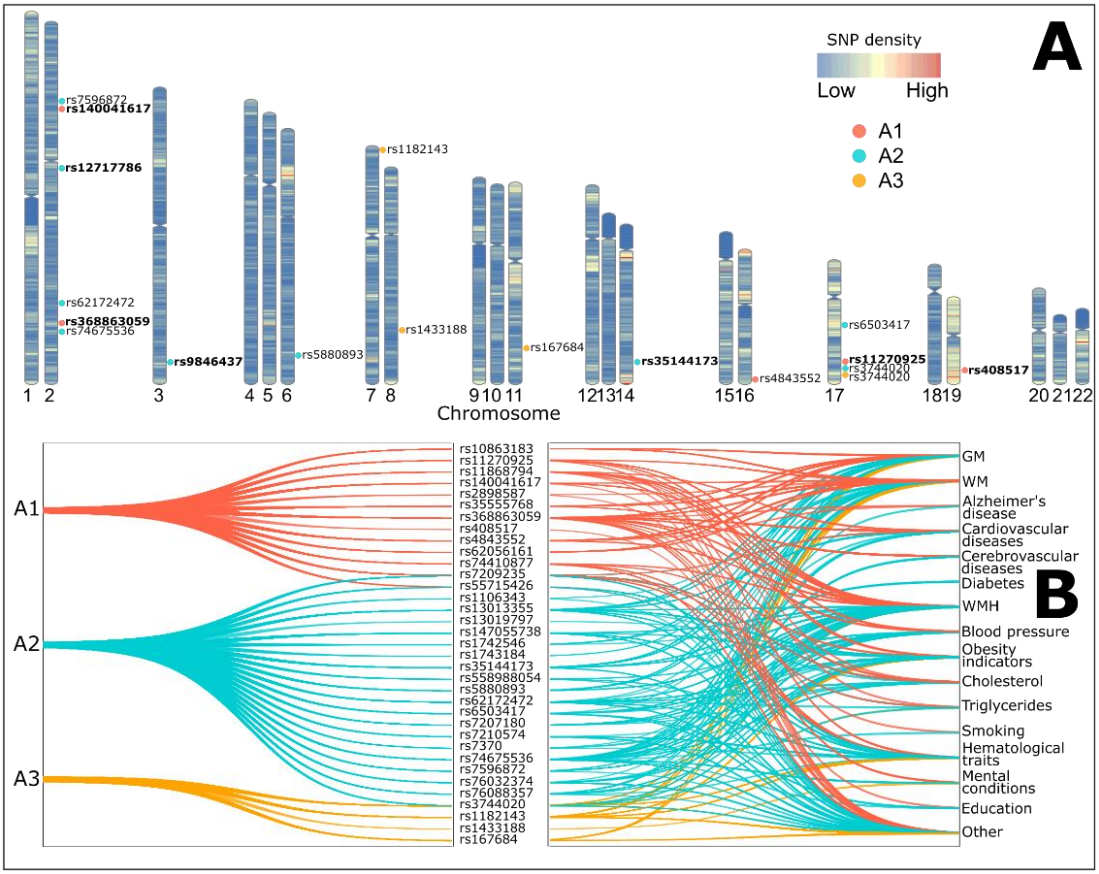
Εν συνεχεία, ο ρυθμός μεταβολής των διαφόρων χαρακτηριστικών αξιολογήθηκε μέσω της εφαρμογής γραμμικών μοντέλων μικτών επιδράσεων (linear mixed-effects models), τα οποία περιλάμβαναν τυχαίες τομές (random intercept) για το κάθε άτομο. Επιπροσθέτως, πραγματοποιήθηκε ανάλυση επιβίωσης (survival analysis) τύπου Kaplan-Meier για την εκτίμηση του χρόνου μετάβασης από τη γνωσιακά υγιή κατάσταση στην ήπια γνωστική διαταραχή (mild cognitive impairment, MCI). Τα αποτελέσματα της ανάλυσης έδειξαν ότι οι υποομάδες A2 και A3 παρουσίασαν τον υψηλότερο ρυθμό γνωστικής έκπτωσης και την ταχύτερη μετάβαση σε MCI (**Εικόνα 3**).

Μετά την εξαγωγή των υποομάδων, διενεργήθηκαν μελέτες συσχέτισης εύρους γονιδιώματος (genome wide association studies, GWAS) χρησιμοποιώντας γενετικά δεδομένα από τη UK Biobank, προκειμένου να εντοπιστούν οι συσχετίσεις μεταξύ των πιθανοτήτων των Smile-GAN υποομάδων και μονονουκλεοτιδικών πολυμορφισμών (single nucleotide polymorphisms, SNPs). Οι συσχετίσεις εξετάστηκαν με τη χρήση γραμμικής παλινδρόμησης, προσαρμοσμένης για την ηλικία, το φύλο, το ICV, και τις πρώτες 40 γενετικές κύριες συνιστώσες (principal components), χρησιμοποιώντας το λογισμικό Plink 2. Δεδομένης της παρατηρούμενης διαχρονικής σταθερότητας της συσταδοποίησης, οι GWAS διεξήχθησαν από κοινού στο συνολικό ηλικιακό εύρος των 45-85 ετών. Έπειτα πραγματοποιήθηκε λειτουργική επισημείωση και γονιδιακή χαρτογράφηση των στατιστικά σημαντικών πολυμορφισμών μέσω της πλατφόρμας FUMA (functional mapping and annotation).



Εικόνα 3: Διαχρονικά αποτελέσματα. *Α) Ρυθμός μεταβολής ανά έτος των επιδόσεων σε γνωστικές δοκιμασίες. Οι ρυθμοί μεταβολής υπολογίστηκαν με τη χρήση γραμμικών μοντέλων μικτών επιδράσεων (linear mixed-effects models). Οι συγκρίσεις των ρυθμών μεταβολής μεταξύ των υποομάδων πραγματοποιήθηκαν μέσω της μεθόδου Wald. Το N αναφέρεται στον αριθμό των ατόμων που διαθέτουν τουλάχιστον 4 διαχρονικές μετρήσεις για το εικονιζόμενο χαρακτηριστικό. Εφαρμόστηκε διόρθωση του ποσοστού ψευδών ανακαλύψεων (False Discovery Rate, FDR) για πολλαπλές συγκρίσεις με κατώφλι τιμής σημαντικότητας 0,05. Οι ρυθμοί μεταβολής των βαθμολογιών του TMT-B παρουσιάζονται με ανεστραμμένη κλίμακα, έτσι ώστε η ταχύτερη γήρανση του εγκεφάλου (που αντικατοπτρίζεται από την ταχύτερη είτε ατροφία, συσσώρευση βλαβών ή γνωστική έκπτωση) να έχει την ίδια κατεύθυνση σε όλα τα γραφήματα. Β) Οι καμπύλες επιβίωσης Kaplan-Meier δείχνουν την πιθανότητα παραμονής σε γνωσιακά υγιή κατάσταση (cognitively unimpaired, CU) και αποφυγής μετάβασης σε ήπια γνωστική διαταραχή (mild cognitive impairment, MCI) για τα άτομα με ηλικία πρώτης σάρωσης μεταξύ 65-75 ετών. Το N υποδηλώνει τον αριθμό των ατόμων σε κάθε χρονικό διάστημα. Το τεστ Log-rank χρησιμοποιήθηκε για τη σύγκριση των καμπυλών επιβίωσης των Smile-GAN υποομάδων. Η μόνη στατιστικά σημαντική διαφορά είναι μεταξύ των καμπυλών των υποομάδων A1 και A3 (τιμή σημαντικότητας=0,01). Τα διαχρονικά αποτελέσματα για την A0 δεν παρουσιάζονται, καθώς προήλθαν από διαφορετική μεθοδολογία σε σύγκριση με τις υποομάδες Smile-GAN. Επιπρόσθετα, ο περιορισμένος αριθμός ατόμων της A0 με διαχρονικά αποτελέσματα καθιστά τα αποτελέσματα μη αξιόπιστα.*

Οι GWAS έδειξαν ότι οι πιθανότητες των Smile-GAN υποομάδων έχουν στατιστικά σημαντικές συσχετίσεις με μονονουκλεοτιδικούς πολυμορφισμούς που είχαν προηγουμένως συσχετιστεί με διάφορα κλινικά χαρακτηριστικά, συμπεριλαμβανομένων φαινοτύπων που προέκυψαν από την απεικόνιση της μικροδομής της λευκής ουσίας (A1-3), της ατροφίας της φαιάς ουσίας (A1-3), των βλαβών της λευκής ουσίας (A1-3), καθώς και παράγοντες κινδύνου για καρδιαγγειακή νόσο (A1-2) και νόσο Αλτσχάιμερ (A1-2) (**Εικόνα 4**). Είναι αξιοσημείωτο ότι εντοπίστηκαν κοινοί πολυμορφισμοί μεταξύ των υποομάδων A1 και A2 οι οποίοι είχαν διαφορετικές επιδράσεις στις δύο υποομάδες. Συγκεκριμένα, ο πολυμορφισμός rs72932727 που σχετίζεται με AD είχε προστατευτική δράση για την υποομάδα A1 (beta=0.1±0.02; p-value=6.49E-09), ενώ αντίθετα αποτέλεσε παράγοντα κινδύνου για την υποομάδα A2 (beta=-0.09±0.02; p-value=4.05E-07). Παρομοίως, οι rs7209235 και rs55715426, οι οποίοι σχετίζονται με βλάβες στη λευκή ουσία, επέδειξαν προστατευτική δράση για την υποομάδα A1, ενώ λειτούργησαν ως παράγοντες κινδύνου για την A2 (rs7209235: A1:beta=-0.07±0.01, p-value=2.31E-09, και A2:beta=0.1±0.01, p-value=1.73E-15; rs55715426: A1: beta=-0.09±0.02, p-value=4.09E-08, και A2: beta=0.13±0.02; p-value=1.04E-15).



Εικόνα 4: Γενετικές αναλύσεις των πιθανοτήτων των Smile-GAN υποομάδων (A1, A2 και A3). *Α) Οι μελέτες συσχέτισης εύρους γονιδιώματος (genome wide association studies, GWAS) εντόπισαν μονονουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms, SNPs) που σχετίζονται με τις πιθανότητες των Smile-GAN υποομάδων (A1, A2 και A3). Στις εν λόγω μελέτες χρησιμοποιήθηκε το κατώτατο όριο τιμών σημαντικότητας σε επίπεδο γονιδιώματος (5E-08). Το γονιδίωμα αναφοράς είναι το Genome Reference Consortium Human Build 37 (GRCh37). Το διάγραμμα ιδεογραμμάτων αντιστοιχεί στα 22 αυτοσωμικά χρωμοσώματα. Β) Φαινοτυπικές συσχετίσεις από τον κατάλογο GWAS. Οι ευρισκόμενοι μονονουκλεοτιδικοί πολυμορφισμοί συσχετίστηκαν με διάφορα κλινικά χαρακτηριστικά συμπεριλαμβανομένων μετρήσεων της φαιάς (π.χ. (υπο)φλοιώδης όγκος, πάχος και επιφάνεια φλοιού), και της λευκής ουσίας (π.χ. ανισοτροπία στη διάχυση του νερού), καρδιαγγειακές παθήσεις (π.χ. στεφανιαία νόσος και έμφραγμα του μυοκαρδίου), νόσο Αλτσχάιμερ, αιματολογικά χαρακτηριστικά (π.χ. αριθμός αιμοπεταλίων και λευκών αιμοσφαιρίων), ψυχικές διαταραχές (π.χ. συμπεριφορά ανάληψης κινδύνου και απόπειρες αυτοκτονίας), και εκπαίδευση.*

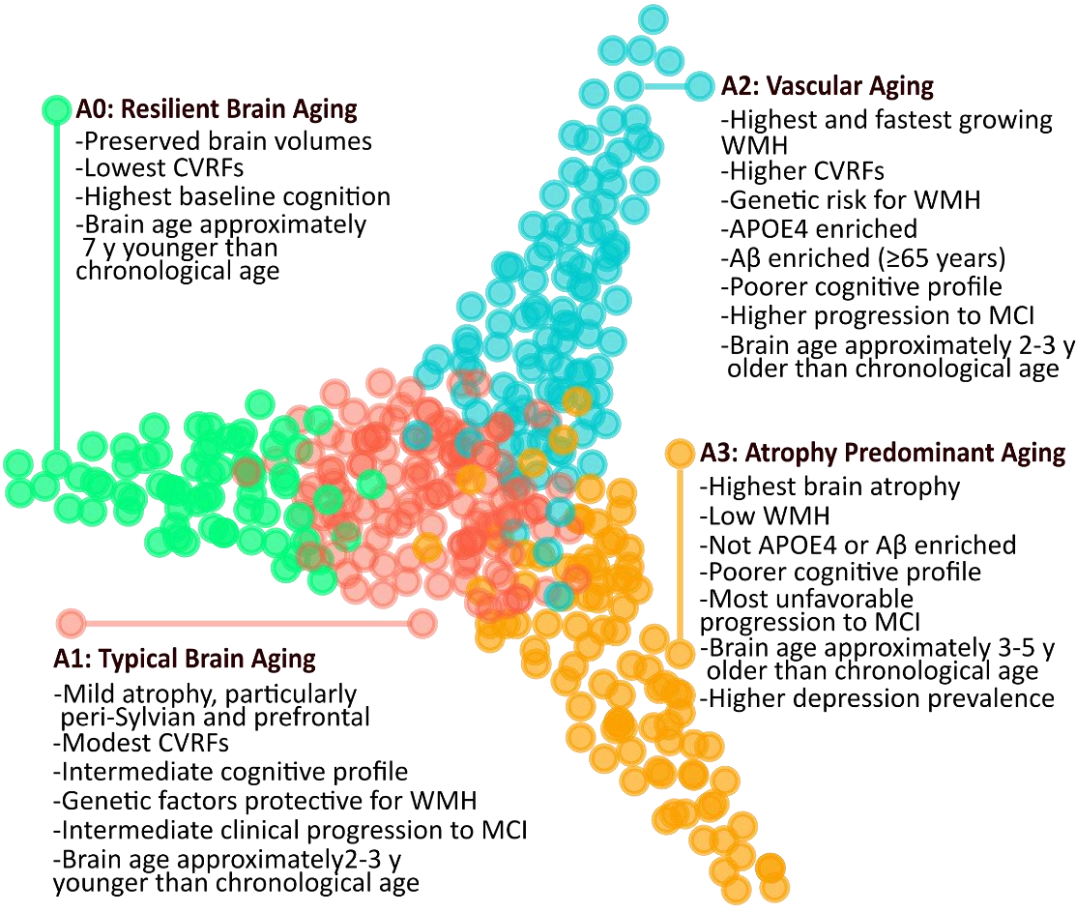
Συμπερασματικά, η παρούσα εργασία αξιοποίησε τις πρόσφατες εξελίξεις στον τομέα της βαθιάς μάθησης και της ανάλυσης δεδομένων μεγάλης κλίμακας προκειμένου να διερευνήσει το φάσμα μεταξύ της φυσιολογικής εγκεφαλικής γήρανσης και της πρώιμης παθολογίας. Συγκεκριμένα, οι πρόοδοι στη συγκέντρωση και εναρμόνιση πολυάριθμων συνόλων δεδομένων επέτρεψαν τη δημιουργία ενός εναρμονισμένου δείγματος 27.402 ατόμων ηλικίας 45-85 ετών, χωρίς διαγνωσμένη γνωστική εξασθένηση, προερχόμενο από 11 μελέτες. Επιπλέον, εφαρμόστηκε μια τεχνική ημι-επιβλεπόμενης συσταδοποίησης για την εξέταση της ετερογένειας των πρώιμων, συχνά ανεπαίσθητων, νευροανατομικών αλλαγών σε αυτόν τον πληθυσμό.

Παρά το γεγονός ότι η ανάλυση πραγματοποιήθηκε ξεχωριστά για κάθε ηλικιακή ομάδα, τα ευρήματά έδειξαν ότι οι πρώιμες δομικές αλλοιώσεις είναι σχετικά ομοιόμορφες και συνεπείς σε ολόκληρο το εξεταζόμενο ηλικιακό φάσμα. Συγκεκριμένα, οι παρατηρήσεις κατέδειξαν ποικιλομορφία δομικών αλλοιώσεων στον εγκέφαλο που σχετίζονται με τη γήρανση, οι οποίες κατηγοριοποιήθηκαν σε τέσσερις υποομάδες, ανιχνεύσιμες από τη μέση ηλικία και με συνεπή μοτίβα μεταξύ των εξεταζόμενων ηλικιακών ομάδων.

Πιο αναλυτικά, εντοπίστηκε μια υποομάδα (A0) με ανθεκτικότητα στη γήρανση, χωρίς ένδειξη εγκεφαλικής ατροφίας, βλαβών στη λευκή ουσία, γνωστικής έκπτωσης ή παραγόντων κινδύνου για καρδιαγγειακά νοσήματα, μια υποομάδα (A1) με τυπικά χαρακτηριστικά γήρανσης, συμπεριλαμβανομένων μέτριων επιπέδων ατροφίας και βλαβών λευκής ουσίας και, δύο υποομάδες με χαρακτηριστικά προχωρημένης γήρανσης: η μία (A2) εμφάνισε αυξημένους παράγοντες κινδύνου καρδιαγγειακών νοσημάτων και εναπόθεση Αβ, και η άλλη (A3) παρουσίασε έντονη και διάχυτη ατροφία, κύρια προερχόμενη από περιβαλλοντικούς παράγοντες και τρόπο ζωής (**Εικόνα 5**) . Είναι αξιοσημείωτο ότι παρά τις διαφορές στα χαρακτηριστικά ατροφίας, οι υποομάδες A2 και A3 εμφάνισαν συγκρίσιμα χειρότερες επιδόσεις στις γνωστικές λειτουργίες σε σχέση με την A0. Στην περίπτωση αυτή, η ατροφία και οι βλάβες της λευκής ουσίας φαίνεται να δρουν συνδυαστικά, προξενώντας γνωστική έκπτωση, γεγονός που μπορεί να εξηγήσει τη μειωμένη ατροφία στην υποομάδα A2 σε σύγκριση με την υποομάδα A3.

Καμία από τις παραπάνω υποομάδες δεν μπορεί να χαρακτηριστεί ως πρώιμο στάδιο της νόσου του Αλτσχάιμερ, γεγονός που υποδηλώνει ότι υπάρχουν πολλαπλά μονοπάτια επιταχυνόμενης γήρανσης του εγκεφάλου, τα οποία μπορεί να οδηγήσουν στη νόσο. Συνολικά, τα ευρήματα αναδεικνύουν κυρίαρχες υποομάδες ανθεκτικότητας και ευπάθειας στην εμφάνιση και επιδείνωση της γνωστικής έκπτωσης καθώς και την εμφάνιση AD.

Η παρούσα μελέτη παρουσιάζει αρκετά δυνατά σημεία, όπως είναι το ευρύ και ποικιλόμορφο δείγμα το οποίο καλύπτει ευρύ φάσμα ηλικιών, καθώς και η εφαρμογή προηγμένων μεθόδων εναρμόνισης και βαθιάς μάθησης. Επιπροσθέτως, η εύρεση πολλών μονονουκλεοτιδικών πολυμορφισμών που σχετίζονται με την ύπαρξη βλαβών λευκής ουσίας και την εγκεφαλική ατροφία ευθυγραμμίζεται με το νευροαπεικονιστικό προφίλ των υποομάδων. Παρόλα αυτά, η μελέτη παρουσιάζει και ορισμένες αδυναμίες.



Εικόνα 5: Σχηματική σύνοψη των βασικών χαρακτηριστικών των υποομάδων γήρανσης του εγκεφάλου.

Πρώτον, η περιορισμένη διαθεσιμότητα δεδομένων Αβ και η ανεπαρκής διαθεσιμότητα μετρήσεων της tau πρωτεΐνης και βιοδεικτών που σχετίζονται με άλλες νευροεκφυλιστικές νόσους πέραν του AD, ενδέχεται να συμβάλουν σε κενά στην ερμηνεία των υποομάδων. Δεύτερον, η απουσία μακροχρόνιας παρακολούθησης των ατόμων αποτρέπει την εξαγωγή ισχυρών συμπερασμάτων σχετικά με την κλινική εξέλιξη και τη μετάβαση σε MCI. Τρίτον, όσον αφορά τη σύνθεση του δείγματος, παρατηρείται το φαινόμενο 'οροφής' (ceiling effect), καθώς τα άτομα με πιο σοβαρές νευροανατομικές αλλοιώσεις είναι πιο πιθανό να διαγνωσθούν με γνωστική εξασθένηση, αποκλείοντας έτσι την ένταξή τους στο δείγμα. Τέταρτον, αν και έχουν παρατηρηθεί μορφολογικές και συσχετιστικές ομοιότητες των υποομάδων στις διαφορετικές δεκαετίες, δεν μπορεί να αποδειχθεί με απόλυτη σαφήνεια η ισοδυναμία τους, καθώς χρησιμοποιήθηκαν διαφορετικά μοντέλα και ομάδες αναφοράς (A0) ανά δεκαετία, και δεν υπήρξε μακροχρόνια παρακολούθηση των ατόμων σε διάστημα δεκαετιών. Τέλος, η ομάδα αναφοράς (A0), παρά την εκπροσώπηση ενός πιο ανθεκτικού στη γήρανση πληθυσμού, εξακολουθεί να παρουσιάζει ένα επίπεδο παθολογίας που αυξάνεται με την πάροδο των δεκαετιών.

2. Ανάπτυξη καινοτόμου μεθόδου βασισμένης σε μη-αρνητική παραγοντοποίηση πινάκων που ενσωματώνει συγχρονικά και διαχρονικά δεδομένα για τη μελέτη των μοτίβων εγκεφαλικής γήρανσης.

Μέχρι σήμερα, οι τεχνικές συσταδοποίησης έχουν προσφέρει σημαντικές δυνατότητες όσον αφορά τη μελέτη της ετερογένειας που παρατηρείται στη γήρανση του εγκεφάλου, καθώς και στις σχετικές νευροεκφυλιστικές διαταραχές. Ωστόσο, η κατανόηση της ετερογένειας παραμένει μια δύσκολη πρόκληση. Οι υπάρχουσες μεθοδολογίες βασίζονται αποκλειστικά σε συγχρονικά δεδομένα, δηλαδή μία μέτρηση ανά άτομο, για την εύρεση μοτίβων γήρανσης, αγνοώντας τη δυναμική εξέλιξη της γήρανσης και των λοιπών παθολογικών καταστάσεων σε χρονικό βάθος. Η παρούσα διατριβή εισάγει μία νέα μέθοδο, ονόματι από κοινού μη-αρνητική παραγοντοποίηση συγχρονικών και διαχρονικών πινάκων (Coupled Cross-sectional and Longitudinal – Non-negative matrix factorization, CCL-NMF), για τον εντοπισμό διακριτών μοτίβων εγκεφαλικής γήρανσης μέσω κοινής βελτιστοποίησης της ανακατασκευής συγχρονικών και διαχρονικών δεδομένων. Αυτή η μεθοδολογία μπορεί να εφαρμοστεί για τη μελέτη διαφόρων χαρακτηριστικών που σχετίζονται με την εγκεφαλική γήρανση όπως η σταδιακή συρρίκνωση (ατροφία) της φαιάς ουσίας, η προοδευτική εναπόθεση αμυλοειδούς β και tau πρωτεϊνών, η αυξανόμενη συσσώρευση βλαβών της λευκής ουσίας στον εγκέφαλο, κ.λπ. Στη παρούσα διατριβή, η συγκεκριμένη μεθοδολογία εφαρμόζεται για τη μελέτη της εγκεφαλικής ατροφίας με τη γήρανση.

Αν και τα συγχρονικά δεδομένα είναι ευρέως διαθέσιμα και καταγράφουν σωρευτικές και μακροχρόνιες αλλαγές στον εγκέφαλο λόγω γήρανσης, επιτρέπουν την προσεγγιστική εκτίμηση των αποκλίσεων των ηλικιωμένων εγκεφάλων από την νέα και υγιή κατάσταση τους λόγω της έλλειψης εξατομικευμένων βάσεων σύγκρισης. Εφόσον είναι διαθέσιμη μία μέτρηση ανά άτομο, οι αποκλίσεις προκύπτουν από συγκρίσεις με ευρύτερες πληθυσμιακές κατανομές. Αντίθετα, τα διαχρονικά δεδομένα προσφέρουν μια άμεση, εξατομικευμένη ποσοτικοποίηση των εγκεφαλικών αλλαγών με την πάροδο του χρόνου, παρέχοντας πολύτιμες πληροφορίες για τη δυναμική εξέλιξη των νευροβιολογικών διαδικασιών. Ωστόσο, τα διαχρονικά σύνολα δεδομένων είναι σπανιότερα. Η μεθοδολογία CCL-NMF είναι σχεδιασμένη για να αντιμετωπίσει αυτούς τους περιορισμούς, αναλύοντας συνδυαστικά τους δύο τύπους δεδομένων με στόχο την εξαγωγή διακριτών διαστάσεων (dimensions) που αναπαριστούν τις συντονισμένες μεταβολές που συμβαίνουν στον εγκέφαλο με τη γήρανση και την παθολογία. Η εν λόγω μεθοδολογία βασίζεται στη μη-αρνητική παραγοντοποίηση πίνακα (non-negative matrix factorization, NMF), μια τεχνική ευρέως χρησιμοποιούμενη σε ποικιλία ερευνητικών τομέων, η οποία

επιδεικνύει εξαιρετικές δυνατότητες χάρη στον περιορισμό της μη-αρνητικότητας. Αυτός ο περιορισμός οδηγεί σε αναπαράσταση των δεδομένων σε μέρη (part-based representation), όπου τα μέρη συνδυάζονται με προσθετικό τρόπο για να σχηματίσουν ένα σύνολο. Στη συγκεκριμένη περίπτωση, η παραγοντοποίηση διεξάγεται από κοινού σε δύο πίνακες: ο πίνακας συγχρονικών αποκλίσεων (cross-sectional map, C-map) αναπαριστά τις αποκλίσεις των εγκεφαλικών χαρακτηριστικών των ηλικιωμένων ατόμων (πληθυσμός στόχος) σε σχέση με έναν υγιή και νεότερης ηλικίας πληθυσμό (πληθυσμός αναφοράς), ενώ ο πίνακας διαχρονικών αλλαγών (longitudinal map, L-map) αποτυπώνει τους ρυθμούς αλλαγής των εγκεφαλικών χαρακτηριστικών λόγω γήρανσης σε ατομικό επίπεδο. Η μεθοδολογία CCL-NMF περιλαμβάνει την παραγοντοποίηση κάθε πίνακα σε δύο νέους πίνακες μικρότερων μεγεθών: ο πρώτος πίνακας απεικονίζει τα μοτίβα εγκεφαλικών αλλαγών (πίνακας λεξικό/βάση) και είναι κοινός για τους δύο τύπους δεδομένων και ο δεύτερος πίνακας (πίνακας συντελεστών/βαρών) προσδιορίζει τον βαθμό έκφρασης κάθε μοτίβου σε ατομικό επίπεδο και διαφοροποιείται για τα συγχρονικά και τα διαχρονικά δεδομένα.

Η εκτίμηση του C-map υλοποιείται μέσω μιας κανονιστικής προσέγγισης μοντελοποίησης (normative modeling), η οποία εκπαιδεύεται σε έναν υγιή πληθυσμό μέσης ηλικίας με σκοπό τη δημιουργία ενός κανονιστικού χώρου αναφοράς. Κατόπιν, το μοντέλο εφαρμόζεται σε ένα ηλικιωμένο πληθυσμό και οι αποκλίσεις των ηλικιωμένων ατόμων από τον κανονιστικό χώρο αντικατοπτρίζουν τις εγκεφαλικές αλλαγές που οφείλονται σε σωρευτικές επιδράσεις γενετικών, περιβαλλοντικών και παθολογικών παραγόντων. Ο L-map, ο οποίος αποτυπώνει τις διαχρονικές αλλαγές στον εγκέφαλο που σχετίζονται με τη γήρανση, προκύπτει μέσω ενός στατιστικού μοντέλου που υπολογίζει τον ρυθμό αλλαγής των μελετώμενων χαρακτηριστικών, όπως αναλύεται λεπτομερώς στη συνέχεια. Τα διαχρονικά δεδομένα αντικατοπτρίζουν τις εγκεφαλικές αλλαγές που πιθανώς σχετίζονται με τις υποκείμενες νευροπαθολογικές διεργασίες που εξελίσσονται σε ατομικό επίπεδο. Με την ενσωμάτωση αυτών των δύο συμπληρωματικών τύπων δεδομένων, η μεθοδολογία CCL-NMF προσφέρει ένα ολοκληρωμένο πλαίσιο για τον χαρακτηρισμό της ετερογένειας της γήρανσης.

Ακολούθως, παρουσιάζεται αναλυτικά το μοντέλο CCL-NMF, όπως απεικονίζεται στην **Εικόνα 6**. Το εν λόγω μοντέλο συνίσταται σε δύο κύριες φάσεις. Στην πρώτη φάση, υπολογίζονται οι συγχρονικές αποκλίσεις και οι διαχρονικές αλλαγές. Στη συνέχεια, αυτά τα δύο είδη πληροφοριών αξιοποιούνται προκειμένου να εξαχθούν οι διαστάσεις της γήρανσης.

Κανονιστική μέθοδος μοντελοποίησης για την εξαγωγή του χάρτη συγχρονικής απόκλισης (C-map)

Έστω **S1** ο πληθυσμός αναφοράς και **S2** ο πληθυσμός στόχος. Ο C-map εμπεριέχει τις αποκλίσεις του **S2** από τον κανονιστικό χώρο που διαμορφώνεται από τον **S1** χρησιμοποιώντας έναν αντιπαραθετικό αυτοκωδικοποιητή AA (adversarial autoencoder). Η βασική ιδέα αυτής της μεθόδου είναι ότι επειδή ο AA εκπαιδεύεται αποκλειστικά σε δεδομένα του **S1**, μαθαίνει να κωδικοποιεί και να ανακατασκευάζει με ακρίβεια τα δεδομένα του **S1**, ενώ η ακρίβεια του θα είναι περιορισμένη κατά την ανακατασκευή των δεδομένων του **S2**. Ειδικότερα, το σφάλμα μεταξύ της εισόδου και της ανακατασκευασμένης εκτίμησης/εξόδου αναμένεται να αντικατοπτρίζει την απόκλιση του **S2** από τον **S1**.

Η αρχιτεκτονική του AA περιλαμβάνει έναν κωδικοποιητή (encoder E) με δύο κρυφά επίπεδα νευρώνων, το καθένα με 110 νευρώνες, και έναν λανθάνοντα χώρο (latent space) διάστασης 10 νευρώνων. Ο αποκωδικοποιητής (decoder D) και το διαχωριστικό δίκτυο (discriminator D_z) έχουν ίδια δομή, με δύο κρυφά επίπεδα των 100 νευρώνων το καθένα. Ο λανθάνων χώρος κανονικοποιείται ώστε να ταιριάζει με μια γκαουσιανή κατανομή. Όλα τα κρυφά επίπεδα νευρώνων χρησιμοποιούν τη συνάρτηση leaky Rectified Linear Unit (ReLU) με μη γραμμικότητα, ενώ ο λανθάνων χώρος και το επίπεδο εξόδου (output layer) του αποκωδικοποιητή χρησιμοποιούν μια γραμμική συνάρτηση ενεργοποίησης (activation function).

Η εκπαίδευση του AA έχει δύο φάσεις:

1) Φάση ανακατασκευής: Σε αυτή τη φάση, ελαχιστοποιείται η συνάρτηση κόστους/απώλειας ανακατασκευής (reconstruction loss), εξασφαλίζοντας ότι η έξοδος ταιριάζει όσο το δυνατόν καλύτερα με την είσοδο. Ο κωδικοποιητής απεικονίζει τα δεδομένα (**x**) στον λανθάνων χώρο (**z**) και ο αποκωδικοποιητής τα ανακατασκευάζει. Η συνάρτηση απώλειας ανακατασκευής είναι:

$$\mathrm{L}_{\mathrm{recon}} = \|\mathbf{x} - \mathrm{D}(\mathrm{E}(\mathbf{x}))\|_2^2 \quad (1)$$

2) Φάση κανονικοποίησης: Εδώ πραγματοποιείται η αντιπαραθετική (adversarial) εκπαίδευση, η οποία επιβάλλει στον λανθάνοντα χώρο (**z**) να ταιριάζει με την πρότερη (prior) γκαουσιανή κατανομή (P(**z**)) ώστε το διαχωριστικό δίκτυο να μη μπορεί να διακρίνει τα πραγματικά δείγματα από την prior κατανομή (P(**z**)) από τα ψεύτικα δείγματα που προέρχονται από τον κωδικοποιητή. Η συνάρτηση αντιπαραθετικής απώλειας είναι:

$$\mathrm{L}_{\mathrm{adv}} = \mathbb{E}[\log(\mathrm{D}_{\mathrm{z}}(\mathbf{z}))] + \mathbb{E}[\log(1 - \mathrm{D}_{\mathrm{z}}(\mathrm{E}(\mathbf{x})))] \quad (2)$$

Ο κωδικοποιητής προσπαθεί να ελαχιστοποιήσει αυτή την συνάρτηση απώλειας για να ξεγελάσει το διαχωριστικό δίκτυο.

Για την εκπαίδευση του ΑΑ, χρησιμοποιείται ο βελτιστοποιητής (optimizer) Adam για 1000 εποχές, με εφαρμογή πρώιμης διακοπής (early stopping) μετά από 50 εποχές. Στο πλαίσιο αυτού του βελτιστοποιητή, που βασίζεται στην καθοδική κλίση, εφαρμόζεται προσέγγιση μικροπαρτίδας (minibatch), με μέγεθος παρτίδας 200. Ένας κυκλικός ρυθμός μάθησης ενισχύει την αποτελεσματικότητα της εκπαίδευσης, διευκολύνοντας τη σύγκλιση με λιγότερες εποχές. Ο αρχικός ρυθμός μάθησης είναι 0,0001, με μέγιστο ρυθμό μάθησης 0,005. Ο κύκλος ρυθμού μάθησης ακολουθεί ένα βασικό τριγωνικό σχήμα με συντελεστή μείωσης πλάτους 0,98.

Αρχικά, τα χαρακτηριστικά διορθώνονται ως προς το φύλο και το ICV. Τα μοντέλα για τη γραμμική διόρθωση εκπαιδεύονται στα δεδομένα των αρχικών σαρώσεων του **S1** και εφαρμόζονται στα δεδομένα των αρχικών σαρώσεων του **S1** και των αρχικών και διαχρονικών σαρώσεων του **S2.** Προτού εφαρμοστεί ο ΑΑ, τα χαρακτηριστικά τυποποιούνται σε βαθμολογίες z (z-scores). Ξανά εδώ, τα μοντέλα τυποποίησης εκπαιδεύονται στα δεδομένα των αρχικών σαρώσεων του **S1** και εφαρμόζονται στα δεδομένα των αρχικών σαρώσεων των **S1** και **S2,** αφού ο ΑΑ αφορά μόνο στις συγχρονικές (αρχικές) σαρώσεις. Ο πληθυσμός του **S1** χωρίζεται σε τρία υποσύνολα: $\mathbf{S1_{train}}$, $\mathbf{S1_{val}}$ και $\mathbf{S1_{heldout}}$, με αναλογία διαχωρισμού 65%, 15% και 20% αντίστοιχα. Ο ΑΑ εκπαιδεύεται στο $\mathbf{S1_{train}}$ και επικυρώνεται στο $\mathbf{S1_{val}}$, πριν εφαρμοστεί στο **S2**. Η μέση τετραγωνική απόκλιση (mean squared deviation, $\mathrm{MSD} = \frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}(\mathrm{x_j} - \hat{\mathrm{x}}_\mathrm{j})^2$, όπου R είναι ο αριθμός των χαρακτηριστικών - εδώ των περιοχών του εγκεφάλου -, x είναι η είσοδος και $\hat{\mathrm{x}}$ είναι η ανακατασκευασμένη έξοδος) χρησιμοποιείται για την επιλογή των ατόμων με τις μεγαλύτερες αποκλίσεις. Δεδομένου ότι η παρούσα διατριβή επικεντρώνεται στη διαδικασία της απώλειας όγκου που σχετίζεται με την ατροφία, οι περιοχές του εγκεφάλου με αρνητικές αποκλίσεις (δηλαδή μεγαλύτερη έξοδο από είσοδο) δεν περιλαμβάνονται στον υπολογισμό της MSD. Οι συμμετέχοντες του **S2** με MSD≥75ο εκατοστημόριο του MSD του $\mathbf{S1_{heldout}}$, οι οποίοι χαρακτηρίζονται ως **S3**, επιλέγονται ως εκείνοι με σημαντική νευροπαθολογική απόκλιση από τα κανονιστικά δείγματα και χρησιμοποιούνται στο NMF. Ο C-map, ο οποίος έχει μέγεθος ίσο με τον αριθμό των εγκεφαλικών περιοχών επί τον αριθμό των ατόμων **S3**, περιλαμβάνει τις αποκλίσεις, με τις αρνητικές αποκλίσεις να καταστέλλονται αντικαθιστώντας τες με την τιμή 0.

<u>Στατιστική μέθοδος για την εξαγωγή του χάρτη διαχρονικής αλλαγής (L-map)</u>
Γραμμικό μοντέλο μικτών επιδράσεων χρησιμοποιείται για τον υπολογισμό του χάρτη διαχρονικής αλλαγής των χαρακτηριστικών των **S3** ατόμων που έχουν

πολλαπλές σαρώσεις. Για το άτομο i και το χρονικό σημείο j, κάθε χαρακτηριστικό μπορεί να μοντελοποιηθεί ως εξής:

$$Y_{i,j,t} = \beta_{0_j} + \beta_{1_j} \mathrm{Time}_{i,j,t} + \beta_{2_j} X_{i,j} + \gamma_{0_{i,j}} + \gamma_{1_{i,j}} \mathrm{Time}_{i,j,t} + \varepsilon_{i,j,t} \quad (3)$$

όπου $Y_{i,j}$ είναι η τιμή του χαρακτηριστικού στο j-οστό χρονικό σημείο για το άτομο i και $X_{i,j}$ είναι ο πίνακας των συνοδευτικών μεταβλητών (covariates) για τις σταθερές επιδράσεις (fixed effects) (π.χ. χρονικό σημείο, αρχική ηλικία κ.λπ.). Οι τυχαίες παράμετροι της κλίσης (slope) και της τομής (intercept) που διαφέρουν ανά άτομο ακολουθούν διμεταβλητή κανονική κατανομή.

Ο υπολογισμός του ρυθμού μεταβολής κάθε χαρακτηριστικού κάθε ατόμου περιλαμβάνει δύο συνιστώσες: τη μέση πληθυσμιακή κλίση β_{1_j} από τον όρο των σταθερών επιδράσεων και την τυχαία κλίση $\gamma_{1_{i,j}}$ ειδικά για το άτομο i. Έτσι, ο τελικός ρυθμός μεταβολής δίνεται ως: $\beta_{1_j} + \gamma_{1_{i,j}}$. Ο L-map, ο οποίος έχει μέγεθος ίσο με τον αριθμό των χαρακτηριστικών επί τον αριθμό των ατόμων με διαχρονικές μετρήσεις, αποτυπώνει τους ρυθμούς αλλαγής των χαρακτηριστικών στο χρόνο. Στο πλαίσιο της παρούσας ανάλυσης, η εστίαση είναι στους ρυθμούς εγκεφαλικής ατροφίας, και γι' αυτό επιλέγονται να κρατηθούν μόνο οι αρνητικές τιμές, ενώ οι θετικές τιμές -λίγες σε αριθμό και μικρές σε μέγεθος- μηδενίζονται. Για να διασφαλιστεί η συμβατότητα με τον περιορισμό μη-αρνητικότητας που απαιτείται για τα δεδομένα εισόδου στο NMF, το πρόσημο του L-map αναστρέφεται ώστε να περιλαμβάνει αποκλειστικά μη-αρνητικές τιμές.

<u>Από κοινού (Joint) μη-αρνητική παραγοντοποίηση</u>

Μετά τον υπολογισμό των C-map και L-map, πραγματοποιείται η από κοινού παραγοντοποίηση τους με τη μέθοδο NMF. Για τον πίνακα συγχρονικής απόκλισης X_C μεγέθους DxN_C και τον πίνακα διαχρονικής αλλαγής X_L μεγέθους διάστασης DxN_L, όπου το D είναι ο αριθμός των εγκεφαλικών χαρακτηριστικών και N_C (N_L) ο αριθμός των ατόμων με συγχρονικές αποκλίσεις (διαχρονικές αλλαγές), στόχος είναι η εξαγωγή K διαστάσεων που αποτυπώνουν τα μοτίβα εγκεφαλικής γήρανσης, χρησιμοποιώντας τη μέθοδο NMF. Η προσέγγιση αυτή λειτουργεί υπό την υπόθεση ότι οι δύο τύποι δεδομένων μοιράζονται τις ίδιες διαστάσεις γήρανσης (πίνακας λεξικό W μεγέθους DxK). Ωστόσο, κάθε τύπος δεδομένων έχει ξεχωριστούς συντελεστές: H_C μεγέθους KxN_C για τα συγχρονικά δεδομένα και H_L μεγέθους KxN_L για τα διαχρονικά δεδομένα. Το μοντέλο μπορεί να εκφραστεί ως εξής:

$$X_C \approx WH_C, X_L \approx WH_L, \text{subject to } W > 0, H_C > 0, H_L > 0 \quad (4)$$

Η συνάρτηση απώλειας είναι:

$$L = \alpha \|X_C - WH_C\|_F^2 + \|X_L - WH_L\|_F^2 \quad (5)$$

όπου α είναι ένας συντελεστής στάθμισης (weighting coefficient), ο οποίος καθορίζεται από την συγκεκριμένη εφαρμογή ή τα δεδομένα που χρησιμοποιούνται και εξισορροπεί τη συμβολή των C-map και L-map στη διαδικασία εκμάθησης του λεξικού W.

Πρόκειται για μια αμοιβαία περιορισμένη διπλή παραγοντοποίηση (mutually constrained dual factorization) των συγχρονικών και διαχρονικών δεδομένων, η οποία βελτιστοποιείται με τη χρήση του κανόνα πολλαπλασιαστικής ενημέρωσης (multiplicative update rule).

$$w_{ij} \leftarrow w_{ij} \frac{(\alpha X_C H_C{}^T + X_L H_L{}^T)_{ij}}{(\alpha W X_C H_C{}^T + W X_L H_L{}^T)_{ij}} \quad (6)$$

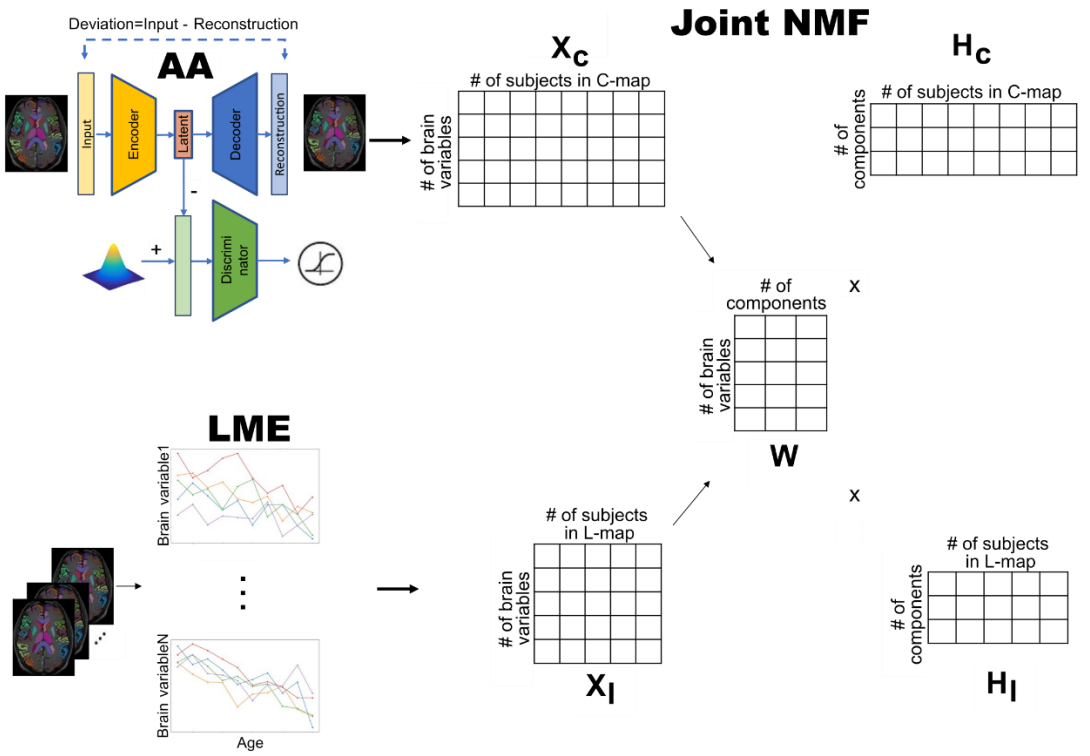
$$h_{ij}^I \leftarrow h_{ij}^I \frac{(W^T X_I)_{ij}}{(W^T W H_I)_{ij}}, (I = C, L) \quad (7)$$

Πριν το NMF, τα (μη μηδενικά στοιχεία των) X_C και X_L αναπροσαρμόζονται με χρήση της MinMax μεθόδου (MinMax scaling) στο εύρος [0,1] για να εξασφαλιστεί το ομοιόμορφο εύρος των χαρακτηριστικών.

Η αρχικοποίηση του W πραγματοποιείται με τη χρήση τυχαίων τιμών, οι οποίες προέρχονται από μια ομοιόμορφη κατανομή στο διάστημα [0, 0,5]. Οι αρχικοί πίνακες κανονικοποιούνται με τη χρήση ενός διαγώνιου πίνακα S, ο οποίος προκύπτει από τις l2-νόρμες (l2-norms) των στηλών του W_{init}:

$$W_{init}{}' = W_{init} S^{-1}, H_{C_{init}}{}' = S H_{C_{init}}, H_{L_{init}}{}' = S H_{L_{init}} \quad (8)$$

Αυτό η διαδικασία κανονικοποίησης επαναλαμβάνεται σε κάθε επανάληψη για να διασφαλιστεί η αριθμητική σταθερότητα και η σύγκλιση του αλγορίθμου.



Εικόνα 6: Διαγραμματική απεικόνιση του μοντέλου CCL-NMF. *Ο αντιπαραθετικός αυτοκωδικοποιητής (AA) υπολογίζει τον χάρτη συγχρονικής απόκλισης (C-map) του πληθυσμού-στόχου S2 (εδώ, ενός ηλικιωμένου πληθυσμού) από τον κανονιστικό χώρο που ορίζεται από τον πληθυσμό αναφοράς S1 (εδώ, έναν υγιή πληθυσμό μέσης ηλικίας). Ο AA εκπαιδεύεται αποκλειστικά στα δεδομένα του S1 και μαθαίνει να τα ανακατασκευάζει με ακρίβεια. Κατά τη διαδικασία ανακατασκευής δεδομένων από τον S2, ο AA παράγει ένα σφάλμα που καταγράφει την ατομική απόκλιση του S2 από τον S1. Οι συμμετέχοντες του S2 που παρουσιάζουν τις μεγαλύτερες αποκλίσεις προσδιορίζονται ως πληθυσμός S3, και χρησιμοποιούνται στο NMF. Τα μοντέλα γραμμικών μικτών επιδράσεων (linear mixed-effects models, LME) χρησιμοποιούνται για τον υπολογισμό του χάρτη διαχρονικής αλλαγής των χαρακτηριστικών (L-map) των S3 ατόμων που έχουν πολλαπλές μετρήσεις στο χρόνο. Με χρήση του NMF, ο C-map, μεγέθους DxN_C, και ο L-map, L_C, μεγέθους DxN_L, όπου D είναι ο αριθμός των αρχικών χαρακτηριστικών και N_C (N_L) είναι ο αριθμός των ατόμων με συγχρονικές αποκλίσεις (διαχρονικές αλλαγές), παραγοντοποιούνται σε ένα κοινό πίνακα λεξικό W, μεγέθους DxK, και ξεχωριστούς πίνακες συντελεστών (H_C, μεγέθους KxN_C, και H_L, μεγέθους KxN_L): $X_C \simeq WH_C$, $X_L \simeq WH_L$, ώστε $W>0$, $H_C>0$, $H_L>0$. K είναι ο αριθμός των εξαγόμενων διαστάσεων γήρανσης. Το 'x' μεταξύ των W και H πινάκων συμβολίζει τον πολλαπλασιασμό πινάκων.*

Δεδομένου του χαρακτήρα του προβλήματος, το οποίο αφορά μη-επιβλεπόμενη μάθηση χωρίς σαφώς καθορισμένη λύση, η επικύρωση της μεθόδου διεξήχθη με τη χρήση ημι-συνθετικών δεδομένων, στα οποία προσομοιώθηκαν συγκεκριμένα μοτίβα ατροφίας. Συγκεκριμένα διερευνήθηκε η ικανότητα του CCL-NMF μοντέλου να ανιχνεύει τα προσομοιωμένα μοτίβα και συγκρίθηκε με αυτή ενός NMF μοντέλου που χρησιμοποιεί σαν είσοδο μόνο τον C-map και ενός NMF μοντέλου που χρησιμοποιεί μόνο τον L-map. Η σύγκριση φανέρωσε ότι το CCL-NMF που λαμβάνει πληροφορία και από τα δύο είδη δεδομένων ανακατασκευάζει με καλύτερο τρόπο τα προσομοιωμένα μοτίβα ατροφίας.

Στη συνέχεια, το CCL-NMF εφαρμόστηκε για τη μελέτη της ετερογένειας της εγκεφαλικής ατροφίας, χρησιμοποιώντας 119 όγκους ROI της φαιάς ουσίας προερχόμενους από T1-εικόνες μαγνητικής τομογραφίας. Τα δεδομένα προήλθαν από την κοινοπραξία iSTAGING, συγκεκριμένα τις παρακάτω μελέτες:

ADNI, AIBL, BIOCARD, BLSA, CARDIA, OASIS, Penn-PMC, SHIP, UK Biobank, WHIMS, WRAP, και HANDLS (Healthy Aging in Neighborhoods of Diversity across the Life Span). Η προεπεξεργασία των T1-εικόνων, η εξαγωγή των ROI όγκων και η εναρμόνιση με χρήση της μεθόδου ComBat-GAM περιεγράφηκε προηγουμένως.

Η ομάδα αναφοράς **S1** (N=977) περιλάμβανε γνωσιακά υγιή άτομα ηλικίας κάτω των 50 ετών, χωρίς παράγοντες καρδιαγγειακού κινδύνου, με μέση ηλικία 39.88±8.09 χρονών και ποσοστό γυναικών 54.86%. Αντίθετα, η ομάδα-στόχος **S2** (N=48.949) αποτελούνταν από άτομα ηλικίας άνω των 50 ετών, με μέση ηλικία 65.41±7.92 χρονών και 53.98% ποσοστό γυναικείου πληθυσμού, εκ των οποίων το 94.23% είναι γνωσιακά υγιείς (**Πίνακας 3**). Χρησιμοποιήθηκαν γραμμικά μοντέλα μικτών επιδράσεων για την εκτίμηση των διαχρονικών ρυθμών μεταβολής των χαρακτηριστικών με την τοποθεσία, την αρχική ηλικία και τον αρχικό όγκο των ROIs ως συνοδευτικές μεταβλητές. Για να μειωθεί η αβεβαιότητα εκτίμησης, τα μοντέλα μικτών επιδράσεων περιλάμβαναν άτομα που είχαν τουλάχιστον τρεις διαχρονικές μετρήσεις.

Πίνακας 3: Δημογραφικά χαρακτηριστικά για τους S1 και S2 πληθυσμούς

	Sample size		Age (years)		Sex (%males)		Diagnosis (%CU)	
	Target	Reference	Target	Reference	Target	Reference	Target	Reference
ADNI	2391	-	73.1±7.2	-	52.4	-	36.4	-
AIBL	922	4	73.1±6.4	45.4±2.5	43.5	25	76	100
BIOCARD	259	-	60.8±8	-	40.9	-	97.7	-
BLSA	916	100	70.1±9.5	40.5±7.4	47.3	42	97.5	100
CARDIA	534	170	53.9±2.3	47.2±2.3	46.4	50.6	1	100
HANDLS	147	33	58.6±5.8	42.9±4.4	44.9	57.6	1	100
OASIS	1097	10	71.8±9.2	47±2	45	20	73.3	100
PENN	959	-	73.9±8	-	43.1	-	20.8	-
SHIP	1810	660	63±7.9	37.6±8.1	48	44.1	100	100
UK BIOBANK	38582	-	64.5±7.3	-	47.1	-	100	-
WHIMS	1080	-	69.6±3.6	-	0	-	100	-
WRAP	252	-	62.1±5.8	-	29	-	99.6	-

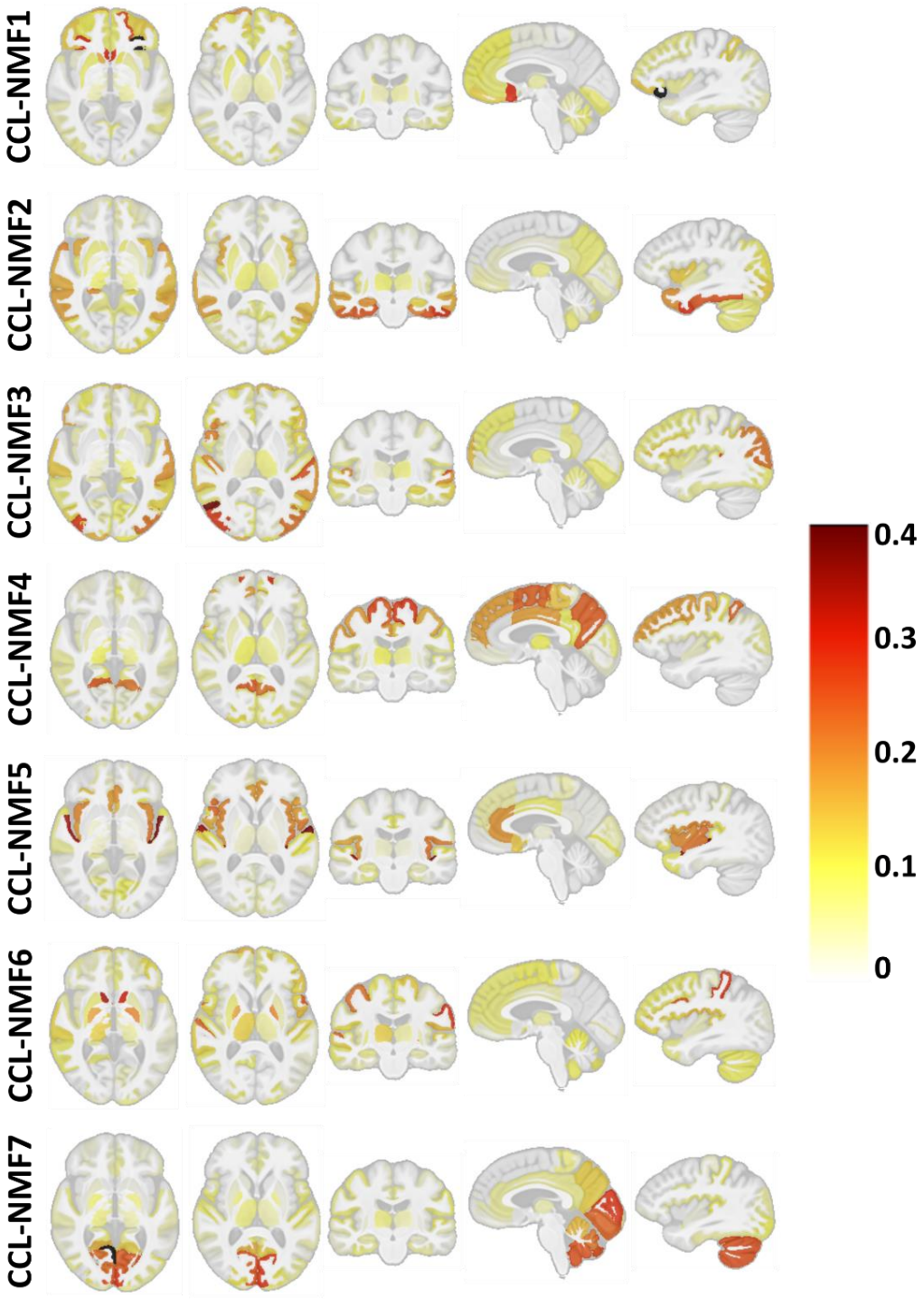
Από τα 48.949 άτομα της ομάδας **S2**, 13.950 (ομάδα **S3**) υπερέβησαν το 75ο εκατοστημόριο της MSD, και 1.063 εξ αυτών είχαν τρεις ή περισσότερες διαχρονικές μετρήσεις, συμβάλλοντας στην κατασκευή του L-map. Η αρχική ηλικία αυτών των συμμετεχόντων ήταν 72,81±8,08 έτη, με μέσο χρόνο παρακολούθησης 4,71±3,79 έτη.

Η ανάλυση CCL-NMF, που πραγματοποιήθηκε με εύρος αριθμού διαστάσεων (K) από 2 έως 15, προσδιόρισε ότι η βέλτιστη λύση επιτυγχάνεται με την επιλογή επτά διαστάσεων (K=7) για την εγκεφαλική γήρανση βάσει του δείκτη αναπαραγωγιμότητας (reproducibility index), της αραιότητας (sparsity) και το σταθμισμένου σφάλματος ανακατασκευής (weighted reconstruction error), όπως ορίζεται από την σχέση (5) .

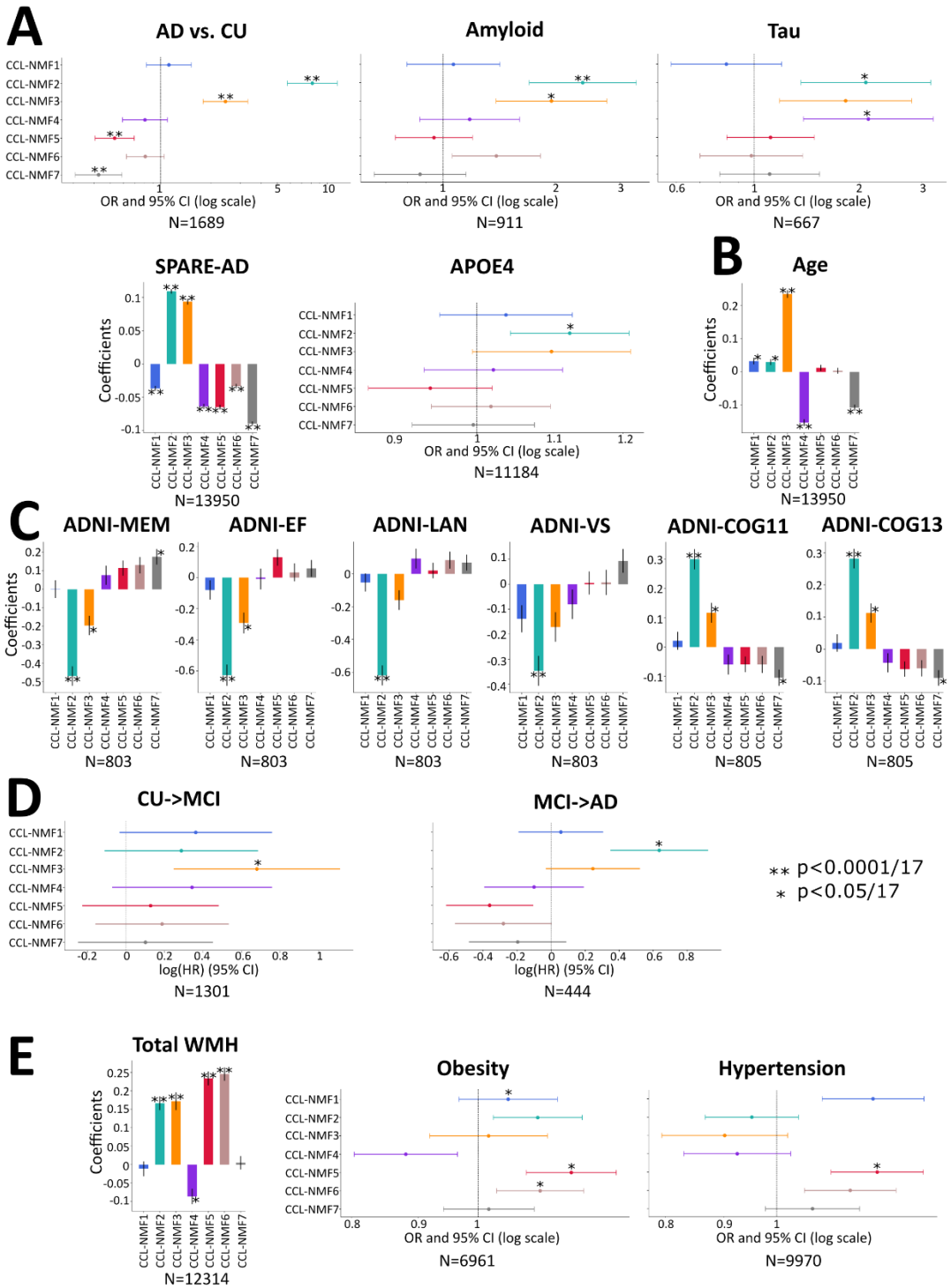
Οι επτά διαστάσεις αποκάλυψαν διακριτά μοτίβα ατροφίας της φαιάς ουσίας (**Εικόνα 7**), τα οποία σχετίστηκαν με κλινικούς, γνωστικούς και παθολογικούς δείκτες (**Εικόνα 8**).

- Το **CCL-NMF1** εμφάνισε ατροφία στα βασικά γάγγλια (basal ganglia), τον μέσω κοιλιακό προμετωπιαίο φλοιό (ventromedial prefrontal cortex) και το μεταιχμιακό σύστημα (limbic system).
- Το **CCL-NMF2** παρουσίασε ατροφία στον μέσο κροταφικό (temporal) λοβό, συμπεριλαμβανομένου του κροταφικού πόλου και της ατρακτοειδούς έλικας (fusiform gyrus), και συνδέθηκε στενά με βιοδείκτες του AD και γνωστική έκπτωση καθώς και με τη μετάβαση από MCI σε AD.
- Το **CCL-NMF3** σχετίστηκε με ατροφία στην κάτω (inferior) μετωπιαία (frontal) και ινιακή (occipital) έλικα, καθώς και σε τμήματα της κροταφικής έλικας. Το CCL-NMF3 εμφάνισε σημαντική συσχέτιση με την αύξηση της ηλικίας, και μέτρια συσχέτιση με την αμυλοειδή β πρωτεΐνη, τη γνωστική έκπτωση και τις βλάβες της λευκής ουσίας. Επιπλέον, το CCL-NMF3 εκφράστηκε περισσότερο στον πληθυσμό σε σχέση με τις άλλες διαστάσεις ατροφίας. Τα παραπάνω χαρακτηριστικά υποδεικνύουν ότι αυτή η διάσταση ενδεχομένως αντανακλά τυπικά χαρακτηριστικά γήρανσης και δε σχετίζεται με συγκεκριμένη παθολογία.
- Το **CCL-NMF4** εμφάνισε ατροφία στην ανώτερη και μέση μετωπιαία έλικα, το προσφηνοειδές λόβιο (precuneus), τον προσαγώγιο φλοιό (cingulate cortex) και ανώτερο βρεγματικό (parietal) λοβό. Το CCL-NMF4 είχε μεγαλύτερη έκφραση σε άτομα μέσης ηλικίας και παρουσίασε συσχέτιση με την παθολογία tau. Ωστόσο, το μικρό μέγεθος του δείγματος tau περιορίζει την εξαγωγή αξιόπιστων συμπερασμάτων.
- Το **CCL-NMF5** παρουσίασε ατροφία στις περισυλβιακές περιοχές και στον πρόσθιο φλοιό του προσαγωγίου (anterior cingulate gyrus). Επιπλέον, αυτή η διάσταση παρουσίασε ισχυρές συσχετίσεις με παράγοντες κινδύνου καρδιαγγειακής νόσου, όπως οι βλάβες της λευκής ουσίας, η παχυσαρκία και η υπέρταση.
- Το **CCL-NMF6** παρουσίασε ατροφία στον επικλινή πυρήνα (nucleus accumbens) και σχετίστηκε σημαντικά με τις βλάβες της λευκής ουσίας και την παχυσαρκία.

- Το **CCL-NMF7** εμφάνισε ατροφία την παρεγκεφαλίδα (cerebellum) και τον μέσω ινιακό λοβό, χωρίς ωστόσο στατιστικά σημαντικές συσχετίσεις με AD βιοδείκτες ή παράγοντες κινδύνου καρδιαγγειακής νόσου. Είναι πιθανό η συγκεκριμένη διάσταση ατροφίας να σχετίζεται με άλλες παθολογικές καταστάσεις για τις οποίες δεν ήταν διαθέσιμοι βιοδείκτες.



Εικόνα 7: Λεξικό CCL-NMF σε μορφή χαρτών εγκεφάλου για K=7 (N_C =13.950, N_L =1.063). *Κόκκινα (λευκά) χρώματα υποδηλώνουν υψηλότερη (χαμηλότερη) συνεισφορά/βάρος της εικονιζόμενης περιοχής στη διάσταση CCL-NMF.*



Εικόνα 8: Συσχετίσεις των συγχρονικών συντελεστών CCL-NMF με Α) χαρακτηριστικά και δείκτες AD, Β) την ηλικία, C) γνωστικές δοκιμασίες, D) την πιθανότητα μετάβασης από γνωσιακά υγιή κατάσταση (cognitively unimpaired, CU) σε ήπια γνωστική διαταραχή (mild cognitive impairment, MCI) και από MCI σε AD, και Ε) παράγοντες κινδύνου καρδιαγγειακής νόσου. *Διόρθωση Bonferroni χρησιμοποιήθηκε για την αντιμετώπιση του προβλήματος των πολλαπλών ελέγχων (multiple testing) (N=17). Ο χρόνος παρακολούθησης για πιθανή μετάβαση CU->MCI και MCI->AD ήταν 5.37±4.29 και 2.6±2.65 έτη, αντίστοιχα. Το N υποδηλώνει το μέγεθος του δείγματος για την αντίστοιχη ανάλυση που παρουσιάζεται στο γράφημα.*

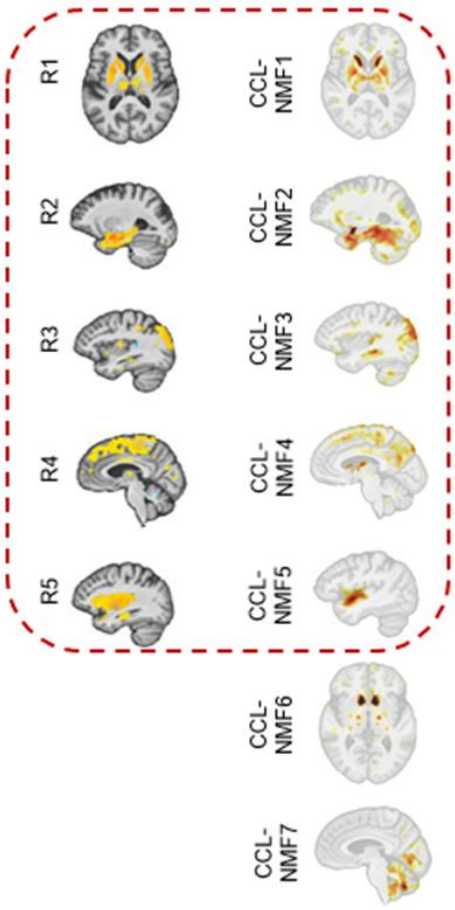
Στη συνέχεια, πραγματοποιήθηκε μια συγκριτική ανάλυση μεταξύ των αναπαραστάσεων εγκεφαλικής ατροφίας που προήλθαν από τη μέθοδο CCL-NMF και από το Surreal-GAN. Αυτές οι δύο προσεγγίσεις εφαρμόστηκαν στο ίδιο σύνολο δεδομένων, το iSTAGING, και στη συνέχεια αξιολογηθηκε η προβλεπτική τους ικανότητα για διάφορα κλινικά χαρακτηριστικά και την πιθανότητα μετάβασης από ήπια γνωστική διαταραχή σε νόσο Αλτσχάιμερ.

Το Surreal-GAN αποτελεί μια ημι-επιβλεπόμενη μέθοδο συσταδοποίησης που βασίζεται σε GAN και έχει αναπτυχθεί ως επέκταση του Smile-GAN. Η βασική του διαφοροποίηση αφορά στην περιγραφή της ετερογένειας μέσω ενός συνόλου συνεχών διαστάσεων, όπου ο δείκτης R (R-index) αποτυπώνει την ένταση/σοβαρότητα κάθε μοτίβου ατροφίας στο κάθε άτομο. Όμοια με το Smile-GAN, το Surreal-GAN χρησιμοποιεί αποκλειστικά συγχρονικά δεδομένα για την μελέτη της ετερογένειας. Παρά τις θεμελιώδεις διαφορές μεταξύ των μεθόδων CCL-NMF και Surreal-GAN, η εφαρμογή τους στα ίδια δεδομένα επιτρέπει μία αξιόπιστη σύγκριση των αποτελεσμάτων.

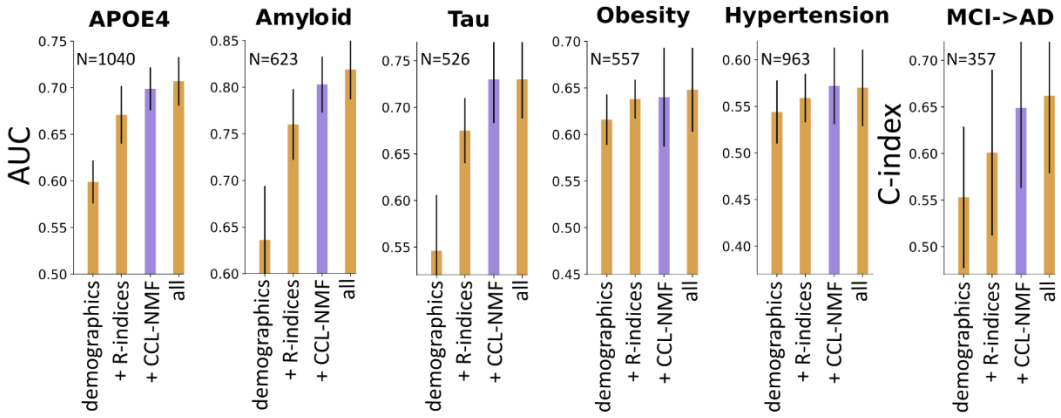
Σύμφωνα με τα ευρήματα, το CCL-NMF εντόπισε επτά διαστάσεις ατροφίας, ενώ το Surreal-GAN πέντε. Στην **Εικόνα 9**, παρατηρείται ότι οι δύο αναπαραστάσεις ευθυγραμμίζονται μερικώς, γεγονός που υπογραμμίζει την σταθερότητα (robustness) των εξαγόμενων μοτίβων. Ωστόσο, το CCL-NMF προσφέρει μια πιο πλούσια και διευρυμένη αναπαράσταση αποτυπώνοντας την ατροφία στον επικλινή πυρήνα και στην παρεγκεφαλίδα και στον μέσο ινιακό λοβό σε διακριτές διαστάσεις.

Συγκρίνοντας τη προβλεπτική ικανότητα των δύο αναπαραστάσεων για διάφορα κλινικά χαρακτηριστικά και την πιθανότητα μετάβασης από MCI σε AD χρησιμοποιώντας μοντέλα παλινδρόμησης και ανάλυσης επιβίωσης και πραγματοποιώντας διαστρωματωμένη διασταυρούμενη επικύρωση 5 υποσυνόλων (5-fold stratified cross-validation), βρέθηκε ότι τα μοντέλα που συμπεριέλαβαν είτε τους δείκτες R (στήλη 2) είτε τους συντελεστές του CCL-NMF (στήλη 3) υπερείχαν των μοντέλων που χρησιμοποίησαν αποκλειστικά δημογραφικά χαρακτηριστικά (στήλη 1) (**Εικόνα 10**). Ειδικότερα, τα μοντέλα που συμπεριέλαβαν τους συντελεστές CCL-NMF παρουσίασαν σταθερά καλύτερες επιδόσεις από εκείνα που χρησιμοποίησαν τους δείκτες R του Surreal-GAN. Αυτό φανερώνει ότι η διαχρονική πληροφορία που ενσωματώνει το CCL-NMF εξάγει μια πιο πλούσια αναπαράσταση που φέρει μεγαλύτερη πληροφορία σε σχέση με τις εξεταζόμενες παθολογίες σε σύγκριση με την αναπαράσταση του Surreal-GAN, η οποία εξάγεται αποκλειστικά από συγχρονικά δεδομένα. Στο παρόν σημείο, είναι σημαντικό να τονιστεί ότι η περιορισμένη διαθεσιμότητα διαχρονικών δεδομένων στην συγκεκριμένη

εφαρμογή, με αναλογία περίπου 1:13 σε σύγκριση με τα συγχρονικά δεδομένα, δεν αναδεικνύει την καθοριστική σημασία της ενσωμάτωσής τους για την εξαγωγή των διαστάσεων γήρανσης. Σε μελλοντικές εφαρμογές που θα περιλαμβάνουν περισσότερες διαχρονικές παρατηρήσεις, αναμένεται ότι η επίδοση του CCL-NMF θα βελτιωθεί σημαντικά, καθώς θα είναι σε θέση να εξάγει αναπαραστάσεις που θα περιλαμβάνουν μεγαλύτερη πληροφορία, ξεπερνώντας κατά πολύ το Surreal-GAN.



Εικόνα 9: Μοτίβα ατροφίας για τις CCL- NMF και τις Surreal-GAN διαστάσεις που βρέθηκαν μέσω t-tests που πραγματοποιήθηκαν σε χάρτες ογκοστοιχείων για κάθε διάσταση CCL-NMF (Surreal-GAN), ενώ προσαρμόστηκαν για την ηλικία, το φύλο, το ICV και τις υπόλοιπες CCL-NMF (Surreal-GAN) διαστάσεις. *Εφαρμόστηκε διόρθωση του ποσοστού ψευδών ανακαλύψεων (False Discovery Rate, FDR) για πολλαπλές συγκρίσεις με κατώφλι τιμής σημαντικότητας 0,001. Η αυξανόμενη ερυθρότητα του ογκοστοιχείου υποδεικνύει ισχυρότερη συσχέτιση με τη συγκεκριμένη διάσταση. Οι πέντε πρώτες διαστάσεις του CCL-NMF παρουσιάζουν ομοιότητες με τις πέντε διαστάσεις του Surreal-GAN.*

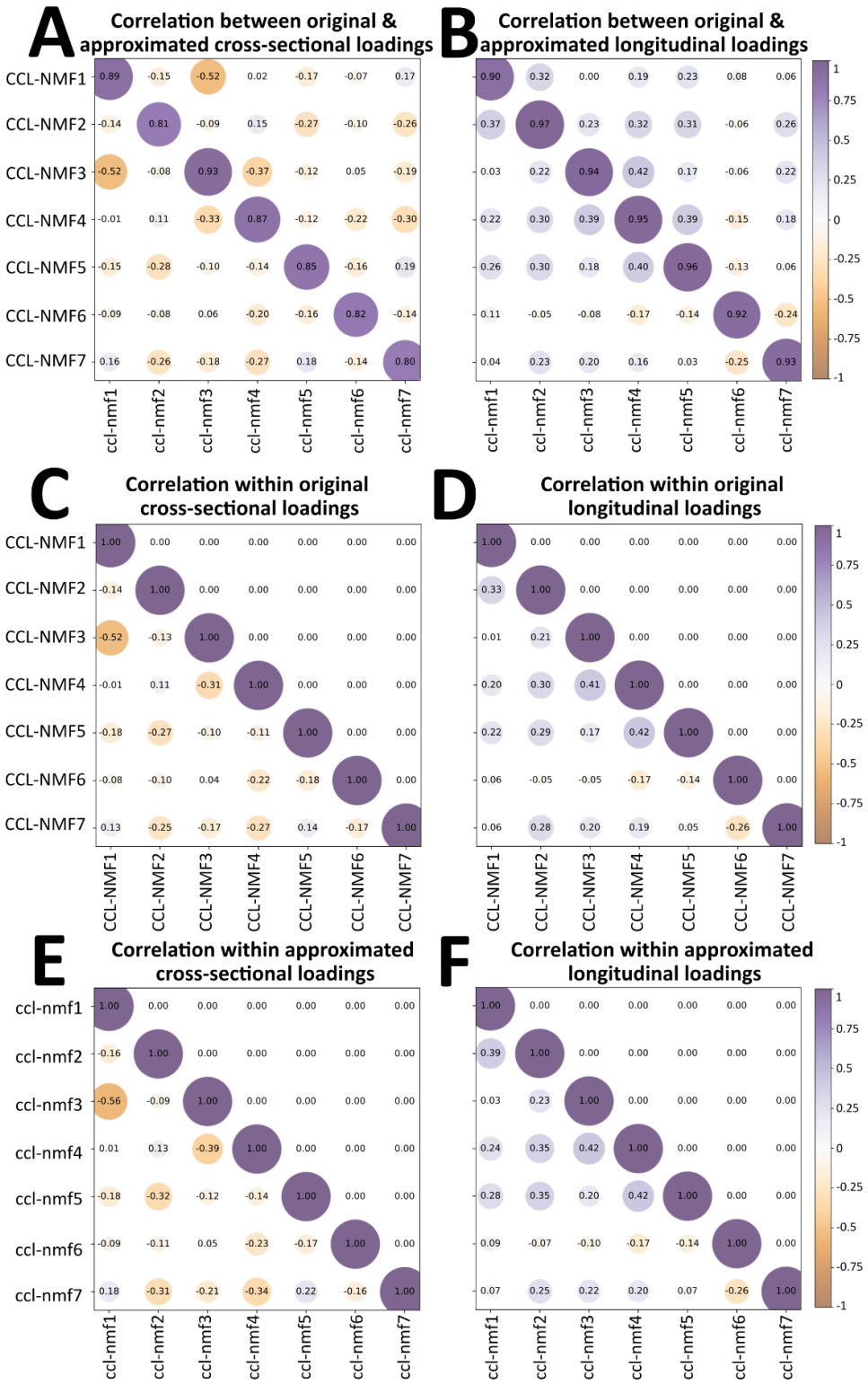


Εικόνα 10: Περιοχή κάτω από την καμπύλη (area under the curve, AUC) και δείκτης συμφωνίας (concordance index, C-index) για μοντέλα που προβλέπουν διάφορα κλινικά χαρακτηριστικά και την πιθανότητα μετάβασης από MCI σε AD. *Τα μοντέλα χρησιμοποιούν διαφορετικά χαρακτηριστικά για τη πρόβλεψη. Τα χαρακτηριστικά αυτά είναι 1) μόνο δημογραφικά χαρακτηριστικά, 2) δημογραφικά χαρακτηριστικά και δείκτες R (R-indices), 3) δημογραφικά χαρακτηριστικά και συντελεστές CCL-NMF και 4) όλα τα προηγούμενα. Το N υποδεικνύει το μέγεθος του δείγματος για την αντίστοιχη ανάλυση που παρουσιάζεται στο γράφημα.*

Η ανάπτυξη ενός μοντέλου, το οποίο μπορεί να αναπαραστήσει με αξιοπιστία την ετερογένεια της εγκεφαλικής γήρανσης, συνιστά έναν σημαντικό ερευνητικό στόχο. Εξίσου σημαντική είναι η δυνατότητα εύκολης υιοθέτησής του και εφαρμογής του σε νέα σύνολα δεδομένων και ποικιλία ερευνητικών περιβαλλόντων. Παρόλο που το μοντέλο δεν είναι υπολογιστικά απαιτητικό, περιλαμβάνει 2 στάδια και απαιτεί έναν πληθυσμό αναφοράς, γεγονός που μπορεί να δυσχεράνει τη χρήση του από νέους χρήστες. Για να διευκολυνθεί η διαδικασία, αναπτύχθηκαν μοντέλα παλινδρόμησης για την εκτίμηση των συντελεστών των επτά διαστάσεων, βασισμένα σε όγκους ROIs και δημογραφικά χαρακτηριστικά. Αυτή η προσέγγιση επιτρέπει τον υπολογισμό προσεγγιστικών συντελεστών χωρίς την ανάγκη εκ νέου εφαρμογής του μοντέλου σε νέα δεδομένα.

Οι Spearman συσχετίσεις μεταξύ των αρχικών και προσεγγιστικών συντελεστών προερχόμενων από μοντέλα παλινδρόμησης με χρήση διαστρωματωμένης διασταυρούμενης επικύρωσης 5 υποσυνόλων κυμάνθηκαν από 0,8 έως 0,93 για τους συγχρονικούς και από 0,9 έως 0,97 για τους διαχρονικούς συντελεστές (**Εικόνα 11A-B**). Αυτά τα ευρήματα υποδεικνύουν ότι οι προσεγγιστικοί συντελεστές συμφωνούν στενά με τους αρχικούς, διατηρώντας παράλληλα τις δομές συνδιακύμανσης (**Εικόνα 11C-F**). Η δυνατότητα αυτή διευκολύνει τους χρήστες στην εκτίμηση των συντελεστών απευθείας από τα δικά τους σύνολα δεδομένων, χωρίς να απαιτείται τεχνογνωσία για την υλοποίηση μοντέλου από την αρχή. Επιπλέον, αξίζει να σημειωθεί ότι τα μοντέλα παλινδρόμησης αυτής της μελέτης χρησιμοποίησαν δεδομένα που δεν έχουν εναρμονιστεί (raw ROI volumes), εισάγοντας έτσι μια σημαντική διευκόλυνση για τους νέους χρήστες, οι οποίοι απαλλάσσονται από την ανάγκη εναρμόνισης των δεδομένων τους—

μια διαδικασία που αποτελεί πρόκληση σε μελέτες που ασχολούνται με ποικιλία συνόλων δεδομένων—πριν από την εφαρμογή των παλινδρομήσεων.



Εικόνα 11: Spearman συσχετίσεις μεταξύ Α) των πραγματικών και των προσεγγιστικών συγχρονικών συντελεστών, Β) των πραγματικών και των προσεγγιστικών διαχρονικών συντελεστών, C) των πραγματικών συγχρονικών συντελεστών, D) των πραγματικών διαχρονικών συντελεστών, Ε) των προσεγγιστικών συγχρονικών συντελεστών, και F) των προσεγγιστικών διαχρονικών συντελεστών CCL-NMF. *Τα κεφαλαία γράμματα (πεζά) αντιπροσωπεύουν τους πραγματικούς (προσεγγιστικούς) συντελεστές.*

Συνολικά, το μοντέλο CCL-NMF γεφυρώνει το χάσμα μεταξύ της πολυπλοκότητας του μοντέλου μηχανικής μάθησης και της εύκολης προς τον χρήστη εφαρμογής. Με αυτόν τον τρόπο, διευκολύνεται η ευρύτερη υιοθέτηση και εφαρμογή των ευρημάτων σε διάφορα ερευνητικά περιβάλλοντα και από χρήστες με διαφορετικά γνωστικά υπόβαθρα.

Συμπερασματικά, το προτεινόμενο μοντέλο εισάγει ένα ευέλικτο πλαίσιο δύο φάσεων για τον εντοπισμό των ετερογενών εγκεφαλικών αλλαγών με τη γήρανση, αξιοποιώντας τόσο συγχρονικά όσο και διαχρονικά δεδομένα. Αυτή η προσέγγιση διαφοροποιείται από τα παραδοσιακά μοντέλα που βασίζονται αποκλειστικά σε συγχρονικά δεδομένα, προσφέροντας νέα προοπτική στην κατανόηση της συνθετότητας της εγκεφαλικής γήρανσης. Επιπλέον, το μοντέλο αποφεύγει την αυστηρή κατηγοριοποίηση των ατόμων σε υποομάδες, όπως παρατηρείται σε προσεγγίσεις όπως το Smile-GAN και τις συνηθισμένες μεθόδους συσταδοποίησης. Αντ' αυτού, επιτρέπει τη συν-έκφραση διαφορετικών διαστάσεων εγκεφαλικής γήρανσης στο ίδιο άτομο. Το μοντέλο δεν απαιτεί υψηλή υπολογιστική και πόρους εξασφαλίζοντας ευρεία προσβασιμότητα στη χρήση του, ενώ η ευελιξία του σχετικά με τον αριθμό των διαστάσεων επιτρέπει την διερεύνηση αναπαραστάσεων ποικίλης πολυπλοκότητας. Παρόλο που στην παρούσα μελέτη το μοντέλο εστιάζει στην ετερογένεια της εγκεφαλικής ατροφίας που σχετίζεται με τη γήρανση, η γενική του δομή καθιστά δυνατή την εφαρμογή του σε διάφορες εγκεφαλικές διαταραχές οι οποίες παρουσιάζουν μονότονη (monotonic) εξέλιξη στο χρόνο. Επιπλέον, η χρήση της ευρέως γνωστής στην ερευνητική κοινότητα NMF μεθόδου διευκολύνει τον πειραματισμό με την ενσωμάτωση όρων κανονικοποίησης στο μοντέλο, οι οποίοι προσαρμόζονται σύμφωνα με τα χαρακτηριστικά των εγκεφαλικών διαταραχών που μελετώνται και παρέχουν αυστηρότερο καθορισμό των διαστάσεων ετερογένειας.

Η εφαρμογή του μοντέλου σε ένα εκτενές και ποικιλόμορφο σύνολο δεδομένων από το iSTAGING αποκάλυψε επτά διακριτές, αναπαραγώγιμες (reproducible) και σημαντικού κλινικού ενδιαφέροντος διαστάσεις εγκεφαλικής ατροφίας. Οι συντελεστές έκφρασης αυτών των διαστάσεων δύνανται να συμβάλουν στην δημιουργία ενός πιο εξατομικευμένου προφίλ του ασθενούς. Οι συγκρίσεις με ένα προηγμένο μοντέλο βαθιάς μάθησης, το οποίο εφαρμόστηκε στο ίδιο σύνολο δεδομένων, ανέδειξαν τη βελτιωμένη προβλεπτική ικανότητα του CCL-NMF όσον αφορά βιοδείκτες και κλινικά χαρακτηριστικά που σχετίζονται με AD και καρδιαγγειακή νόσο. Επιπρόσθετα, η παρούσα μελέτη παρέχει μια πρακτική προσέγγιση για τους ερευνητές, επιτρέποντάς την εκτίμηση των συντελεστών αυτών των επτά διαστάσεων ατροφίας στα διαθέσιμα σύνολα δεδομένων τους, μέσω απλουστευμένων και εύκολα εφαρμόσιμων μοντέλων.

Ωστόσο, η προϋπόθεση της μη αρνητικότητας που επιβάλλει η μέθοδος NMF περιορίζει την εφαρμογή του μοντέλου σε διαταραχές που εξελίσσονται μονότονα. Αυτή η προϋπόθεση μπορεί να περιορίσει την ευρεία χρήση του μοντέλου ή ενδέχεται να απαιτήσει κατάλληλες προσαρμογές σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν φυσικά μικτά πρόσημα ή περιέχουν θόρυβο.

3.Μελλοντικά βήματα

Στο μέλλον, η μεθοδολογία CCL-NMF αναμένεται να τροποποιηθεί κατάλληλα ώστε να εφαρμοστεί σε δεδομένα μορφής ογκοστοιχείων (voxels), επιτρέποντας την ανίχνευση χωρικά λεπτομερών μοτίβων που συχνά αποκρύπτονται όταν χρησιμοποιούνται δεδομένα μορφής προκαθορισμένων ανατομικών περιοχών (ROIs). Επιπλέον, το μοντέλο θα επεκταθεί μέσω της ενσωμάτωσης όρων κανονικοποίησης, με σκοπό την προώθηση της αραιότητας των μοτίβων καθώς και της χωρικής γειτνίασης (spatial contiguity) των ογκοστοιχείων στα εξαγόμενα μοτίβα.

Ένα ακόμα βήμα θα περιλαμβάνει την ενσωμάτωση γενετικών δεδομένων με τη χρήση ατλάντων σε μορφή ογκοστοιχείων που φέρουν γενετική πληροφορία, όπως οι Allen Brain Atlases, για τη διερεύνηση αλληλεπιδράσεων μεταξύ γονιδίων και ανατομίας του εγκεφάλου που σχετίζονται με δομικές και λειτουργικές αλλαγές κατά τη γήρανση και τον νευροεκφυλισμό. Η χρήση τέτοιων ατλάντων αναμένεται να συμβάλει στην αποκάλυψη γενετικών παραγόντων που επηρεάζουν την ανθεκτικότητα και την ευπάθεια στη γνωστική παρακμή. Επιπλέον, το μοντέλο θα εφαρμοστεί σε δεδομένα άλλων απεικονιστικών μεθόδων για τη μελέτη της ετερογένειας των χαρακτηριστικών της διάχυσης του νερού στις νευρικές ίνες της λευκής ουσίας, της λειτουργικής συνδεσιμότητας και της εναπόθεσης αμυλοειδούς και tau πρωτεΐνης. Η μελέτη των διαφόρων πτυχών της εγκεφαλικής γήρανσης αναμένεται να προσφέρει μια πιο ολοκληρωμένη εικόνα, καταγράφοντας συγχρόνως δομικές, λειτουργικές και μοριακές αλλαγές.

Τέλος, θα εξετασθούν οι συσχετίσεις των εξαγόμενων μοτίβων με βιοδείκτες πέραν του AD, συμπεριλαμβανομένων άλλων τύπων άνοιας, της νόσου του Πάρκινσον (Parkinson’s disease) και της πολλαπλής σκλήρυνσης (multiple sclerosis). Αυτή η προσέγγιση αναμένεται να συμβάλει στον εντοπισμό κοινών μηχανισμών μεταξύ των διαφόρων νευροεκφυλιστικών διαταραχών, αποκαλύπτοντας υποομάδες ασθενών που διατρέχουν κίνδυνο για μικτές παθολογίες, και να ενισχύσει την κατανόηση του τρόπου με τον οποίο οι νευροεκφυλιστικές διεργασίες αλληλεπιδρούν, επηρεάζοντας τη γήρανση του εγκεφάλου και τη γνωστική έκπτωση.

Table of contents

List of figures

List of tables

Preface

The present thesis is the culmination of my long-standing interest in the complexities of brain aging and its implications for cognitive decline and neurodegenerative diseases, particularly Alzheimer's disease (AD). Brain aging is shaped by various neurobiological processes, environmental factors, and genetic predispositions, presenting significant challenges in unraveling its heterogeneity. This work has aimed to advance understanding of the intricate and heterogeneous brain changes occurring with aging using state-of-the-art machine learning (ML) techniques and large-scale datasets.

Throughout this research, I have worked at the intersection of cutting-edge ML and large-scale neuroimaging datasets. Leveraging deep learning and big data analytics, this thesis first targets an area that has received limited attention: the heterogeneity of brain aging in its early stages before the onset of clinical symptoms. While most research has focused on disease, such as Alzheimer's, heterogeneity, this study emphasizes understanding variability at preclinical stages. By investigating neuroanatomical patterns early in aging, this research seeks to advance early diagnosis, improve risk stratification, and support more personalized interventions. The analysis reveals several subtypes of early neuroanatomical brain changes, highlighting the diversity in brain aging and shedding light on the role of cardiovascular and lifestyle factors in accelerating neurodegeneration.

Additionally, this work advances current knowledge by presenting a novel model to dissect the heterogeneity of brain changes associated with aging and disease. Utilizing a methodology grounded in non-negative matrix factorization (NMF), the approach circumvents rigid subtype classifications, acknowledging that individuals can simultaneously exhibit multiple aging patterns with varying degrees of severity. This provides a more nuanced representation of the aging brain. The key innovation of this approach lies in integrating both cross-sectional and longitudinal data, addressing the limitations of prior methods focused solely on cross-sectional data, and enabling a more comprehensive understanding of the interplay between static and dynamic brain changes across individuals.

The thesis is organized as follows:

- *Chapter 1* introduces brain aging, focusing on cognitive decline, the main neuropsychological symptom of brain aging. Alzheimer's disease, along with its prodromal stage, mild cognitive impairment, is also discussed, as it is the most common neurodegenerative disease associated with cognitive deterioration. Finally, it delves into the factors driving cognitive decline and the onset of AD.

- *Chapter 2* provides an overview of semi-supervised clustering approaches to explore disease heterogeneity via patterns or transformations between a reference and a target domain. Following the discussion on clustering methods, the chapter introduces normative modeling approaches, which establish individualized baselines to assess deviations from typical patterns. Together, these methods offer complementary perspectives for studying disease heterogeneity, with semi-supervised clustering identifying subgroups within the population and normative models highlighting deviations at the individual level.

- *Chapter 3* investigates the heterogeneity of neuroanatomical brain changes during the early asymptomatic phases of aging by utilizing advancements in deep learning and big data analytics. Identifying neuroanatomical heterogeneity before clinical symptoms arise could offer prognostic insights into neurodegenerative disease susceptibility and improve patient management and clinical trial recruitment. To achieve this, a novel semi-supervised clustering method called Smile-GAN is applied to T1- and T2-weighted magnetic resonance imaging data from a large, harmonized multi-cohort sample of middle-to-late age cognitively unimpaired individuals (N=27,402), coordinated by the iSTAGING consortium. The neuroanatomical heterogeneity is studied separately in four-decade-long age intervals spanning 45-85 years. Finally, the identified subgroups correlate with genetic and lifestyle risk factors, biomedical measures, and cognitive decline trajectories. The study identifies distinct early subgroups of neuroanatomical change that exhibit consistent patterns across age decades, reflecting varying expressions of brain resilience and degeneration. These findings provide critical insights into how early neuroanatomical differences may influence susceptibility to and progression of neuropathological conditions.

- *Chapter 4* presents a novel methodology to address the limitations of existing approaches in disentangling the heterogeneity of brain changes related to aging and neurodegenerative diseases. This method utilizes NMF to decompose brain changes into distinct components, integrating both cross-sectional and longitudinal information through a mutually constrained NMF decomposition framework. By optimizing the reconstruction of both maps of brain change, the joint NMF approach captures the complex interplay between static and dynamic aspects of brain alterations. This methodology is validated with semi-synthetic data before being used to delineate atrophy heterogeneity in an aging population (N=48,949), having as reference a healthy middle-aged cohort (N=977), both drawn from the iSTAGING consortium. The analysis identifies brain atrophy components correlated with AD biomarkers, cognitive performance, cardiovascular risk factors, and disease progression, revealing significant neuroanatomical patterns linked to clinical phenotypes and expanding current knowledge about brain aging heterogeneity. Additionally, by

providing individualized expression levels across components, the approach contributes to establishing personalized therapeutic interventions tailored to patient profiles, paving the way for more targeted and effective treatment strategies. Finally, the model enables easy out-of-sample application by employing regression-based estimation of the expression levels of the derived components, enhancing its applicability in research and clinical settings.

- *Chapter 5* summarizes the conclusions and the contributions of the thesis and reflects on the next steps and future research directions.

1 Theoretical background

Aging is an intrinsic feature of an organism's life cycle, marked by a gradual physical deterioration and an elevated risk of various diseases, including cancer, cardiovascular and neurological diseases, and death. The aging process unfolds at different paces among different species, with variations observed not only among individuals of the same species but also among different tissues of an individual. At the biological level, aging is characterized by the build-up of molecular and cellular damage, resulting in structural and functional dysfunctions in cells and tissues, such as loss of mitochondrial homeostasis, impaired intracellular communication, senescence, and reduced regenerative capacity. The rate and fate of aging are determined by the interplay between the organism, its genes, and the environment[1].

1.1 Brain aging and neurodegeneration

The brain is highly vulnerable to the effects of aging, resulting in alterations to its structure and function. The most prevalent macrostructural age-related changes include brain atrophy[2][3][4], disruption of white matter (WM) integrity, accumulation of white matter lesions[5][6], and alterations in functional connectivity[7][8].

Specifically, cerebral atrophy occurs due to morphological modifications that decrease dendrite arborization complexity, such as dendritic shortening and loss of dendritic spines. This reduces synaptic density and transmission, contributing significantly to cognitive decline[9]. Extensive research indicates that the human brain volume tends to decrease as individuals age, typically at an approximate rate of 5% per decade beyond the age of 40[10], with the potential for this decline to accelerate further, especially after reaching 70 years of age[11]. Additionally, ventricular enlargement arises due to increased space between folds and a loss of gyrification[11].

Typical alterations in WM with age involve degeneration of oligodendrocytes, myelin breakdown, decreased remyelination, and mild reactive astrocytic gliosis linked to white matter lesions. White matter lesions (often called white matter hyperintensities (WMH) as they appear as hyperintense regions on fluid-level attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) scans) in different anatomic brain regions can contribute to different domain functional deterioration. For instance, frontal lobe white matter lesions are associated with decreased information processing speed, visual-motor skills,

verbal fluency, categorization, and mental sequencing. In contrast, subcortical white matter lesions are primarily linked to depression as age increases[12]. Moreover, WM changes are related to small vessel disease, resulting in ischemia, cell death, and enlargement of perivascular spaces that impede the brain's glymphatic waste product drainage[13].

On the whole, studies employing resting-state functional MRI (rsfMRI) have revealed that older adults exhibit decreased within-network functional connectivity (especially within the default mode network but also other brain networks such as salience, dorsal attention, and sensorimotor networks[14][15][16][17]) and increased between-network functional connectivity (such as visual - somatomotor, visual - cingulo-opercular, as well as between dorsal attention components and components from both motor and salience networks[18][19]) compared to younger adults. Moreover, older adults display reduced network segregation/modularity, suggesting less distinct functional divisions across whole-brain networks and decreased local efficiency (increased path length to neighboring nodes). Expanding previous results to brain connectivity during cognitive tasks, task-activation fMRI studies have captured age-related changes in brain network connectivity during cognitive tasks[20][21].

Importantly, aging is the primary risk factor for most neurodegenerative diseases, such as Parkinson's disease (PD), Alzheimer's disease (AD), and frontotemporal lobar degeneration (FTLD)[22]. Neurodegenerative diseases and their associated cognitive deficits are prevalent among older populations, impacting both their lifespan and quality of life. Various hallmarks of aging are related to the pathogenesis of neurodegenerative diseases (**Figure 1.1**). For example, genomic instability, telomere attrition, altered intercellular communication, epigenetic alterations, mitochondrial dysfunction, abnormal protein synthesis, and cell senescence contribute significantly to aging and increase susceptibility to neurodegenerative diseases. Genetics, environmental, and lifestyle risk factors also affect the possible onset and progression of neurodegenerative diseases[23].

Motivated by cognitive decline as the key symptom of brain aging, this work studies Alzheimer's disease, the most common neurodegenerative disease associated with cognitive deterioration. The next section discusses the spectrum from normal brain aging to Alzheimer's disease.

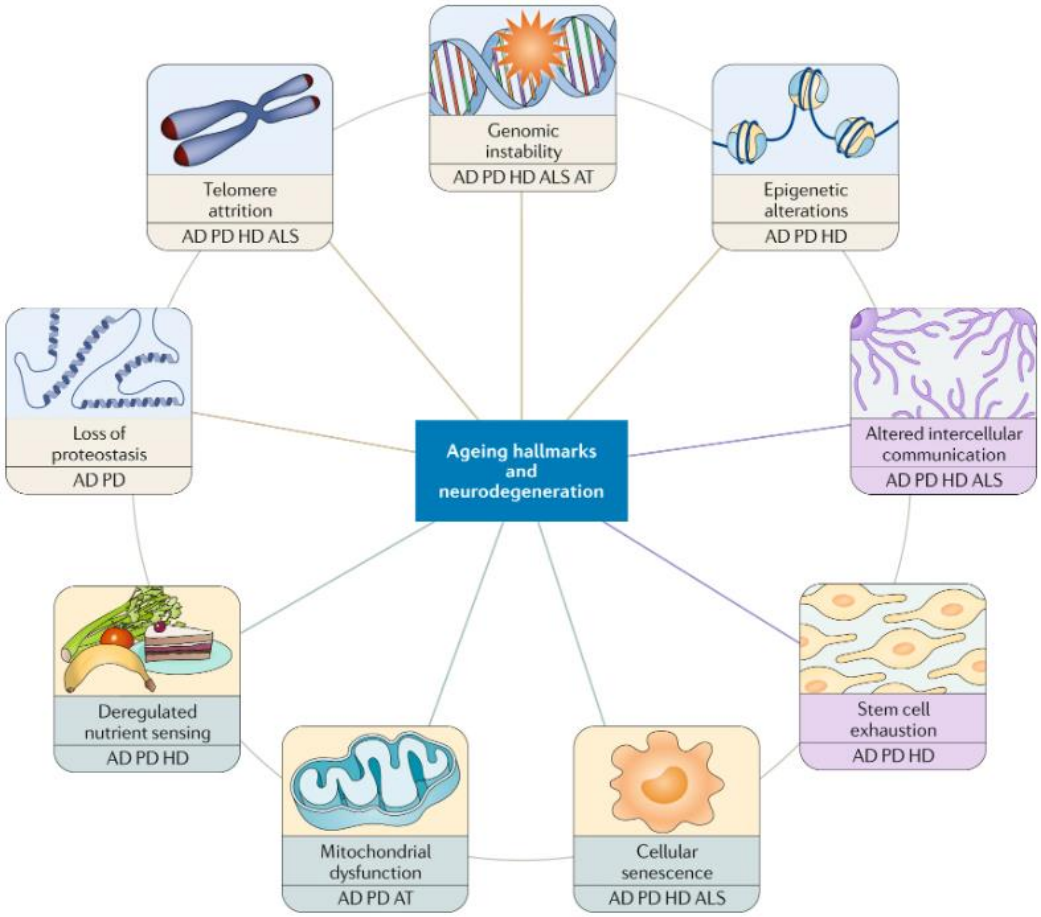


Figure 1.1: Aging hallmarks found in common neurodegenerative diseases. *Abbreviations: AD, Alzheimer's disease; PD, Parkinson's disease; HD, Huntington's disease; ALS, amyotrophic lateral sclerosis; AT, ataxia telangiectasia[23].*

1.2 The continuum from normal brain aging to Alzheimer's disease

1.2.1 Normal brain aging

Despite reaching consensus on the diagnostic criteria for categorizing individuals as Alzheimer's disease patients and some proposed criteria for mild cognitive impairment (MCI), understanding of normal brain aging remains limited. The precise boundary distinguishing normal aging from MCI is a subject of ongoing debate. There are multiple ways to approach the concept of 'normal' in the context of brain aging. One perspective defines it as the absence of any comorbidities. While this approach is valid for studying aging, it may not represent the broader population; disease-free brains are rare, especially in the older population. Individuals without any functional impairment have been referred to as 'super-normal', but this is also very rare. Therefore, the focus primarily rests on what is termed 'typical brain aging. This category encompasses individuals being relatively functional in their everyday tasks; still,

they tend to experience a noticeable decline in performance in several cognitive domains with aging, also reflected by neuroimaging and neuropathological markers changes. These individuals possibly suffer from comorbid conditions such as diabetes, hypertension, and cardiovascular disease (CVD)[24]. To conclude, concepts and terms such as 'normal brain aging' are not well and strictly established and should be cautiously approached.

1.2.2 Mild cognitive impairment

The concept of a grey zone/boundary between cognitive changes observed in normal brain aging and those typically found in Alzheimer's disease has spurred extensive research in the aging and dementia field. Presumably, a continuum exists between normality and the early signs of Alzheimer's disease. This transitional phase is usually referred to as 'mild cognitive impairment'.

Based on current consensus, the diagnosis of MCI requires clinical data indicating a change in cognitive abilities[25][26][27]. This data is typically gathered through interviews with the examined individual or their next of kin. The subjective cognitive complaints are then further substantiated by objective cognitive assessments, such as neuropsychological test batteries. Objective cognitive impairment is identified by below-average performance in one or more cognitive measures, indicating deficits in specific cognitive areas or domains. While there is no gold standard for selecting neuropsychological test batteries, it is essential to ensure that all major cognitive areas, including executive functions, attention, language, memory, and visuospatial skills, are thoroughly evaluated.

The clinical assessment of cognitive complaints in MCI cases often overlaps with the diagnosis of dementia. However, the two conditions contrast in the additional criterion for MCI patients to maintain their independence in functional abilities. To evaluate this aspect, a comprehensive interview is usually conducted with the individual and closest relative, focusing on assessing activities of daily living (ADL) such as breathing, personal cleansing, dressing, toileting, etc., and instrumental activities of daily living (IADL) such as going around on their own, preparing their meals, cleaning their house, etc. When the individual starts experiencing mild difficulties in IADL, this typically indicates MCI. On the other hand, the ability to perform basic ADL is generally supposed to remain intact in individuals with MCI[28].

1.2.3 Alzheimer's disease

Alzheimer's disease is a neurodegenerative disorder marked by the abnormal accumulation and deposition of amyloid β (Aβ) peptides, forming extracellular

plaques, and of hyperphosphorylated tau, which forms intracellular neurofibrillary tangles (NFTs). These changes lead to the loss of synapses and neurons, culminating in a gradual decline in cognitive and functional abilities[29] (**Figure 1.2**). Alzheimer's disease is the leading cause of dementia, accounting for an estimated 60% to 80% of cases[30]. Aging is the most prominent risk factor for AD, with the incidence doubling every five years after the age of 65[31]. Around 44 million people suffered from AD worldwide in 2020, and this number is steadily rising, doubling approximately by 2050[32].

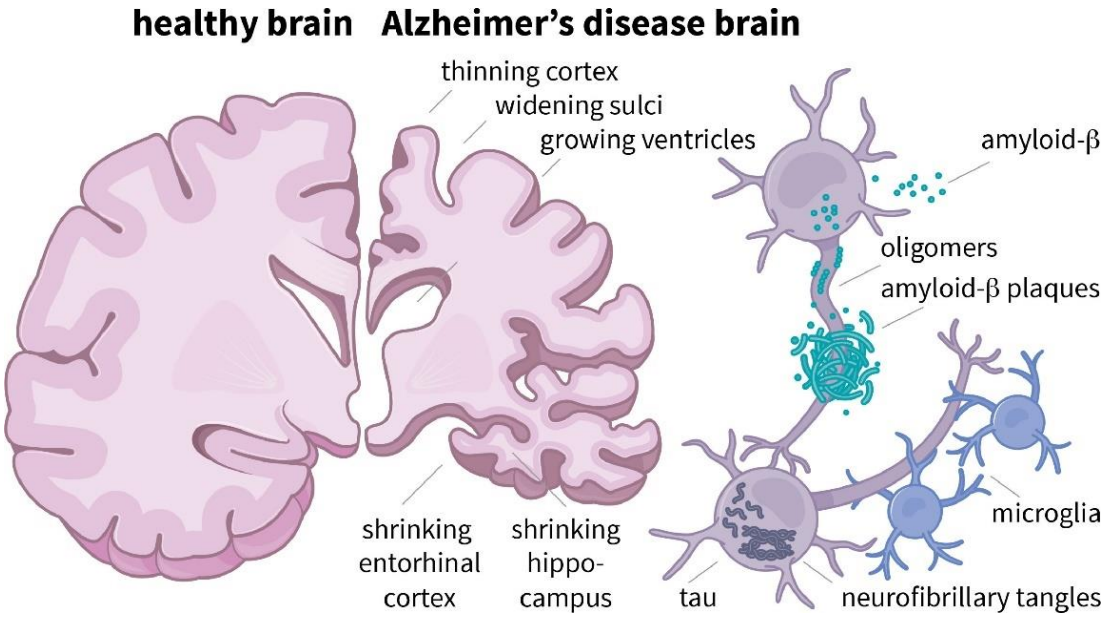


Figure 1.2: Alzheimer's disease neuropathology*[33]*

Alzheimer's disease pathogenesis is influenced by a complex interplay of multiple factors, with genetics playing an essential role. The AD-related genetic variants can be grouped into two categories, as depicted in **Figure 1.3**:
1]Rare, highly penetrant mutations in amyloid β A4 precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) lying on one end of the spectrum are responsible for autosomal dominant familial AD (ADAD), often with early onset. Among these mutations, PSEN1 accounts for approximately 81%, APP for about 14%, and PSEN2 for roughly 6% of ADAD cases [34].
2]Common alleles found through genome-wide association studies (GWAS) in the opposite end of the spectrum are primarily associated with late-onset sporadic AD (SAD). The term 'sporadic', contrary to 'familial', indicates that the patients rarely report the presence of first-degree relatives affected by the disease. While these common alleles have a low individual effect on disease susceptibility, their cumulative impact contributes to the overall genetic predisposition for the disease. Among those genes, the only gene consistently found to be associated with late-onset sporadic AD across multiple genetic

studies is the apolipoprotein E (APOE) gene, specifically the ε4 allele of APOE[35]. The APOE gene provides instructions for making a protein called apolipoprotein E involved in lipid metabolism, which is immunochemically colocalized to amyloid plaques, vascular amyloid deposits, and neurofibrillary tangles in AD[36]. Besides APOE, various independent AD GWAS have yielded evidence for over 80 risk loci[37].

Familial AD constitutes a small proportion of AD cases, while more than 90% of AD patients appear to be sporadic and to have disease onset between the ages of 60 and 65[38]. These sporadic forms of AD are regarded as multifactorial disorders, implying that an individual's susceptibility arises from the intricate interplay among environmental, lifestyle, genetic, and epigenetic factors. The next section will discuss various factors that trigger a cascade of events culminating in cognitive weakness and Alzheimer's disease.

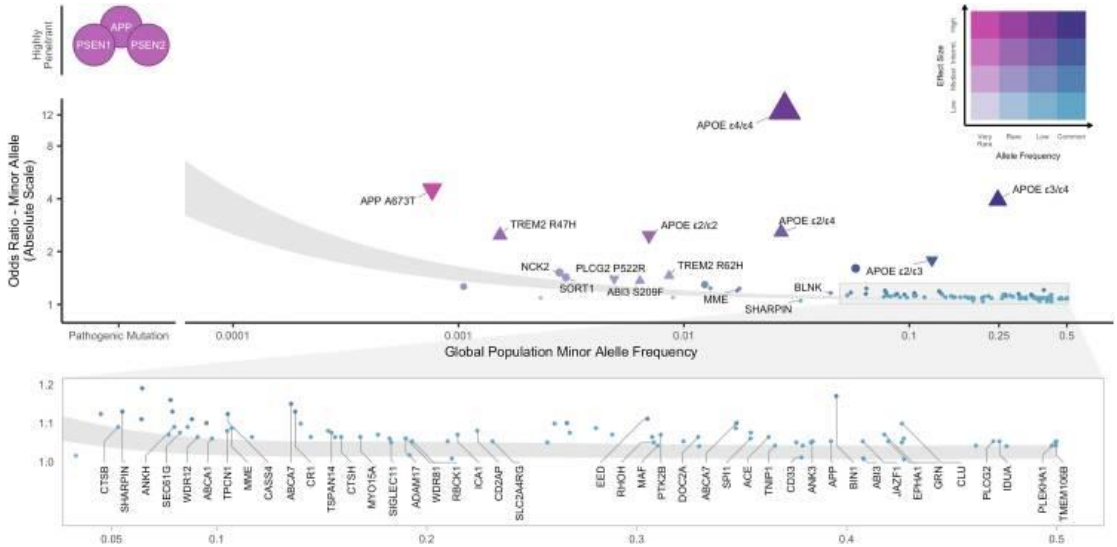


Figure 1.3: Genetic framework of AD. *Rare, highly penetrant mutations in APP, PSEN1, and PSEN2 associated with autosomal dominant familial AD at one end of the spectrum vs common and low-frequency variants with small effect size discovered by GWAS of late-onset sporadic AD at the other end of the spectrum[37].*

1.3 Factors driving cognitive decline and Alzheimer's disease

Over twenty years, the quest for viable treatments capable of decelerating cognitive decline and Alzheimer's disease progression has predominantly relied on the amyloid hypothesis[39][40]. This hypothesis briefly suggests that the amyloid β protein plays a pivotal role in triggering a series of events that ultimately lead to cognitive deterioration and dementia. Clinical trials have focused on techniques to reduce amyloid, encompassing both active and passive immunization and inhibiting enzymes responsible for amyloid

production. However, there has been a lot of discussion regarding the efficacy of such drugs. Several of these trials succeeded in decreasing the accumulation of amyloid β in the brain, but this was not reflected by improved clinical outcomes. There were clinical trials encompassing immunotherapies that had adverse effects because of not targeting the correct amyloid β variants[41]. Such examples raised concerns about whether the amyloid-lowering drugs could really enhance cognition. On the other hand, recent data showcased the potential advantages of amyloid-lowering immunotherapies. A phase 1b study of aducanumab demonstrated pronounced removal of Aβ from the brain and a deceleration in clinical deterioration[42]. Similar results were observed in a phase 2 trial involving donanemab[43].

1.3.1 Amyloid plaques

In individuals without AD, amyloid β is formed by sequential cleavages of amyloid β precursor protein by β- and γ-secretases. Then, it is discharged outside the cell, broken down, or eliminated. Nevertheless, as individuals get older or when pathologies arise, the capacity to metabolize Aβ diminishes, accumulating Aβ peptides. Among the accumulated Aβ, Aβ 40 and Aβ 42 are the major components. The rise of Aβ 42 or Aβ 42 /Aβ 40 ratio promotes the formation of Aβ fibrils. Over time, these accumulated Aβ fibrils aggregate into senile plaques, leading to neurotoxicity and triggering tau pathology, which is eventually followed by neurodegeneration (observed as brain shrinkage, decreased metabolic activity, or elevated levels of neurofilament light in cerebrospinal fluid or blood plasma), culminating in cognitive decline[44].

1.3.2 Tau tangles

Substantial evidence derived from biomarkers has shown that it is not just Aβ but the synergy between Aβ and tau that is linked to impaired brain function, atrophy, and cognitive decline. Cognitively unimpaired (CU) individuals with increased Aβ and tau deposits in the brain experience cognitive decline, especially in episodic memory, faster than those without any biomarker or only one of the abnormal biomarkers[45][46]. One step forward, there are studies indicating early pathological tau changes that occur in the brainstem and further propagate to medial temporal regions before evidence of Aβ accumulation[47]. The tau accumulation in the medial temporal lobe is linked to neurodegeneration and memory loss. This specific pathological manifestation of tau without the presence of Aβ has been categorized as primary age-related tauopathy (PART); it is still uncertain whether PART is an independent tau-related condition, separate from Alzheimer's disease, or if it is an earlier stage within the AD continuum[48]. In any case, this data strengthens the hypothesis that the simultaneous presence of amyloid and tau accumulation triggers

cognitive deterioration and the onset of dementia. Nevertheless, multiple other factors could potentially obfuscate the link between Aβ and cognition.

1.3.3 Vascular disease

Vascular disease constitutes an age-related pathological condition that often plays a pivotal role in cognitive decline. Brain MRI changes in this disease usually manifest as white matter hyperintensities. While WMH is also present in Alzheimer's disease cases, investigations involving cognitively unimpaired individuals or subjects with MCI reveal that lesions and infarctions contribute to cognitive decline independently or synergistically with amyloid and tau[49][50][51][52]. Furthermore, vascular risk factors such as smoking and hypertension contribute to cognitive deterioration independently of both amyloid levels and the presence of white matter hyperintensities[53][54]. This indicates the prominent vascular component of cognitive decline, potentially acting through pathways beyond the scope of MRI resolution or via mechanisms involving inflammation.

1.3.4 Non-Alzheimer's disease-related proteins

Besides vascular disease, several other protein aggregates, such as α-synuclein, associated with dementia with Lewy Bodies (DLB) and multiple system atrophy (MSA), and transactive response DNA binding protein of 43 kDa (TDP-43), associated with amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration, are usually found in the brains of older subjects. These proteins may be related to different types of dementia, but they may co-occur with AD. Regardless of the presence of AD pathology, such non-AD pathologies heighten the vulnerability of the aging brain to cognitive decline and dementia [55].

1.3.5 Neuroinflammation

In post-mortem examination of Alzheimer's disease cases, extracellular Aβ plaques and intracellular neurofibrillary tangles are encircled by activated microglia[56]. Microglia are the brain's resident immune cells, continuously surveilling the cerebral microenvironment to address pathogens and injuries[57]. Although microglial activation is prevalent in numerous neurodegenerative diseases, the exact role of microglia is quite complex and has not been fully explored. Depending on their surroundings, activated microglia can adopt either a protective or neurotoxic phenotype. The protective phenotype involves phagocytosis, which aids in clearing Aβ fibrils and neuronal debris, reshaping synapses, and releasing growth factors. Conversely, the neurotoxic phenotype releases cytokines, which can induce or contribute to tissue damage and disease onset or progression[58]. The link between Aβ and tau accumulation might entail microglia activation. In a study involving cell

cultures, soluble Aβ oligomers were observed to activate microglial cells[59]. Research involving transgenic Alzheimer's mouse models has demonstrated that microglia activation occurs before tau accumulation[60]. Microglia activation induces tau hyperphosphorylation by releasing cytokines, ultimately leading to neurofibrillary tangles[61]. Additionally, there are several genetic variants associated with the immune response that have been identified as risk factors for Alzheimer's disease, indicating the possible initiating role of inflammation in the amyloid cascade [62][63]. For example, TREM2, a gene modulating inflammatory responses and neuroprotection in microglia and regulating myeloid cell maturation, proliferation, and survival, has multiple variants with diverse roles in the development and progression of AD[64].

1.3.6 Lifestyle and environmental factors

Although exploring associations between biomarkers offers convincing support for a chain of pathological events, there remains significant unexplained variability in these relationships. Some individuals experience a sudden and sharp decline in their cognitive abilities; others can maintain their cognitive performance throughout their lifespan. The concept of individual resilience is not new; in the late 1960s, researchers first noted a divergence between the extent of brain pathology and cognitive function before death in several brain samples[65]. Resilience denotes preserving intact cognitive function despite neuropathological changes that typically result in significant clinical impairment. Resistance is when pathological features are notably absent even when they might be anticipated based on group-level data[66][67]. Previous studies suggest that resistance and resilience may be promoted by lifestyle factors such as diet, physical activity, social engagement, cognitive stimulation as well as environmental factors, including avoiding exposure to heavy metals, nanoparticles, and toxic pesticides, socioeconomic factors, genetics, and epigenetics mechanisms[68][69][70][71][72]. For example, sleep could act as a resistance mechanism by aiding in the clearance of amyloid[71], while body mass index, smoking, and alcohol consumption were considered risk factors connected to decreased resistance[73]. On the other hand, intellectual enrichment might primarily serve as a mechanism of resilience, potentially linked to lower Alzheimer's disease pathology[69].

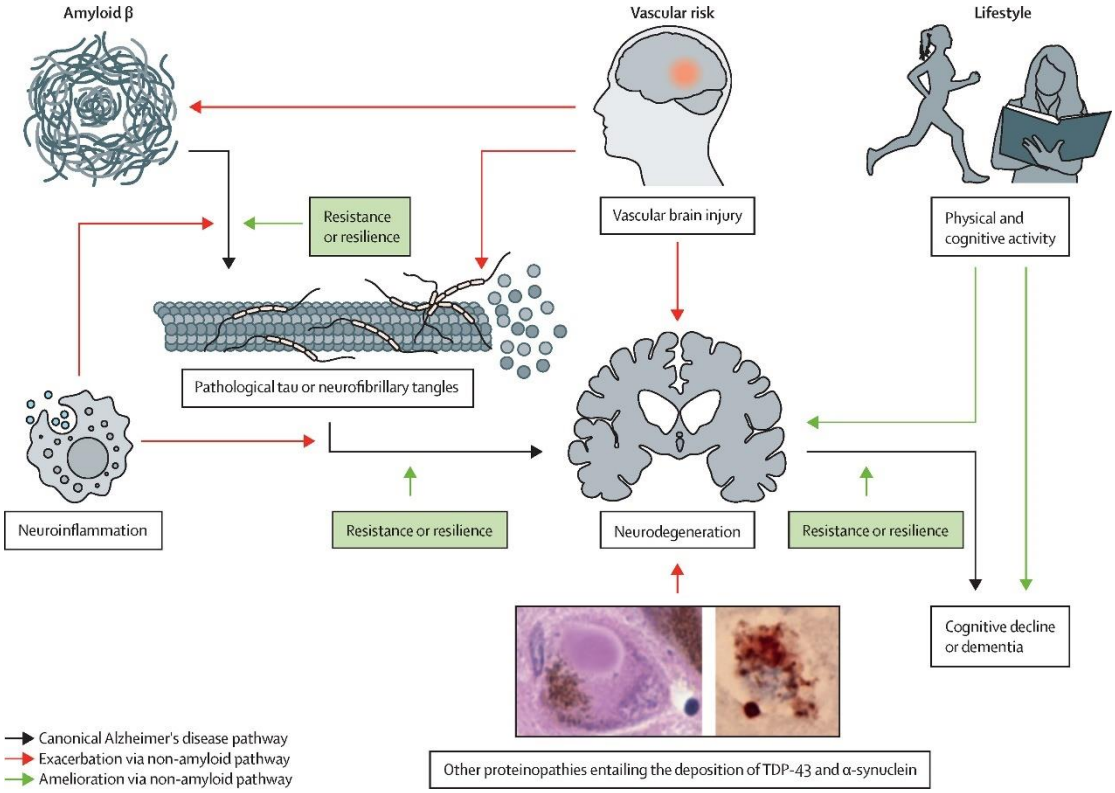


Figure 1.4: The intricacy of Alzheimer's disease pathophysiology. *The canonical amyloid pathway illustrated by black arrows is affected by various factors that either intensify (red arrows) or alleviate (green arrows) the likelihood of dementia progression [74].*

Currently, the amyloid hypothesis has served as the predominant model for understanding the pathophysiology of Alzheimer's disease, drawing support from genetic data, preclinical observations, and human biomarker studies. However, solely relying on this hypothesis might not consistently yield therapeutically meaningful outcomes. The complexities and limitations of this hypothesis raise a methodologically challenging question: within an individual or a population sample, how pivotal is the canonical amyloid cascade compared to other contributing factors such as comorbidities, lifestyle, environmental, and genetic factors? **Figure 1.4** illustrates how protective and risk factors can shape the canonical amyloid cascade. Alzheimer's disease, like other common late-life conditions such as CVD, is complex and multifaceted, and multiple therapeutic strategies might potentially treat it. By better observing and quantifying the various pathways and biomarkers contributing to the onset and progression of Alzheimer's disease, starting from the early asymptomatic stages, deeper insights can be gained into the factors at play in each individual, ultimately facilitating the development of tailored treatments for specific individuals at different stages of the disease.

2 Heterogeneity in brain aging

As discussed previously, Alzheimer's disease is a complex and multifaceted process. Although AD is one of the most common age-related neurodegenerative disorders, and it accounts for the majority of dementia cases, it is only one component of the vast brain aging spectrum. Several neuropathologies, including other dementia types (DLB, FTLD, etc.), tauopathies, and vascular pathologies, along with lifestyle, environmental, and genetic factors, induce intricate brain changes, either intensifying or protecting against neurodegeneration. Characterizing this heterogeneity is crucial for individualized predictions, patient management, and stratification into clinical trials. However, most studies overlook this heterogeneity or use a priori-defined neuropathological categories based on clinical diagnoses. This results in a restricted comprehension of underlying biological mechanisms with potential clinical implications.

Over the past decade, the progressively growing amount of clinical, neuroimaging, and molecular biomarker data collected from large-scale observational cohort studies have enabled a deeper investigation and, therefore, understanding of the various manifestations of aging, Alzheimer's disease, and related dementias[75][76][77]. The availability of such large-scale datasets, along with the development of harmonization methods[78] allowing cross-cohort constructive integration of these data, has fostered the advancement of data-driven clustering techniques for studying disease heterogeneity.

A growing body of literature seeks to dissect this heterogeneity using several machine learning approaches. For example, Zhang et al.[79] utilize a Bayesian Latent Dirichlet Allocation (LDA) model to identify latent atrophy patterns in voxel-wise MRI data. While effective in uncovering patterns, the model requires discretization of continuous voxel data, which can introduce artifacts and reduce sensitivity. Additionally, its reliance on cross-sectional data limits its ability to capture individual atrophy progression over time. Similarly, SuStaIn[80] infers subtypes and stages to address temporal and phenotypic heterogeneity but assumes fixed subtypes and arbitrary timescales, potentially oversimplifying complex disease spectra. Its dependence on cross-sectional data may not accurately reflect true temporal dynamics, and its computational constraints restrict input to a limited number of brain regions, hindering the resolution of finer details. Finally, both methods, along with several other techniques based on unsupervised clustering methods such as k-means[81][82] and hierarchical clustering[83][84][85], parse heterogeneity

directly in the patient domain and thus are often limited by disease-irrelevant confounding variability, e.g., neurodevelopmental variability of brain structure across individuals.

More recently, semi-supervised clustering approaches have emerged to address this issue by examining heterogeneity from a different perspective. Such approaches disentangle disease heterogeneity via patterns or transformations between the reference (e.g., healthy controls) and the target domain (e.g., patients), thus minimizing the influence of disease-unrelated confounders that are common to both groups (**Figure 2.1**).

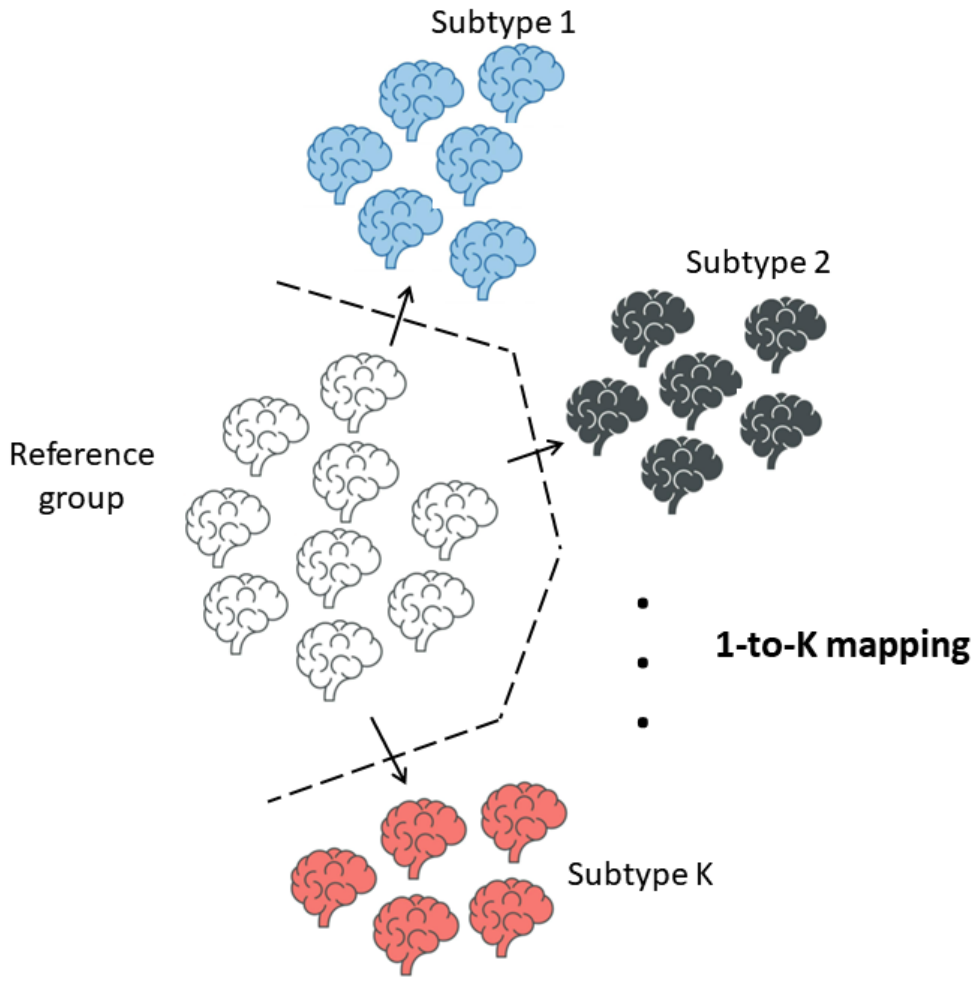


Figure 2.1: Semi-supervised clustering approaches. *The figure is adapted from [86].*

Clustering approaches are useful for identifying subgroups of participants but face several challenges: clustering focuses on group averages and does not fully model individual variation within clusters, treating clusters as atomic units; various partitioning methods can yield different results depending on the measures and algorithms used; some participants may not fit clearly into any class, or some classes may be too small to be meaningful; selecting a unique optimal number of clusters can be problematic, with different metrics

potentially suggesting different solutions; and finally, it is unclear whether healthy participants should be clustered separately or alongside patients, as disease variation may be nested within normal variation.

Normative modeling offers a complementary perspective to the predominant approach for addressing heterogeneity using clustering algorithms. It shifts the focus away from group means to understand cohort variation, emphasizing second-order statistics over first-order statistics. It aims to understand individual variations and map deviations at the individual level independently of clinical labels, offering a distinct and valuable perspective on the heterogeneity domain [87][88][89][90].

Introduced over a century ago, normative growth charts have become essential in pediatric medicine and anthropometry, quantifying individual variation against centiles of a reference population[91][92]. Normative modeling is now a well-established technique for making individual-level inferences in clinical neuroimaging studies[89][93][94]. Normative models can estimate various mappings, such as between behavioral scores and neurobiological measures. They are particularly valuable in studying brain development and aging, given that many brain disorders arise from atypical developmental trajectories[95] and that cognitive decline is linked to brain tissue changes in aging and neurodegenerative diseases[96][97][98]. Normative modeling has been applied in diverse clinical contexts, including charting the development of pre-term infants[99] and exploring the biological heterogeneity in cohorts with brain disorders like schizophrenia[100][101][102], attention-deficit/hyperactivity disorder (ADHD)[103][104][105], and autism[106][107].

In the remainder of this chapter, several semi-supervised approaches will be examined, with a particular focus on the Smile-GAN model, which will be explored in detail in Chapter 3. Normative modeling techniques will also be discussed, emphasizing autoencoders, which will be featured in Chapter 4.

2.1 Semi-supervised clustering approaches

2.1.1 HYDRA (Heterogeneity through Discriminative Analysis)

HYDRA[108] is based on a widely used discriminative method called support vector machines (SVM)[109]. HYDRA combines multiple (K) SVM classifiers piecewise to build a K-facet polytope, where each PT is assigned to the facet/hyperplane that best distinguishes it from the healthy controls (HC). Each facet of the convex polytope can be seen as encoding a distinct subtype, thereby capturing a unique disease effect. This approach serves the dual

purpose of classification and clustering; the classification part involves distinguishing between HC and PT using a convex polytope created by combining several linear max-margin classifiers. The clustering involves grouping PT into clusters based on their associations with the individual linear sub-classifiers.

Estimating this convex polytope involves solving each linear SVM iteratively, adhering to the principles of sample-weighted SVM. The optimization process ends when the sample weights reach stability, indicating the establishment of the optimal polytope.

The overarching objective here is to maximize the margin of the polytope, which can be summarized as:

$$\min_{\{\mathbf{w}_j, b_j\}_{(j=1)}^{K}} \sum_{j=1}^{K} \frac{\|\mathbf{w}_j\|_2^2}{2} + \lambda \sum_{\substack{i|y_i=+1 \\ j}} \frac{1}{K} \max\{0, 1 - \mathbf{w}_j^T \mathbf{x}_i^T - b_j\} + \lambda \sum_{\substack{i|y_i=-1 \\ j}} S_{i,j} \max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i^T + b_j\}$$

where $\mathbf{w}_j$ and b_j are the weight and bias for the j hyperplane, respectively. $S = [S_{i,j}] \epsilon \{0,1\}^{NxK}$ is a binary subtype membership matrix indicating whether the PT sample i (i=1,.., N) belongs to subtype j (j=1,.., K). Finally, λ is a penalty parameter on the training error.

2.1.2 CHIMERA

Similar to HYDRA, CHIMERA[110] seeks a '1-to-K' mapping; however, contrary to HYDRA, which is a discriminative approach, CHIMERA employs a generative probabilistic framework modeling pathological processes through transformations from HC to the PT space/distribution with each transformation T representing a disease subtype. Assuming a set of HC **X** and a set of patients **Y**, **X**′ denotes the transformed HC given as $\mathbf{x}_i' = \mathbf{T}(\mathbf{x}_i)$. Given a disease differentiated in K distinct ways from the HC, the transformation for a single HC point i into the patient domain is specified as follows: $\mathbf{T}(\mathbf{x}_i) = \sum_{k=1}^{K} \lambda_{ki} T_k \mathbf{x}_i$,

In the scenario where the pathological subtypes and, thus, the pathological directions are distinct:

$$\lambda_{ki} = \begin{cases} 1, & \text{for the transformation associated} \\ & \text{with the specific disease subtype affecting } \mathbf{x}_i \\ 0, & \text{otherwise} \end{cases}$$

The matching between **X**′ and **Y** distributions is estimated using a variant of the Coherent Point Drift algorithm[111]. Every HC point transformed into the

PT domain ($\mathbf{x}_i'$) is treated as the centroid of a spherical Gaussian cluster. Actual PT points ($\mathbf{y}_j$) are considered independent and identically distributed data generated by a Gaussian Mixture Model (GMM)[112] with equal weight assigned to each cluster. The aim is to maximize the data likelihood expressing the similarity between the real PT and transformed HC distributions, considering confounders such as age and sex. The Expectation-Maximization algorithm[113] is employed to optimize the loss function. Subsequent clustering inference becomes straightforward once the optimized transformation $\mathrm{T_k}$ is attained, allowing for the assignment of a patient to the subtype membership corresponding to the highest likelihood.

2.1.3 Smile-GAN (SeMI-supervised cLustEring-Generative Adversarial Network)

Smile-GAN[114] is a Generative Adversarial Network (GAN)[115] architecture for clustering a PT group based on its multivariate differences to a HC group. The primary concept lies in learning 1-to-K (number of clusters) mappings from the HC domain X to the PT domain Y. The Smile-GAN architecture is displayed in **Figure 2.2**. Specifically, the Smile-GAN model achieves this idea by learning one mapping function, f, from joint HC domain X and subtype domain Z to the PT domain Y, by transforming HC data **x** to distinct synthesized PT data **y'** = f(**x, z**) that cannot be distinguished from the real PT data by the adversarial network. The data distributions are denoted as $\mathbf{x} \sim \mathrm{p_{HC}}$, $\mathbf{y} \sim \mathrm{p_{PT}}$, $\mathbf{y}' \sim \mathrm{p_f}$, and $\mathbf{z} \sim \mathrm{p_{Sub}}$, respectively, where $\mathbf{z} \sim \mathrm{p_{Sub}}$ is sampled from a discrete uniform distribution and encoded as a one-hot vector with dimension equal to the number of clusters. Under the assumption that a single true function represents the underlying pathology of the real PT variable **y**=h(**x, z**), Smile-GAN enhances the approximation of the mapping function f to closely align with the genuine underlying function by imposing several regularization constraints. These constraints promote sparse transformations assuming that the disease affects only certain specific regions, impose Lipschitz continuity of functions, and introduce a function $\mathrm{g}: \mathrm{Y} \to \mathrm{Z}$ to the model. By including the g function, the mapping functions, with different inputs $\mathbf{z}$, are constrained to capture sufficiently distinct imaging patterns, thus enabling the inverse mapping g to detect the correct latent variable/subtype within the PT group.

Having said that, the objective of the Smile-GAN model is the following:

$$\mathrm{L(D, f, g)} = \mathrm{L_{GAN}(D, f)} + \mu \mathrm{L_{change}(f)} + \lambda \mathrm{L_{cluster}(f, g)}$$

With

$$L_{GAN}(D,f) = E_{\mathbf{y}\sim p_{PT}}[\log(D(\mathbf{y}))] + E_{\mathbf{z}\sim p_{Sub},\mathbf{x}\sim p_{HC}}\left[1 - \log\left(D(f(\mathbf{x},\mathbf{z}))\right)\right]$$
$$L_{change}(f) = E_{\mathbf{x}\sim p_{HC},\mathbf{z}\sim p_{Sub}}[\|f(\mathbf{x},\mathbf{z}) - x\|_1]$$
$$L_{cluster}(f,g) = E_{\mathbf{x}\sim p_{HC},\mathbf{z}\sim p_{Sub}}[l_c(\mathbf{z}, g(f(\mathbf{x},\mathbf{z})))]$$

where l_c denotes the cross-entropy loss with $l_c(\mathbf{a},\mathbf{b}) = -\sum_{i=1}^{k} \mathbf{a}^i \log \mathbf{b}^i$.
The adversarial loss L_{GAN} compels the synthesized PT data to align with the distribution of real PT data. In this process, the discriminator D, which aims to distinguish between synthesized and real PT data, strives to maximize this loss, while concurrently, the mapping function f seeks to minimize it.
As discussed previously, the regularization terms L_{change} and $L_{cluster}$ are employed to constrain the function space where f is learned from.
The mapping function, $f: X * Z \rightarrow Y$, and the clustering function, $g: Y \rightarrow Z$, are learned through the following training procedure:

$$f, g = \arg\min_{f,g} \max_{D} L(D, f, g)$$

After training, the clustering function g is applied to the real PT data to estimate their subtype variables, indicating their respective clustering memberships.

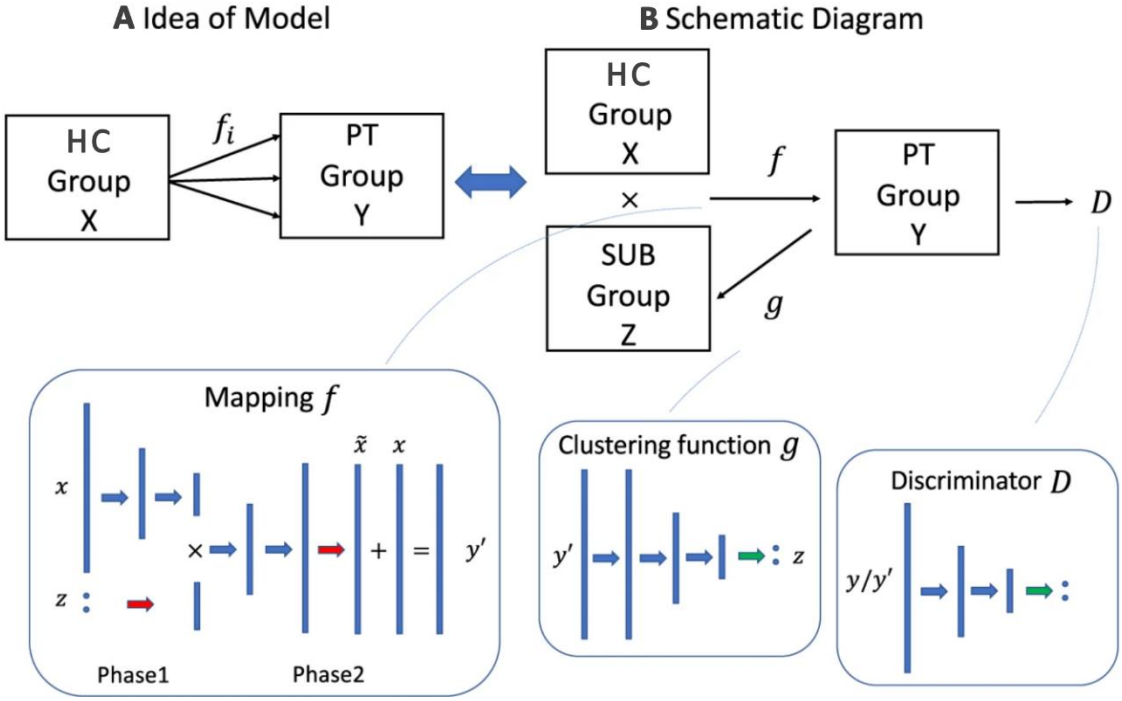


Figure 2.2: Smile-GAN model. *A) Smile-GAN model conceptualization. B) Smile-GAN architecture: Smile-GAN model learns one mapping function, f, from joint HC domain X and subtype domain Z to the PT domain Y while learning another function $g: Y \rightarrow Z$. The discriminator D tries to distinguish between synthesized PT data Y' and real PT data Y [114].*

2.2 Normative modeling approaches

Normative models offer statistical inferences at the individual level by comparing data to an expected 'normative' distribution or trajectory. This approach is frequently employed in growth charts to illustrate developmental changes in body weight and height as a function of age, where deviations from the normative growth curve are identified as outliers at each age point. Specifically, normative modeling extends this concept by establishing mappings between behavioral, demographic, or clinical characteristics and biological measures, offering centile estimates of variation across populations (**Figure 2.3.A**). This approach allows for the positioning of an individual within the normative distribution, thereby identifying the extent to which they deviate as an outlier in a given measure, thus offering a precise method to parse heterogeneity within cohorts.

Normative modeling comprises four key steps (**Figure 2.3.B**): First, a reference cohort is selected alongside relevant variables to define the mapping and population over which variability is assessed. Second, a statistical model is constructed to model the variance in a response variable based on clinically relevant predictors/independent variables within the reference cohort. For instance, a normative model might be developed to relate brain regional volumetrics to demographics such as age and sex using data from a population-based reference cohort. Third, the predictive accuracy of the model is evaluated using metrics such as mean-squared error and explained variance, with validation performed on withheld data through cross-validation to ensure reliable generalizability. Finally, the validated model is applied to assess deviations in samples from a target cohort, such as a patient cohort relative to the reference one.

Normative modeling techniques encompass a variety of methods, such as Gaussian process regression, hierarchical Bayesian regression, quantile regression, support vector regression, and autoencoders. While the primary focus of this thesis is not on normative modeling per se, a brief review of some principal normative models will be provided, with particular emphasis placed on autoencoders, which will be thoroughly discussed in Chapter 4.

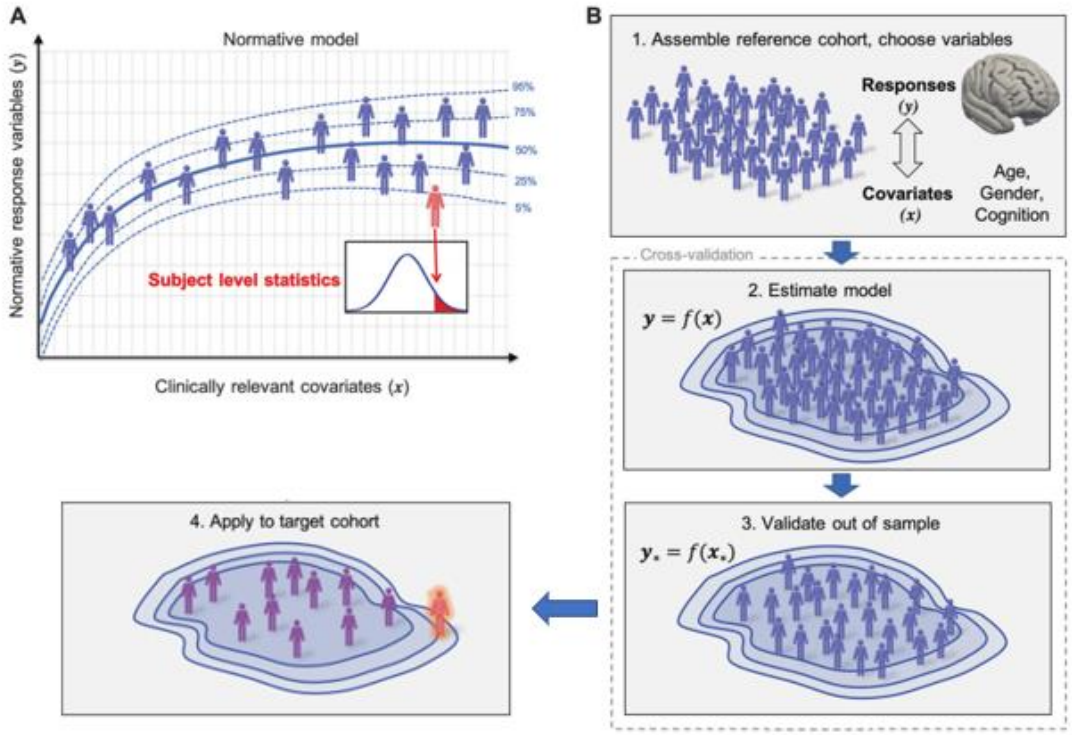


Figure 2.3: Normative modeling. *A) Normative modeling is analogous to the use of growth charts in pediatric medicine, but instead of traditional response variables such as body height, body weight, or head circumference, it utilizes biological measures, such as regional brain activity or neuroimaging phenotype. Similarly, classical predictors/independent variables such as age and sex are replaced with clinically relevant variables. The Gaussian distribution curve provides statistical inference at the individual level relative to the normative model (red figure). B) Normative modeling concept: Following selecting the reference cohort and variables, the normative model is estimated and then validated out-of-sample. The validated model is subsequently applied to a target cohort, such as a patient population. The figure is adapted from [88].*

2.2.1 Gaussian process regression

Gaussian process regression (GPR) is a robust non-parametric Bayesian approach that excels in normative modeling, particularly for understanding and quantifying individual deviations from normative patterns within a population. GPR leverages Bayesian inference principles to predict target variable values based on input features. It assumes the data can be represented by a Gaussian process, a collection of random variables with joint Gaussian distributions. It is completely specified by its mean function m(x) and covariance function (or kernel) k(x, x’):

$$\mathrm{f(x)}\sim\mathrm{GP(m(x),k(x,x'))}$$

with mean function:

$$\mathrm{m(x)} = \mathrm{E[f(x)]}$$

and covariance function:

$$\mathrm{k(x,x')} = \mathrm{E}\big[\big(\mathrm{f(x)} - \mathrm{m(x)}\big)\big(\mathrm{f(x')} - \mathrm{m(x')}\big)\big]$$

In normative modeling, GPR relies on the covariance function or kernel to define relationships between input space points. The choice of kernel, such as Radial Basis Function (RBF), Matérn, or linear kernels, is crucial as it encodes assumptions about the modeled function's properties. This kernel helps capture complex relationships between input variables (e.g., age, gender, genetic factors) and the target biological measure (e.g., cortical thickness). GPR places a prior distribution over potential data-describing functions, updated to a posterior distribution using Bayes' theorem when data is observed. This update provides probabilistic estimates of the target variable at new points, offering predictions and uncertainty estimates.

GPR is robust and flexible, capable of modeling complex, non-linear relationships between variables without assuming a specific functional form. Its ability to provide uncertainty estimates alongside predictions makes it particularly valuable in clinical and research settings where understanding variability and confidence is crucial. Additionally, as a Bayesian method, GPR can be especially effective with smaller datasets, leveraging the prior distribution to inform predictions[116].

2.2.2 Hierarchical bayesian regression

Hierarchical bayesian regression (HBR) is a sophisticated statistical approach that enhances normative modeling by incorporating multiple levels of variability and uncertainty. It provides a nuanced understanding of individual differences by modeling data at various hierarchical levels (e.g., individual, group, population) and integrating multiple sources of information. HBR, also known as multilevel or mixed-effects regression, extends traditional regression models by introducing parameters that vary across more than one level, making it particularly well-suited for complex datasets where observations are nested within larger groups.

In normative modeling, HBR captures variability at multiple levels and models how these levels interact. Consider the scenario where measurements are from individuals nested within groups (e.g., patients within different clinics).
At the individual level:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

where y_{ij} is the outcome for individual i in group j, x_{ij} are the predictors, β_{0j} and β_{1j} are group-specific intercepts and slopes, and ϵ_{ij} are individual-level errors.
At the group level:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad u_{0j} \sim N(0, {\tau_0}^2)$$
$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad u_{1j} \sim N(0, {\tau_1}^2)$$

where γ_{00} and γ_{10} are the overall intercept and slope, and u_{0j} and u_{1j} are group-level random effects.

HBR uses prior distributions to incorporate existing knowledge or assumptions about the parameters at each level, which are updated with data to produce posterior distributions. The Bayesian framework allows for the integration of prior knowledge and continuous updating of priors as new data becomes available. HBR efficiently handles nested data structures and provides credible intervals for estimates at each level, allowing for a nuanced understanding of variability and uncertainty. This approach is particularly valuable in clinical and research settings, where understanding individual deviations from group norms is crucial. For example, in neuroimaging, HBR can model the relationship between brain structural measures and demographic/clinical variables, providing probabilistic estimates and uncertainty quantification for both population-level trends and individual-specific deviations[117].

2.2.3 Autoencoders

Autoencoders are an artificial neural network used for unsupervised learning, focusing on dimensionality reduction and feature extraction. In normative modeling, they provide a robust method to understand and quantify individual deviations from a normative pattern within a population. Comprising an encoder, which compresses input data into a lower-dimensional representation, and a decoder, which reconstructs the input from this compressed form, autoencoders learn efficient data representations by minimizing the difference between the original and reconstructed data (**Figure 2.4**).

Mathematically, let x be the input data. The encoder function f maps x to a latent representation z:

$$\mathrm{z} = \mathrm{f}(\mathrm{x}) = \sigma(\mathrm{W_e x} + \mathrm{b_e})$$

where W_e and b_e are the weights and biases of the encoder, and σ is a non-linear activation function (e.g., ReLU, sigmoid).

The decoder function g maps z back to the reconstructed input $\hat{x}$:

$$\hat{\mathrm{x}} = \mathrm{g}(\mathrm{z}) = \sigma(\mathrm{W_d z} + \mathrm{b_d})$$

where W_d and b_d are the weights and biases of the decoder.

In normative modeling, autoencoders facilitate reducing high-dimensional biological data, such as brain imaging, into a compact latent space, capturing

the most relevant features that describe normative patterns. The encoder-decoder mechanism enables the model to understand the underlying structure and distribution of the reference cohort data, allowing for the identification of deviations based on reconstruction errors.

The commonly used loss function for this purpose is the mean squared error (MSE):

$$L(x, \hat{x}) = \|x - \hat{x}\|^2 = \|x - g(f(x))\|^2$$

These errors, indicating how well the model captures the normative pattern, can highlight individuals whose data significantly deviates from the norm, providing valuable insights into individual variability.

Applying autoencoders in normative modeling involves collecting relevant data from a reference cohort, training the model to minimize reconstruction errors, and validating it using techniques like cross-validation. Once trained, the autoencoder can be applied to new data points to identify deviations from the normative model. This approach is particularly beneficial for capturing complex, non-linear relationships and handling large, unlabeled datasets common in biomedical research. By providing detailed individual-level analysis through reconstruction errors, autoencoders offer a powerful tool for understanding and mapping individual deviations, making them invaluable in clinical and research settings[118].

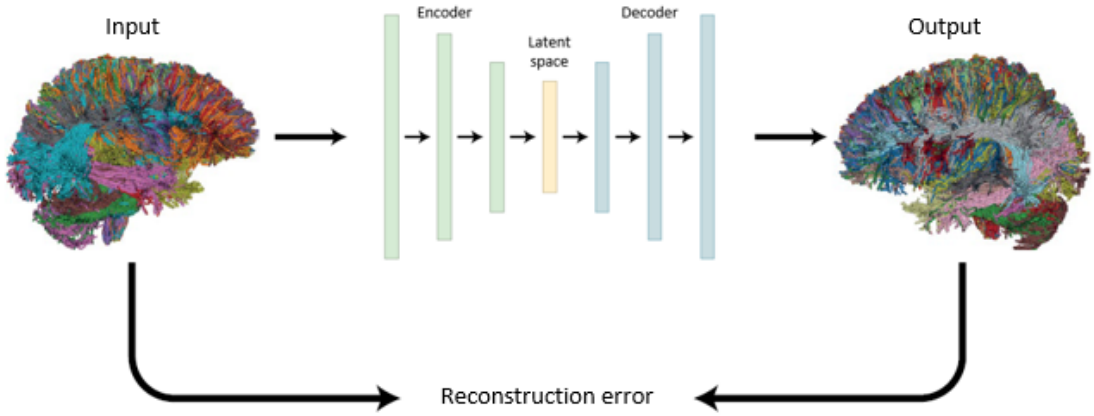


Figure 2.4: Autoencoders. *An autoencoder is a neural network comprising an encoder that receives high-dimensional input data (such as brain images), compresses it into a low-dimensional latent embedding, and a decoder that reconstructs the input data from the compressed representation. In the context of normative modeling, the network is trained on normative data from healthy individuals. The reconstruction error generated by comparing the network's output to its input indicates the brain data deviation from the normative pattern. The figure is adapted from [119].*

3 Study of structural brain change heterogeneity in aging at early asymptomatic phases using a semi-supervised deep learning clustering method

Over the past few decades, there has been a growing understanding of the neurobiological processes associated with various neuropathologies that affect the human brain, including Alzheimer's disease and cerebrovascular disease. However, the mechanisms by which individuals transition from normal aging to pathological manifestations remain poorly understood. This knowledge gap can be attributed, in part, to the limited availability of sufficiently large-scale neuroimaging datasets and the requisite tools for modeling and validating such complex processes. Investigating the neuroanatomical heterogeneity of aging at early stages, before the onset of clinical symptoms, may yield prognostic insights into preclinical stages of neurodegenerative diseases, paving pathways toward patient stratification and promoting precision medicine in clinical trials and healthcare.

This chapter explores the application of a deep learning (DL) method to identify subgroups with common patterns of structural variation in a large and diverse cohort of cognitively unimpaired participants, along with examining the associations of these subgroups with genetics, functional connectivity, white matter microstructure, biomedical measures, and cognitive decline trajectories.

3.1 Introduction

Aging is associated with complex changes in brain structure and function. Diverse genetic, environmental, and pathologic factors may trigger, aggravate, or protect against pathophysiologic processes that underlie neurodegeneration and its clinical manifestation[120]. These factors may act independently, synergistically, or antagonistically. Common age-associated neuropathologies such as Alzheimer’s and vascular-related diseases have long preclinical phases when magnetic resonance imaging can measure early brain changes[121][122]. Understanding early brain structural changes may provide prognostic information about susceptibility to or presence of neurodegeneration and inform patient management and stratification into clinical trials.

The investigation of heterogeneous brain changes in normal to early pathologic brain aging spectrum requires large and diverse databases, which are not typical of individual neuroimaging studies. New harmonization methods allow cross-cohort constructive integration of datasets, enabling rich mega-analyses. Additionally, novel artificial intelligence (AI) methods allow data-driven investigation into subtle patterns of brain change.

This chapter leverages an advanced semi-supervised DL clustering method discussed previously termed Smile-GAN to unravel brain structural heterogeneity in a large, diverse dataset of CU individuals drawn from 11 neuroimaging studies. Heterogeneity is analyzed separately across four decade-long age intervals ranging from 45 to 85 years, with decade intervals chosen to minimize age-related influences during clustering. It is hypothesized that subgroups of early structural brain variability can be identified, which will have distinct associations with genetics, functional connectivity, white matter integrity, biomedical measures, lifestyle risk factors, amyloid β, and trajectories of cognitive decline.

3.2 Methods

3.2.1 iSTAGING data

Data were drawn from the iSTAGING (imaging-based coordinate SysTem for AGIng and NeurodeGenerative diseases)[78][123][124] consortium, a collaborative effort to consolidate neuroimaging, clinical, and cognitive data from >39,000 individuals across the adult lifespan. Here, a total of 58,113 timepoints from 27,402 CU individuals, aged 45-85 years at baseline scan, were included from the following studies: Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging, Biomarker, and Lifestyle (AIBL) Study, Biomarkers of Cognitive Decline Among Normal Individuals (BIOCARD), Baltimore Longitudinal Study of Aging (BLSA), Coronary Artery Risk Development in Young Adults (CARDIA) study, Open Access Series of Imaging Studies (OASIS), University of Pennsylvania Memory Center cohort (Penn-PMC), Study of Health in Pomerania (SHIP), UK Biobank, Women's Health Initiative Memory Study (WHIMS), and Wisconsin Registry for Alzheimer's Prevention (WRAP). The supervisory committee of each study approved its inclusion in this project. The institutional review board of the University of Pennsylvania approved this project. According to the Declaration of Helsinki, all participants gave written informed consent to each study for data acquisition and analysis. Participant demographics for baseline and longitudinal cohorts are shown in **Tables 3.1-2**.

Table 3.1: Demographic summary and volumetric measures of the CU sample (baseline scans). *N=27,402. Abbreviations: WMH, white matter hyperintensities; Aβ, Amyloid β; APOE, Apolipoprotein E.*

Other races: Hispanic/Latino, Native American, Multiracial, unknown, other; information about races is presented as given in the originating studies.

Study	Sample size	Sex (% males)	Race				Age (years)	Aβ (% positive)	APOE-ε4 carriers	Total brain volume (mm^3)	Total WMH volume (mm^3)
			(% Asian)	(% Black)	(% other)	(% White)					
ADNI	670	39.85	1.28	7.40	2.54	88.78	71.64 (55.10 - 84.87)	34.34	39.55	1166260.64 (905059-1568464.41)	3703.72 (0-70106.37)
AIBL	559	39.36	0	0	0	100	72.70 (45.07 - 84.96)	34	29.87	1152897.60 (906669.64 -1510058)	4434.22 (0-78002.12)
BIOCARD	364	36.81	0.82	0.82	0.28	98.08	63.60 (45.34 - 84.91)	22.26	34.34	1207950.51 (943595.91 - 1635686.72)	2479.38 (0-35005.08)
BLSA	815	44.05	4.79	27.03	2,46	65.72	67.53 (45-84)	26.47	34.97	1148360.01 (842437.23 - 1598045.49)	4272.69 (0-71351.88)
CARDIA	828	46.62	0	39.98	0	60.02	51.77 (45-61)	0	49.28	1219338.74 (896746.64 -1659822)	1827.35 (39.10-19571)
OASIS	424	41.98	0.71	14.22	0	85.07	68.75 (46.04 - 84.99)	27.71	36.79	1149065.56 (872067.08 - 1525243.33)	2970.71 (0-34206.71)
PENN	216	33.33	0	24.07	1.39	74.54	69.91 (46-84.07)	36.36	71.76	1207878.47 (961902.49 - 1591351.67)	3967.49 (0-38562.40)
SHIP	2275	47.82	0	0	0	100	59.71 (45-84)	0	0	1247191.50 (849395.63 - 1667079.41)	1931.82 (0-64119.96)

UK Biobank	20043	47.29	1.23	0.50	1.13	97.14	62.60 (45-80)	0	39.94	1244275.83 (898920.93-1817049)	2236.01 (0-49631.96)
WHIMS	965	0	1.56	3.85	2.39	92.20	69.80 (64-83.85)	0	26.01	1108209.93 (752469.43 - 1420114.27)	3442.95 (0-52623.63)
WRAP	243	30.04	0.41	2.06	2.06	95.47	63.65 (49.60 - 77.70)	15.44	39.51	1201032.54 (985467-1569749)	2294.58 (0-25235)

Table 3.2: Demographic summary of the baseline CU sample having longitudinal scans. *N= 3,567. Abbreviations: APOE, Apolipoprotein E.*
Other races: Hispanic/Latino, Native American, Multiracial, unknown, other; information about races is presented as given in the originating studies.

Study	**Sample size**	**Sex (% males)**	**Race**				**APOE-ε4 carriers**
			(% Asian)	**(% Black)**	**(% other)**	**(% White)**	
ADNI	371	42.32	1.74	6.60	2.08	89.58	33.15
AIBL	212	41.51	0	0	0	100	26.42
BIOCARD	198	35.86	1.51	1.52	0	96.97	33.33
BLSA	451	41.02	5.32	28.38	2.22	64.08	29.27
CARDIA	427	47.78	0	35.60	0	64.40	34.43
OASIS	113	47.79	1.77	10.62	0	87.61	25.66
PENN	46	34.78	0	21.74	2.17	76.09	43.48
UK Biobank	1213	48.64	0.91	0.58	0.9	97.61	40.89
WHIMS	415	0	1.69	2.65	2.41	93.25	25.30
WRAP	121	32.23	0.83	0.83	0.82	97.52	35.54

3.2.2 Image pre-processing

A fully automated processing pipeline was applied to extract morphometric variables from structural MRI. T1-weighted image intensity inhomogeneity was corrected[125], followed by multi-atlas skull-stripping[126]. 145 anatomic regions of interest (ROIs) were segmented using a multi-atlas, multi-warp label fusion-based method[127]. Inter-study ROI harmonization was performed using the Neuroharmonize toolbox[78] based on the multi-variate ComBat method combined with generalized additive cubic spline models to capture nonlinear age, sex, and deep learning-based intracranial volume measurement (DLICV)[128] effects. This method reduces acquisition-related effects and preserves variability due to biological covariates. Specifically, each ROI volume was modeled as a nonlinear function of age, sex, and baseline DLICV. Based on the adjusted data, the remaining systematic differences in shift (location) and variance (scale) were attributed to site-specific acquisition settings and adjusted conservatively with an empirical Bayes regularization. The harmonization model was trained on each site baseline scans of CU and then applied to the entire dataset. This method has already been validated in other works[123][129][124].

White matter hyperintensities were segmented from FLAIR and T1-weighted images using a DL-based method[128]. A semi-automated visual quality check tool (https://github.com/CBICA/MRISnapshot) was used to review WMH segmentation manually. The imaging parameters for each study are presented elsewhere[130].

3.2.3 Study design

Subgroups of structural brain measures of CU individuals (N=27,402) were independently examined in four decade-long age intervals spanning 45-85 years; decade intervals were used to mitigate age-related effects during clustering. The first decade spans ages 45 to younger than 55, notated [45,55). Participants older than 85 were excluded due to insufficient sample availability. Within each age interval, the 145 harmonized ROI volumes were linearly corrected for continuous age, sex, and DLICV to avoid biasing the clustering with disease-unrelated neuroanatomical variations. Linear correction was performed due to the limited age range within each interval. WMH volumes were cube-root transformed due to skewness and then adjusted for the same covariates. Corrected data were standardized to z-scores.

Principal component analysis (PCA) was separately applied to age-, sex-, and DLICV-adjusted and harmonized anatomic ROI and cube-root transformed WMH volumes for dimensionality reduction with the ultimate goal of detecting

a group with low atrophy and WMH load called resilient brain agers (A0). Specifically, for anatomic ROIs, PCA was applied to intermediate-resolution gray matter (GM) ROIs, ventricles, and corpus callosum from the MUSE[127] atlas to capture regional rather than focal atrophy patterns, avoiding the high variability of smaller ROIs. The first principal component (ROI-PC1) explained 20, 20, 21, and 20% variance in the data in the [45,55), [55,65), [65,75), and [75,85) age groups, respectively. ROI-PC1 exhibited a positive association with GM ROI volumes and a negative association with ventricular volume. Outliers (defined as one standard deviation (SD) from the mean) were selected as the atrophy-resilient subjects (ROI-A0), while all other participants were classified as atrophy-vulnerable subjects (ROI-rest). For WMH, PCA was applied to 8 lobar cerebral WMH volumes since WMH development varies across the brain. The first principal component (WMH-PC1) explained 52, 56, 64, and 68% of the data variance in the [45,55), [55,65), [65,75), and [75,85) age groups, respectively, and it was positively correlated with lesion volumes. Thus, WMH-PC1 was utilized to distinguish participants with low lesion burden (WMH-A0) from those with high lesion burden (WMH-rest). Participants assigned to ROI-A0 and WMH-A0 groups were ultimately labeled resilient brain agers A0, and everyone else was labeled non-A0. For anatomic ROI- and WMH-PCA, only the first principal component was utilized, as it was the only component interpretable within the context of distinguishing between A0 and non-A0. The dichotomization of ROI- and WMH-PC1, such as whether one or more SD was selected, was empirical, with the sole prerequisite being the assurance of a sufficient sample size for the A0 group defined by the intersection of ROI- and WMH-A0.

Using A0 as a reference, heterogeneity within the remaining samples was investigated by fitting a Smile-GAN model independently for each age group. Smile-GAN was trained jointly on the 145 anatomic ROI and 8 lobar WMH volumes. Clustering methods[131][84][132] used to quantify heterogeneity in neuroimaging are often limited by disease-irrelevant confounding variability. Smile-GAN, by learning a one-to-many mapping from the reference (A0) to the target domains (non-A0), models disease heterogeneity without being confounded by disease-unrelated factors (e.g., demographics) detectable in A0. PCA and Smile-GAN models trained on baseline scans were applied to available longitudinal scans within each age group.

3.2.4 Model longitudinal stability

Since clustering was performed using the model based on the age at the time of scanning, an analysis was conducted to determine whether transitions between study-defined age decades affect clustering stability/reproducibility,

using individuals with longitudinal scans (N=3,567). Longitudinal clustering stability was assessed for participants with scans acquired in multiple age groups, clustered using independently derived models. This stability was compared to the reference stability of longitudinal imaging for participants who remained within a single age group during follow-up.

3.2.5 Genetic analysis

Smile-GAN probability scores were used as phenotypes in GWAS, utilizing imputed genotyping data from the UK Biobank. Multiple linear regressions were performed, controlling for continuous age, sex, DLICV, and the first 40 genetic principal components via Plink 2 (v2.0.0)[133]. Given the observed longitudinal clustering stability, GWAS was performed for the entire age range [45,85) (N=18,282; 47% males). FUMA[134] was used to identify and annotate candidate SNPs, independent significant SNPs, (top) lead SNPs, and genomic loci. The top lead SNP within each locus was queried to determine whether the locus was novel—not previously associated with any clinical traits—and the candidate SNPs were explored for phenome-wide associations using the GWAS Catalog[135]. Additionally, SNP-based heritability estimates (h2) were calculated using genome-wide complex trait analysis (GCTA) (v1.93.2)[136]. Finally, Smile-GAN probability scores were associated with the polygenic risk score (PRS) for two subtypes of late-life depression (LLD1 and LLD2), as developed in previous studies[129][137]. LLD1 is characterized by preserved brain structure, while LLD2 shows diffuse brain atrophy.

3.2.6 Functional connectivity and white matter microstructural integrity

Between-network functional connectivity (FC) was analyzed based on 21 FC networks extracted using group-independent component analysis (ICA) on rsfMRI data from the UK Biobank study[77]. Additionally, fractional anisotropy (FA) maps derived from diffusion tensor imaging (DTI) data from the UK Biobank were used to measure WM microstructural integrity[77]. Mean FA values were extracted from 48 white matter tracts using the Johns Hopkins University tract atlas[138]. Linear regression was applied to associate the Smile-GAN subgroups with the $(20 * 21)/2 = 210$ internetwork FC and 48 FA features, adjusting for age, sex, and subgroup labels as covariates. Given the subgroup consistency across the four age intervals, functional connectivity (N=19,143; 47% males) and fractional anisotropy (N=3,443; 48% males) were examined across the entire 45-85 age range.

3.2.7 Statistical analysis

Voxel-based morphometry (VBM)[139][140] as implemented in Statistical Parametric Mapping (SPM, Version 12, https://www.fil.ion.ucl.ac.uk/spm/) running on MATLAB (R2017b, Mathworks Inc.) was used to compare subgroups in gray matter patterns using tissue density maps (regional analysis of volumes examined in normalized space, RAVENS)[141], considering continuous age, sex, and DLICV as covariates. Multiple-voxel testing was corrected by controlling the Family Wise Error rate (FWE) via random field theory[142] at 0.1%.

Complementary to mass-univariate voxel-based subgroup comparisons, a manifold learning technique called locally linear embedding (LLE)[143][144] was applied to map high-dimensional imaging patterns into a low-dimensional space that allowed visualization of multivariate data. The LLE algorithm was applied to the 145 covariates-adjusted harmonized anatomic ROI volumes in each age group. Standard dense matrix operations were used for the eigenvalue decomposition; the number of neighbors was set to 150, and the number of manifold coordinates was set to 3 for illustration purposes.

The clinical, cognitive, biomarker, and Apolipoprotein E allele associations of the subgroups were examined separately in each age group. Linear and logistic regressions were performed for continuous (e.g., Trail Making Test B) and categorical features (e.g., smokers vs. non-smokers), respectively. For cognitive outcomes having overdispersed and skewed distributions (e.g., mini-mental state examination, MMSE), the beta-binomial distribution[145] was fitted. The regression models included subgroup labels while adjusting for continuous age, sex, and study (and education for cognitive scores). For features showing consistent trends across >1 age group, the data from multiple age groups were pooled together, and subgroup differences were reexamined using one model in the combined dataset over broader age ranges considering the study*age interaction term. Differences across subgroup intercepts were assessed using the Wald test[146]. Multiple comparison corrections were conducted for the number of features by controlling the false discovery rate (FDR)[147] at 5%.

3.2.8 Longitudinal outcomes analysis

Linear mixed-effects (LME) models with subject-specific random intercepts were fitted to estimate the rate of change (RC) per year for atrophy, WMH, cognition, SPARE-AD (spatial pattern of abnormality for recognition of early Alzheimer's disease)[148] — a signature of AD-specific regional brain atrophy, which has also been found to predict progression from normal cognition to MCI

— and SPARE-BA (spatial pattern of atrophy for recognition of brain aging)[149] — a structural MRI-based brain age estimation. Both SPARE models were previously validated[123][124][148][149][150][151]. The LME models included subgroup indicators, time of visit, and their interaction term while adjusting for baseline age, sex, study, education, and DLICV. RC subgroup comparisons were conducted using the Wald test. The longitudinal analyses were conducted considering individuals with ≥4 longitudinal measures to reduce uncertainty in slope estimation.

Development of MCI defined by the individual participating study was used to indicate longitudinal cognitive deterioration. Survival curves for time to progression to MCI were generated using a nonparametric Kaplan–Meier estimator[152]; the log-rank test[153] was used to compare the curves between subgroups.

3.2.9 Amyloid β status

Amyloid cutoffs calculation. Levels of amyloid accumulation were measured using three techniques: CSF (cerebrospinal fluid) Aβ 1-42 (Aβ42), Pittsburgh compound B ([^{11}C]PiB), and [^{18}F]Florbetapir. Two cutoffs in each outcome measurement divided ranges into positive (Aβ+), unclear, and negative (Aβ-) status. These cutoffs were computed for each technique and each site separately, as differences in acquisition and pre-processing methods yielded significantly different outcome distributions. First, a reference study was determined per acquisition technique with well-validated cutoffs. Then, equivalent cutoffs for subsequent studies were calculated by matching the new distribution of CU individuals to that of the reference distribution using normal distribution fits. Age and sex ratio were statistically matched between participants of the pairwise distributions (P-value>0.15).

CSF- Aβ42 measures were provided by ADNI, BIOCARD, and PENN. For ADNI, Aβ42<180pg/mL was labeled as Aβ+ and Aβ42>200pg/mL as Aβ-. Based on prior work, Aβ42<192pg/mL has been considered consistent with the presence of cerebral amyloid using the Luminex platform[154]. Then, the equivalent cutoffs were computed for BIOCARD (311.1pg/mL for Aβ+, 350.5pg/mL for Aβ-) and PENN (230.4pg/mL for Aβ+, 256.8pg/mL for Aβ-).

[^{11}C]PiB standardized uptake value ratio (SUVR) measurements from amyloid-PET were provided by ADNI, AIBL, BIOCARD, BLSA, OASIS, and WRAP. First, for OASIS, [^{11}C]PiB>1.50 (high amyloid burden group according to their previous study) was labeled as Aβ+, and [^{11}C]PiB<1.25 (low burden group) was labeled as Aβ- based on previously reported cutoffs[155]. Equivalent

cutoffs were computed for ADNI (1.55 and 1.44), AIBL (1.31 and 1.20), BIOCARD (1.37 and 1.25), BLSA (1.08 and 1.03), and WRAP (1.28 and 1.18).

[^{18}F]Florbetapir (also known as [^{18}F]AV45) SUVR measurements from amyloid-PET were provided by ADNI and OASIS, and their distributions were similar in terms of the cutoffs. Previously, a cutoff of 1.11 was established[156]. To be conservative, [^{18}F]Florbetapir>1.15 was labeled as Aβ+, and [^{18}F]Florbetapir<1.05 was labeled as Aβ- for both studies.

Final Aβ status definition: CSF- Aβ42 provided by ADNI, PENN, and BIOCARD, [^{11}C]PiB provided by ADNI, AIBL, BIOCARD, BLSA, OASIS, and WRAP, and [^{18}F]AV45 provided by ADNI and OASIS were used as amyloid measures. For studies with more than one amyloid measure, the final Aβ status was defined based on the [^{18}F]AV45, and if this was not available, the [^{11}C]PiB determined the final Aβ status. If both PET measures were unavailable, the Aβ status was defined based on the CSF- Aβ42.

3.2.10 Clinical risk factors

Based on the originating studies, participants being either current or former smokers were grouped under the '*smoker*' label, while those who have never smoked were labeled '*non-smoker*' in this study. Individuals with body mass index (BMI) higher than 30 kg/m^2 were labeled '*obese*' whereas those with 18.5<BMI<24.9 kg/m^2 were labeled as '*normal*'. Based on the originating studies, individuals with hypertension status '*negative/absent*' or '*remote/inactive*' were grouped as '*hypertension negative*'. In contrast, those with '*hypertension positive/present/recent/active*' status were labeled '*hypertension positive*' in this analysis—the same for diabetes and hyperlipidemia status. Contrary to the other studies, SHIP blood biochemistry was collected in a non-fasted state; therefore, SHIP participants were excluded from the analysis of blood measures (high/low-density lipoprotein) and relevant CVD risk factors (diabetes, hyperlipidemia). Individuals with one or two APOE-ε4 alleles were considered APOE-ε4 carriers, while individuals with zero alleles were considered APOE-ε4 non-carriers. For depression status, only the CARDIA and UK Biobank studies were included due to their significantly higher availability compared to other studies and the fact that depression was self-reported in both.

3.3 Results

3.3.1 Consistent accelerated brain aging patterns across age groups.

PCA defined the A0 resilient group as participants with the lowest atrophy and WMH volume within each age group. In reference to A0, Smile-GAN showed optimal stability for three clusters (K=3), measured by the Adjusted Rand Index[157][158] (**Table 3.3**). Two types of phenotypes from this clustering scheme were used for subsequent analyses. The Smile-GAN subgroup probability was the direct model output, representing a continuous variable for each of the three clusters for each participant, with the sum of these three probabilities equaling 1; the Smile-GAN subgroup was decided by taking the highest probability (dominant subgroup).

***Table 3.3**: **Demographic summary of the subgroups in each age group.** Abbreviations: ARI, adjusted rand index; DLICV, deep learning-based intracranial volume measurement; UKBB, UK Biobank. Other races: Hispanic/Latino, Native American, Multiracial, unknown, other; information about races is presented as given in the originating studies.*

Age group	ARI	Subgroup	Sample size	Sex (% males)	Age (years)	Race	Study (%)	Education	DLCIV (mm^3)
[45,55)	0.34 ±0.12	A0	420	47.62	50.89 (45-54.71)	Total:316 Asian:1.58 % Black:5.06 % Other:1.27 % White:92.09 %	ADNI:0 AIBL:0 BIOCARD:0.95 BLSA:2.38 CARDIA:5.24 OASIS:0.71 PENN:0.24 SHIP:24.76 UKBB:65.24 WHIMS:0 WRAP:0.48	Total:380 Level 1:0.26% Level 2:17.11% Level 3:24.47% Level 4:58.16%	1450056 (1110097-1936156.50)
		A1	1721	43.23	50.83 (45-54.90)	Total:1418 Asian:1.06 % Black:6.63 % Other:2.04 % White:90.27 %	ADNI:0 AIBL:0.06 BIOCARD:0.81 BLSA:1.98 CARDIA:10.28 OASIS:0.41 PENN:0.12 SHIP:17.55 UKBB:68.56 WHIMS:0 WRAP:0.23	Total:1600 Level 1:0.06% Level 2:19.88% Level 3:22.19% Level 4:57.87%	1425628 (906000.8-2022055.38)
		A2	1376	43.17	50.82 (45-54.86)	Total:1269 Asian:1.81 % Black:12.69 % Other:1.18 % White:84.32 %	ADNI:0 AIBL:0 BIOCARD:1.45 BLSA:0.73 CARDIA:23.04 OASIS:0.44 PENN:0.15 SHIP:7.78 UKBB:66.13 WHIMS:0 WRAP:0.29	Total:1317 Level 1:0 Level 2:15.57% Level 3:27.48% Level 4:56.95%	1429914 (1035778-1994671.5)
		A3	1708	45.02	50.97 (45-54.92)	Total:1442 Asian:3.12 %	ADNI:0 AIBL:0.06 BIOCARD:1.76 BLSA:3.16	Total:1592 Level 1:0.19% Level 2:17.84%	1433562 (1010006-1944471.13)

						Black:5.83 % Other:2.28 % White:88.77 %	CARDIA:7.14 OASIS:0.82 PENN:0.06 SHIP:15.52 UKBB:71.08 WHIMS:0 WRAP:0.41	Level 3:24.18% Level 4:57.79%	
[55,65)	0.49 ±0.15	A0	832	48.44	59.77 (55-64.96)	Total:755 Asian:0.53 % Black:2.25 % Other:1.33 % White:95.89 %	ADNI:0.60 AIBL:0.24 BIOCARD:1.44 BLSA:2.88 CARDIA:0.72 OASIS:0.84 PENN:0.24 SHIP:9.01 UKBB:82.57 WHIMS:0.48 WRAP:0.96	Total:781 Level 1:0 Level 2:17.16% Level 3:23.43% Level 4:59.41%	1459593 (1132562-1906292.13)
		A1	3332	41.57	59.78 (55-64.96)	Total:3030 Asian:1.35 % Black:1.85 % Other:1.06 % White:95.74 %	ADNI:0.69 AIBL:0.18 BIOCARD:1.50 BLSA:1.95 CARDIA: 1.77 OASIS:1.08 PENN:0.21 SHIP:8.64 UKBB:82.62 WHIMS:0.15 WRAP:1.2	Total:3075 Level 1:0.03% Level 2:16.36% Level 3:22.99% Level 4:60.62%	1.430189 (9.733343-1922659.25)
		A2	2241	42.84	59.84 (55-64.98)	Total:2071 Asian:1.74 % Black:3.58 % Other:1.2% White:93.48 %	ADNI:0.62 AIBL:0.18 BIOCARD:1.34 BLSA:1.52 CARDIA:3.97 OASIS:0.85 PENN:0.49 SHIP:7.18 UKBB:82.11 WHIMS:0.54 WRAP:1.2	Total:2072 Level 1:0.05% Level 2:16.78% Level 3:23.17% Level 4:60%	1.435182 (1.082587-1947952)
		A3	2807	42.32	59.64 (55-64.99)	Total:2567 Asian:1.36 % Black:1.99 % Other:1.44 % White:95.21 %	ADNI:0.78 AIBL:0.43 BIOCARD:1.46 BLSA:2.92 CARDIA:1.28 OASIS:1.35 PENN:0.39 SHIP:7.98 UKBB:81.72 WHIMS:0.25 WRAP:1.43	Total:2598 Level 1:0.15% Level 2:15.20% Level 3:21.56% Level 4:63.09%	1.435505 (9.948574-2003986.63)
[65,75)	0.48 ±0.16	A0	1047	46.61	69.06 (65-74.93)	Total:954 Asian:0.63 % Black:1.68 % Other:1.15 % White:96.54 %	ADNI:4.68 AIBL:2.2 BIOCARD:1.24 BLSA:3.25 CARDIA:0 OASIS:2.39 PENN:1.15 SHIP:6.97 UKBB:68.67 WHIMS:8.79 WRAP:0.67	Total:919 Level 1:0 Level 2:14.15% Level 3:24.81% Level 4:61.04%	1.444433 (1.089337-1866891.88)
		A1	3916	45.58	69 (65-75)	Total:3648 Asian:0.93 % Black:1.81 % Other:0.91 % White:96.35 %	ADNI:3.86 AIBL:3.35 BIOCARD:1.15 BLSA:2.40 CARDIA:0 OASIS:2.09 PENN:1.12 SHIP:5.13 UKBB:73.44	Total:3429 Level 1:0.35% Level 2:14.2% Level 3:24% Level 4:61.45%	1.430995 (9.609970-1945619)

							WHIMS:6.44 WRAP:1.02		
		A2	2822	44.90	69.13 (65-74.89)	Total:2642 Asian:0.79 % Black:2.61 % Other:0.61 % White:95.99 %	ADNI:3.12 AIBL:2.66 BIOCARD:0.82 BLSA:2.41 CARDIA:0 OASIS:1.45 PENN:0.99 SHIP:4.68 UKBB:74.81 WHIMS:8.11 WRAP:0.96	Total:2448 Level 1:0.2% Level 2:15.77% Level 3:26.39% Level 4:57.64%	1.433579 (1.066332-2213104)
		A3	3256	46.16	68.87 (65-74.95)	Total:3018 Asian:0.66 % Black:1.79 % Other:1% White:96.55 %	ADNI:3.44 AIBL:3.56 BIOCARD:1.50 BLSA:2.09 CARDIA:0 OASIS:1.38 PENN:1.2 SHIP:5.65 UKBB:72.79 WHIMS:7.40 WRAP:0.98	Total:2824 Level 1:0.25% Level 2:15.12% Level 3:23.69% Level 4:60.94%	1.430858 (1.027622-2045929.88)
[75,85)	0.55 ±0.12	A0	187	49.20	77.35 (75-84.16)	Total:169 Asian:1.18 % Black:2.96 % Other:0 White:95.86 %	ADNI:9.63 AIBL:12.3 BIOCARD:1.60 BLSA:11.76 CARDIA:0 OASIS:4.81 PENN:0.53 SHIP:7.49 UKBB:43.32 WHIMS:8.02 WRAP:0.53	Total:159 Level 1:1.89% Level 2:15.09% Level 3:19.5% Level 4:63.52%	1.436035 (1.123752-1834656.38)
		A1	689	50.51	77.51 (75-84.99)	Total:628 Asian:2.07 % Black:4.62 % Other:0.63 % White:92.68 %	ADNI:13.35 AIBL:10.16 BIOCARD:2.03 BLSA:12.77 CARDIA:0 OASIS:5.37 PENN:3.34 SHIP:5.52 UKBB:42.96 WHIMS:4.35 WRAP:0.15	Total:585 Level 1:0.34% Level 2:12.48% Level 3:21.71% Level 4:65.47%	1.417356 (1.120537-1886061.13)
		A2	561	48.66	77.62 (75-84.90)	Total:478 Asian:1.26 % Black:6.69 % Other:0.84 % White:91.21 %	ADNI:8.91 AIBL:9.27 BIOCARD:1.07 BLSA:14.26 CARDIA:0 OASIS:3.92 PENN:2.85 SHIP:11.23 UKBB:40.82 WHIMS:7.49 WRAP:0.18	Total:467 Level 1:1.5% Level 2:14.13% Level 3:27.62% Level 4:56.75%	1.424865 (1.051920-1811627)
		A3	487	52.77	77.41 (75-84.97)	Total:431 Asian:1.62 % Black:3.25 % Other:0.93 % White:94.2 %	ADNI:9.45 AIBL:8.83 BIOCARD:2.05 BLSA:9.86 CARDIA:0 OASIS:6.78 PENN:3.29 SHIP:9.03 UKBB:42.92 WHIMS:7.39 WRAP:0.41	Total:421 Level 1:1.19% Level 2:11.4% Level 3:22.57% Level 4:64.84%	1.436247 (1.089950-1848784.75)

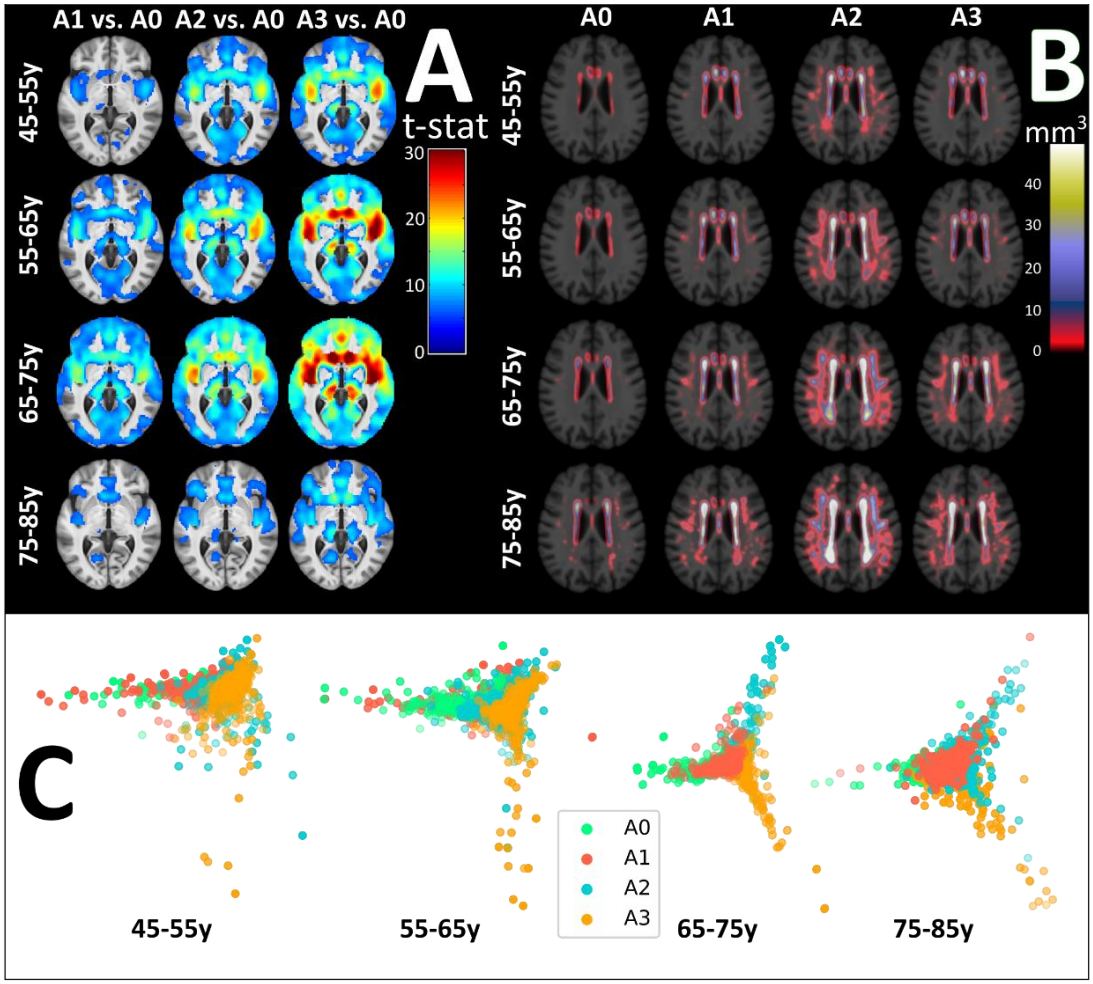


Figure 3.1: Structural profile of the brain aging subgroups for the four age groups. *A) Significant GM volumetric reduction (p_{FWE}<0.001) for the Smile-GAN subgroups compared to the A0 group in each age group. Warmer (cooler) colors indicate regions with severe (low) GM atrophy. An overlay brain template in gray colors is used. B) Average WMH maps computed by averaging WMH RAVENS maps aligned to a common atlas space within each ROI. Pinkish colors indicate regions with lower WMH burden, while whitish colors indicate high WMH burden regions. An overlay brain template in gray colors is used. C) 3D projected LLE-space derived from brain volumetric measures. The data points have been colored based on the subgroup labels. This projection allows visualization of subgroups across the age groups; as a projection, the axes are not directly meaningful.*

Although derived independently by age decade, the Smile-GAN subgroups, A1, A2, and A3, showed consistent differences in atrophy and WMH load compared to A0 (**Figure 3.1A-B**). A1 showed mild, predominantly peri-Sylvian atrophy. A2 displayed greater peri-Sylvian atrophy accompanied by atrophy in orbitofrontal and other prefrontal regions. A3 had diffuse atrophy across the brain, including medial frontal regions and thalamus (**Figure 3.1A**). WMH burden was higher in A2 than in the other subgroups (**Figure 3.1B**). Among A1/A2/A3, A1 had the least atrophy and was the largest subgroup, so it may be considered 'typical' aging. In comparison, A2 (highest lesions) and A3 (most severe atrophy) are considered 'accelerated' aging subgroups. VBM within the [75,85) group showed less prominent between-subgroup differences due to

relatively more advanced atrophy in the [75,85) A0 compared to younger A0 groups and more structural variability in this age group (**Figure 3.1A**). The average brain age difference between the youngest- and oldest-appearing brains (A0 vs. A3) was ~10 years and relatively consistent across age groups (**Figure 3.2**).

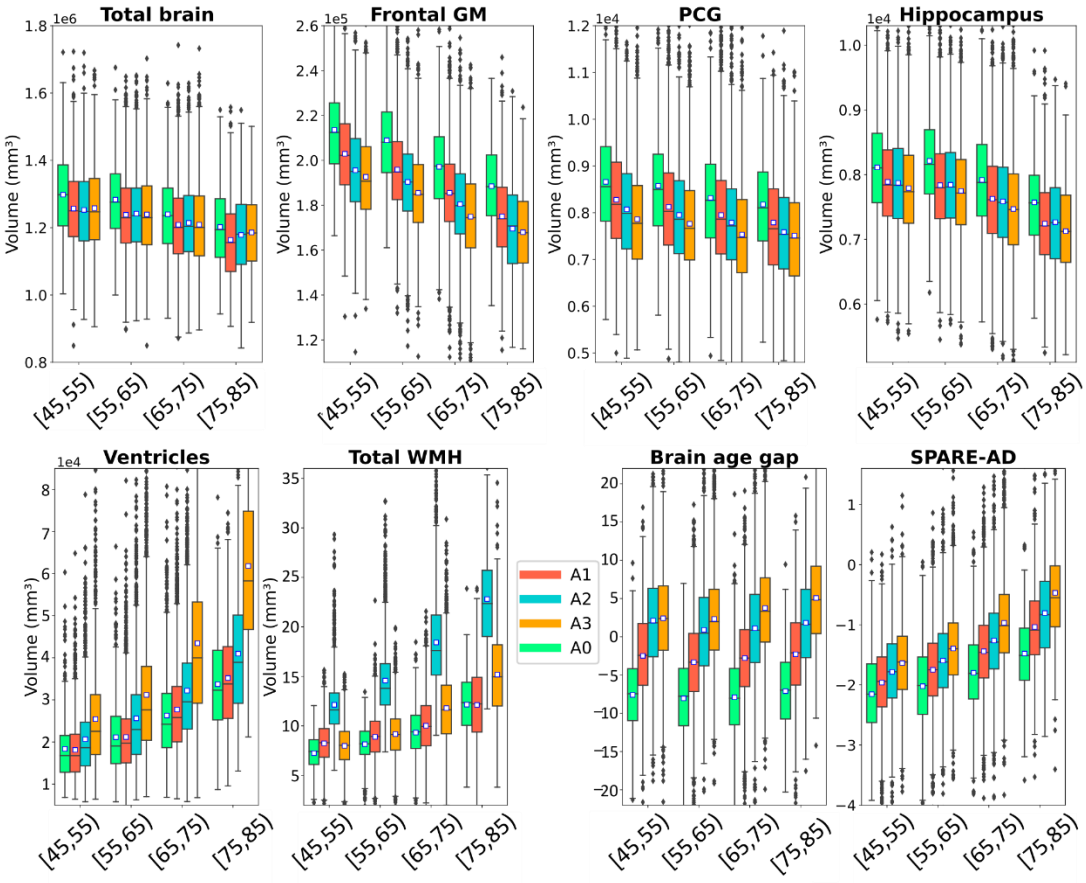


Figure 3.2: Anatomic ROIs, total WMH (cube-root transformed), SPARE-AD, and brain age gap for the subgroups for the four age groups. *The brain age gap is the chronological age subtracted from the structural MRI-based brain age estimation (SPARE-BA). Abbreviations: SPARE-AD, spatial pattern of abnormality for recognition of early Alzheimer's disease; WMH, white matter hyperintensities; GM, gray matter; PCG, posterior cingulate gyrus.*

Longitudinal scans within one age interval showed approximately 85% consistency of cluster assignment. Around 80% longitudinal stability of clustering assignments was observed in participants who aged into the next interval within a follow-up<=3 years, even though independent clustering models were applied to scans at different age intervals (e.g., participants classified as A2 using the [55,65) model were mainly classified as A2 on follow-up scans using the [65,75) model). Furthermore, **Tables 3.4-5** display the mean Smile-GAN probability shifts between two consecutive scans within the same age group and across different age groups, respectively. These findings

indicate that the magnitude of probability changes across decades was comparable to those observed within the same decade.

Table 3.4: Mean Smile-GAN probability changes between two consecutive scans within the same age group. *2,775 subjects have at least two scans within the same study-defined age group.*

Smile-GAN subgroup	Mean Smile-GAN probability change		
	A1(i+1)-A1(i)	A2(i+1)-A2(i)	A3(i+1)-A3(i)
	Age group [45,55)		
A1	0.02±0.25	-0.001±0.17	-0.02±0.16
A2	-0.09±0.22	0.18±0.30	-0.09±0.20
A3	-0.06±0.16	-0.01±0.18	0.07±0.19
	Age group [55,65)		
A1	0.04±0.20	-0.03±0.15	-0.01±0.13
A2	-0.05±0.18	0.06±0.22	-0.002±0.14
A3	-0.08±0.18	-0.04±0.15	0.12±0.21
	Age group [65,75)		
A1	0.06±0.19	-0.06±0.14	0.01±0.13
A2	0.01±0.13	0.04±0.17	-0.05±0.12
A3	-0.07±0.17	-0.002±0.12	0.07±0.20
	Age group [75,85)		
A1	0.07±0.17	-0.03±0.13	-0.03±0.12
A2	-0.01±0.12	0.03±0.15	-0.02±0.09
A3	0.004±0.11	-0.01±0.11	0.005±0.15

Table 3.5: Mean Smile-GAN probability changes between two consecutive scans in different age groups. *1,201 subjects cross study-defined age groups at least once.*

Smile-GAN subgroup	Mean Smile-GAN probability change		
	A1(i+1)-A1(i)	A2(i+1)-A2(i)	A3(i+1)-A3(i)
	Age group [45,55)		
A1	-0.01±0.25	-0.03±0.19	0.04±0.18
A2	-0.09±0.24	0.18±0.31	-0.09±0.24
A3	-0.09±0.23	0.06±0.21	0.03±0.28
	Age group [55,65)		
A1	0.10±0.26	-0.11±0.19	0.002±0.19
A2	-0.02±0.18	0.12±0.24	-0.10±0.19
A3	-0.09±0.23	0.01±0.17	0.08±0.26
	Age group [65,75)		
A1	0.07±0.26	-0.03±0.19	-0.03±0.19
A2	-0.03±0.15	0.02±0.22	0.003±0.16
A3	0.002±0.16	-0.08±0.22	0.08±0.26

Although VBM suggested a primary difference in severity across subgroups, examination of differences in location and severity of atrophy identified unique volumetric fingerprints across subgroups. **Figure 3.1C** shows the 3D projected LLE-space derived from brain volumetric measures, revealing worse atrophy in A1 compared to A0, followed by diverging branches for A2 and A3, especially after age 65. These axes echo the variability of distances between regional measures seen in radial plots (**Figure 3.3**). WMH volumes were not included in LLE analyses; the 2-axes divergence exclusively reflects atrophy subgroups and not the distinct difference in WMH burden.

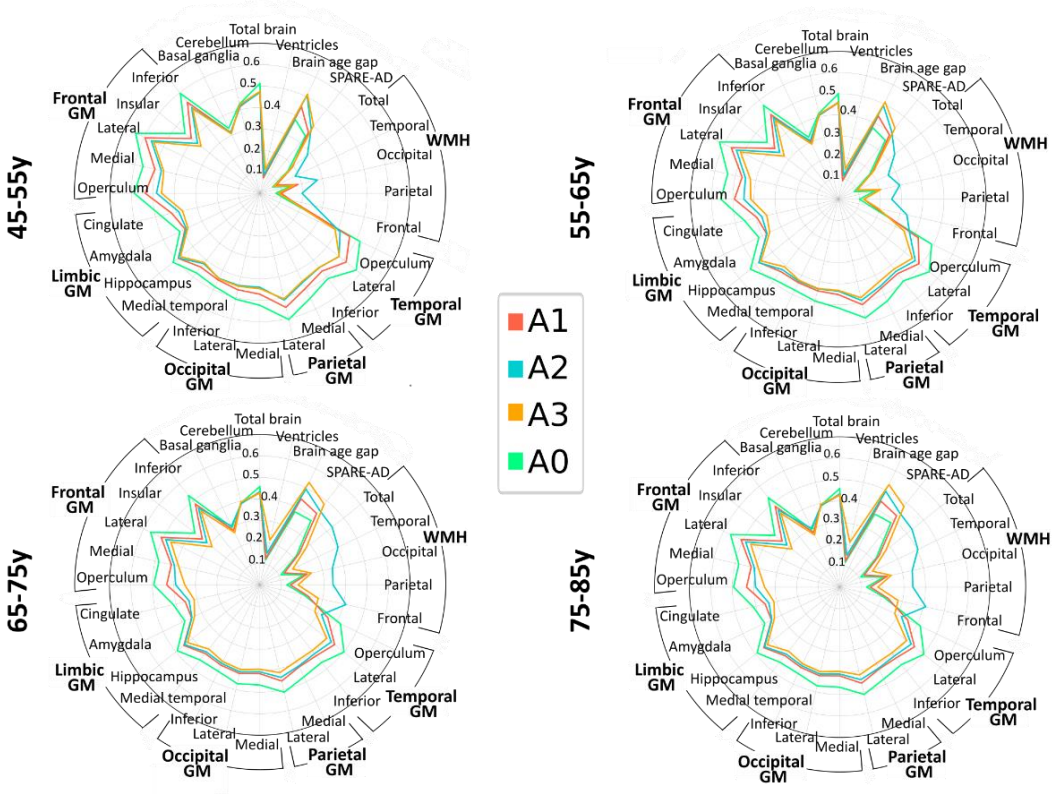


Figure 3.3: Radial plots showing the subgroup-specific mean of the (min-max) scaled values of the ROI/WMH volumes, SPARE-AD, and brain age gap. *WMH volumes were first cube-root transformed. The brain age gap is the chronological age subtracted from the structural MRI-based brain age estimation (SPARE-BA). Abbreviations: SPARE-AD, spatial pattern of abnormality for recognition of early Alzheimer's disease; WMH, white matter hyperintensities; GM, gray matter.*

3.3.2 Clinical, cognitive, biomarker, and APOE-ε4 genotype features

Between-subgroup differences in clinical and cognitive features, Aβ, and APOE-ε4 carrier status were examined separately for each age group. Features that showed consistent trends across >1 age group are summarized in **Figure 3.4** after reanalyzing pooling data. Consistent with the known association of CVD and WMH, subgroup A2 had the highest proportion of participants with CVD risk factors, including hypertension and obesity. Subgroups A2 and A3 showed

similarly higher proportions of smokers and individuals with diabetes than A0 and A1.

Although A2 did not show the most severe atrophy, it had the highest prevalence of APOE-ε4 carriers and, after age 65, the most elevated proportion of cerebral Aβ-positivity (Aβ+). However, trends toward a higher prevalence of APOE-ε4 carriers and higher Aβ+ prevalence in A2 vs. A3 were not statistically significant. Regarding Aβ measures, the only statistically significant difference was the higher prevalence of Aβ+ in A2 compared to A0. These findings suggest that none of the Smile-GAN subgroups were specifically an early AD-related group, but the A2 subgroup had a higher prevalence of AD pathologic change. While participants were selected as not having been diagnosed with cognitive impairment, the A2 and A3 accelerated aging subgroups showed poorer cognitive test performance compared to the A0 and A1 subgroups for ages 55-75. Despite different structural features, A2 and A3 did not differ significantly in cognitive performance across domains. Thus, poorer cognitive performance in A2 and A3 appears to reflect additive effects of atrophy and WMH. Additionally, A3 had the highest proportion of subjects suffering from depression after age 55.

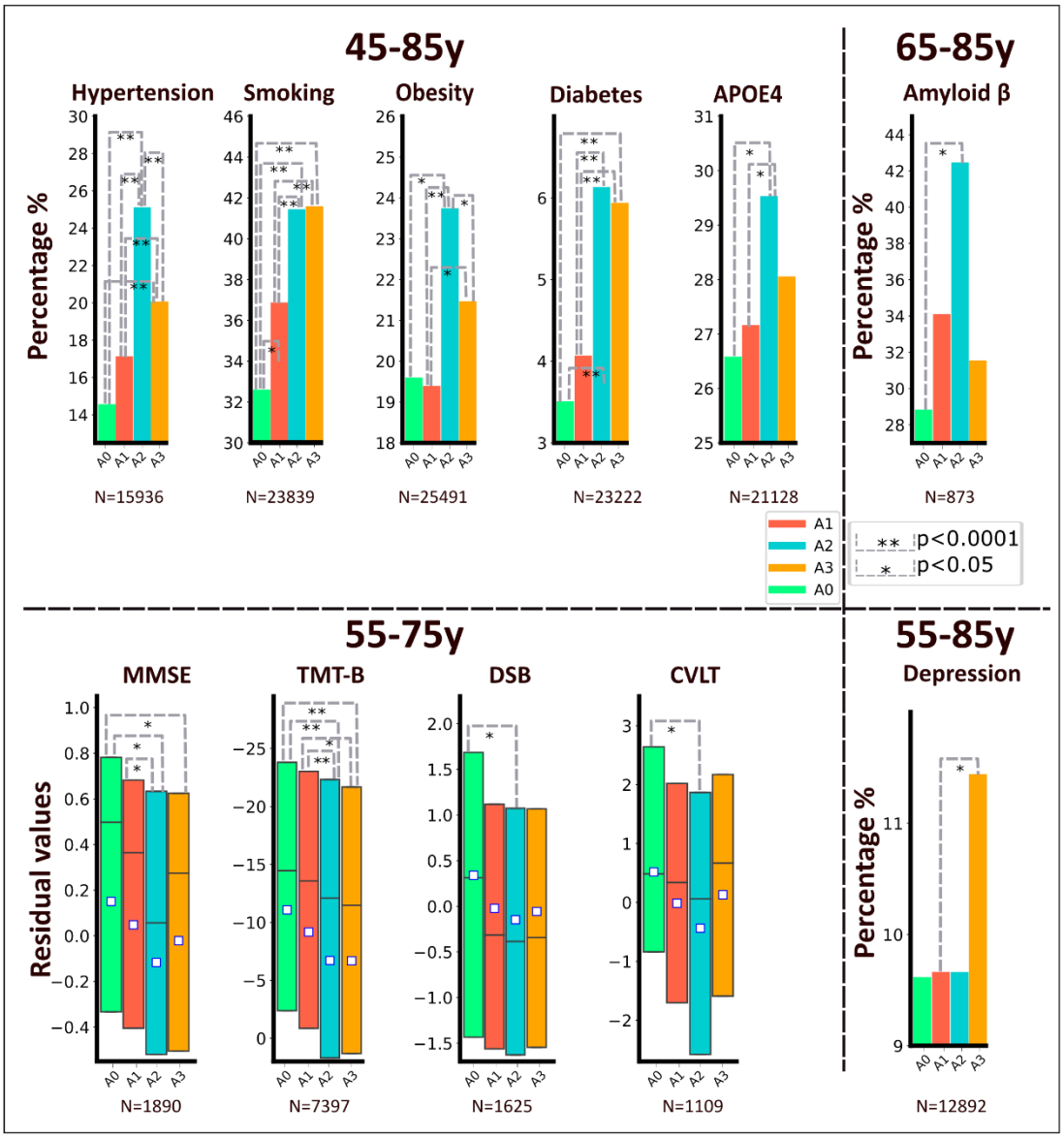


Figure 3.4: Clinical, cognitive, amyloid β, and APOE-ε4 carrier status trends of the brain aging subgroups at baseline. *The plotted features are those non-imaging features that showed consistent trends across more than one age group, presented as a summary after pooling data across age groups. The age ranges above the plots indicate the broader age groups examined. APOE-ε4 carriers were considered those having one or two ε4 alleles. The boxplots show the residuals after adjustment for continuous age, sex, and study (and education for cognitive test scores) for each subgroup. Higher MMSE, DSB, and CVLT indicate better cognition, while lower TMT-B indicates better cognition; TMT scores are presented with an inverted scale, so poorer cognitive performance is in the same direction across the four graphs. The white dot indicates the mean value. The horizontal line shows the median value. The bar plots show the percentage of participants with various risk factors for each subgroup. N indicates the sample size for the graph. FDR correction for multiple comparisons with a p-value threshold of 0.05 was applied. Abbreviations: APOE, Apolipoprotein E; MMSE, mini-mental state examination; TMT-B, trail making test B; DSB, digit span backward; CVLT, California verbal learning test.*

3.3.3 Genome-wide associations of the Smile-GAN probability scores

The Smile-GAN probability scores (A1, A2, and A3) were associated with five, nine, and four genomic loci, respectively. Several loci were previously identified, while others were novel (**Figure 3.5A**). These previously identified loci were associated with various clinical traits, including imaging-derived phenotypes

from white matter microstructure (A1-3)[159], gray matter atrophy (A1-3)[160], WMH (A1-3)[161], CVD risk factors (A1-2)[162], CVD (A1-2)[163], and AD (A1-2)[164] (**Figure 3.5B**). Manhattan plots are shown in **Figure 3.6**.

Several SNPs exerted pleiotropic effects on more than one phenotype/probability with opposite directions of the association effects. For example, A1 (beta=-0.07±0.01; p-value=2.31E-09 and beta=-0.09±0.02; p-value=4.09E-08) and A2 (beta=0.1±0.01; p-value=1.73E-15 and beta=0.13±0.02; p-value=1.04E-15) were associated with the two novel independent variants (rs7209235 and rs55715426 at cytogenetic region:17q25.1), whose mapped genes GALK1 and H3-3B were associated with several CVD biomarkers, including cholesterol[165] and Apolipoprotein B[166]. Therefore, these variants may be protective against CVD for A1 but may serve as risk variants for A2. Furthermore, A1 (beta=0.1±0.02; p-value=6.49E-09) and A2 (beta=-0.09±0.02; p-value=4.05E-07) were both associated with the candidate SNP rs72932727 (cytogenetic position:2q33.2) previously associated with the Alzheimer's disease PRS[164]. Since A2 had the highest prevalence of APOE-ε4 carriers and Aβ+ subjects, opposite to A1, rs72932727 may play a protective role against AD for A1 and may be a risk factor for A2.

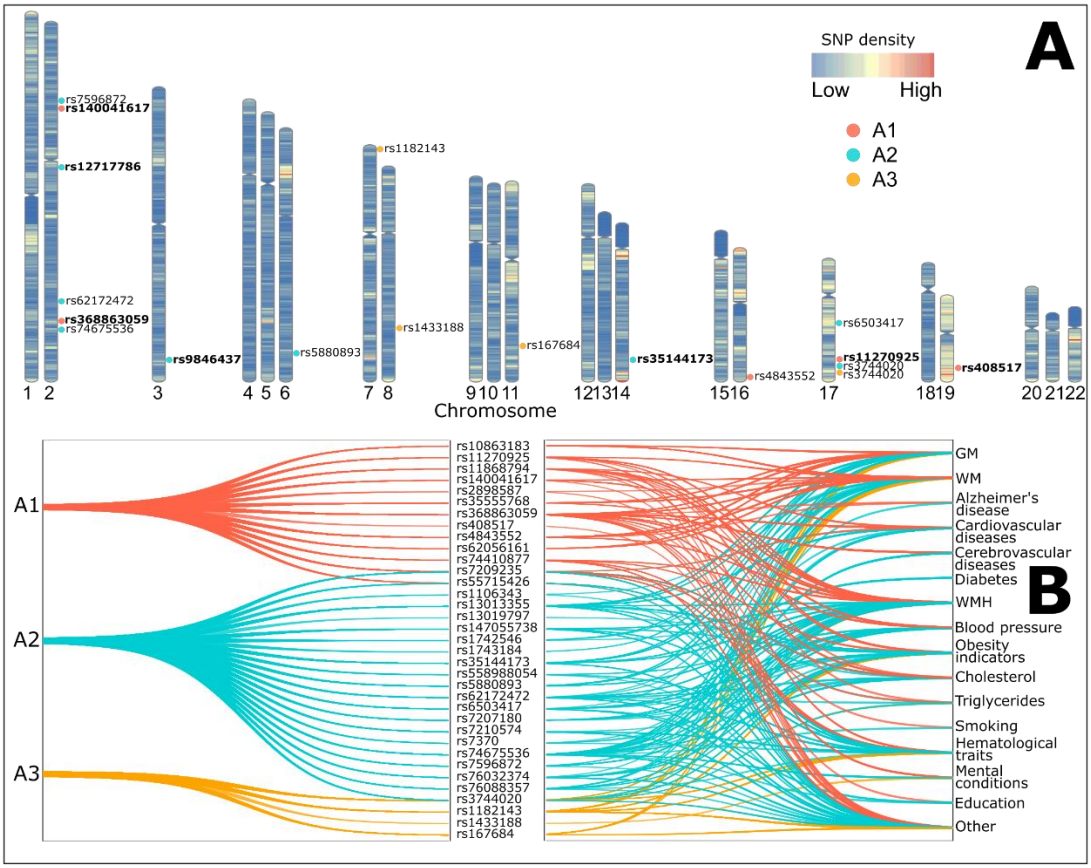


Figure 3.5: Genetic analyses of the Smile-GAN probability scores (A1, A2, and A3). *A) GWAS identified genomic loci (represented by the top lead SNP) associated with the Smile-GAN probability scores (A1, A2, and A3). The genome-wide p-value threshold (5E-08) was used in all GWAS. A locus was denoted as novel (with the lead SNP represented in bold font) if it was not associated with any clinical traits in the GWAS Catalog. The reference genome is Genome Reference Consortium Human Build 37 (GRCh37). The ideogram plot represents all autosomal chromosomes (1 to 22). B) Phenome-wide associations from GWAS Catalog. Independent significant SNPs inside each locus were associated with many clinical traits, which were further classified into high-level groups, including gray matter measures (e.g., (sub)cortical volume, cortical thickness, and surface area), white matter measures (e.g., whole brain restricted isotropic diffusion and whole brain free water diffusion), cardiovascular diseases (e.g., coronary artery disease and myocardial infarction), cerebrovascular diseases (e.g., non-lobar intracerebral hemorrhage and stroke), hematological traits (e.g., platelet, eosinophil, and white blood cell counts), mental conditions (e.g., risk-taking behavior and suicide attempts), etc. In addition, traits such as Alzheimer's disease, WMH, CVD risk factors, and education were also identified. Abbreviations: WM, white matter; WMH, white matter hyperintensities; GM, gray matter.*

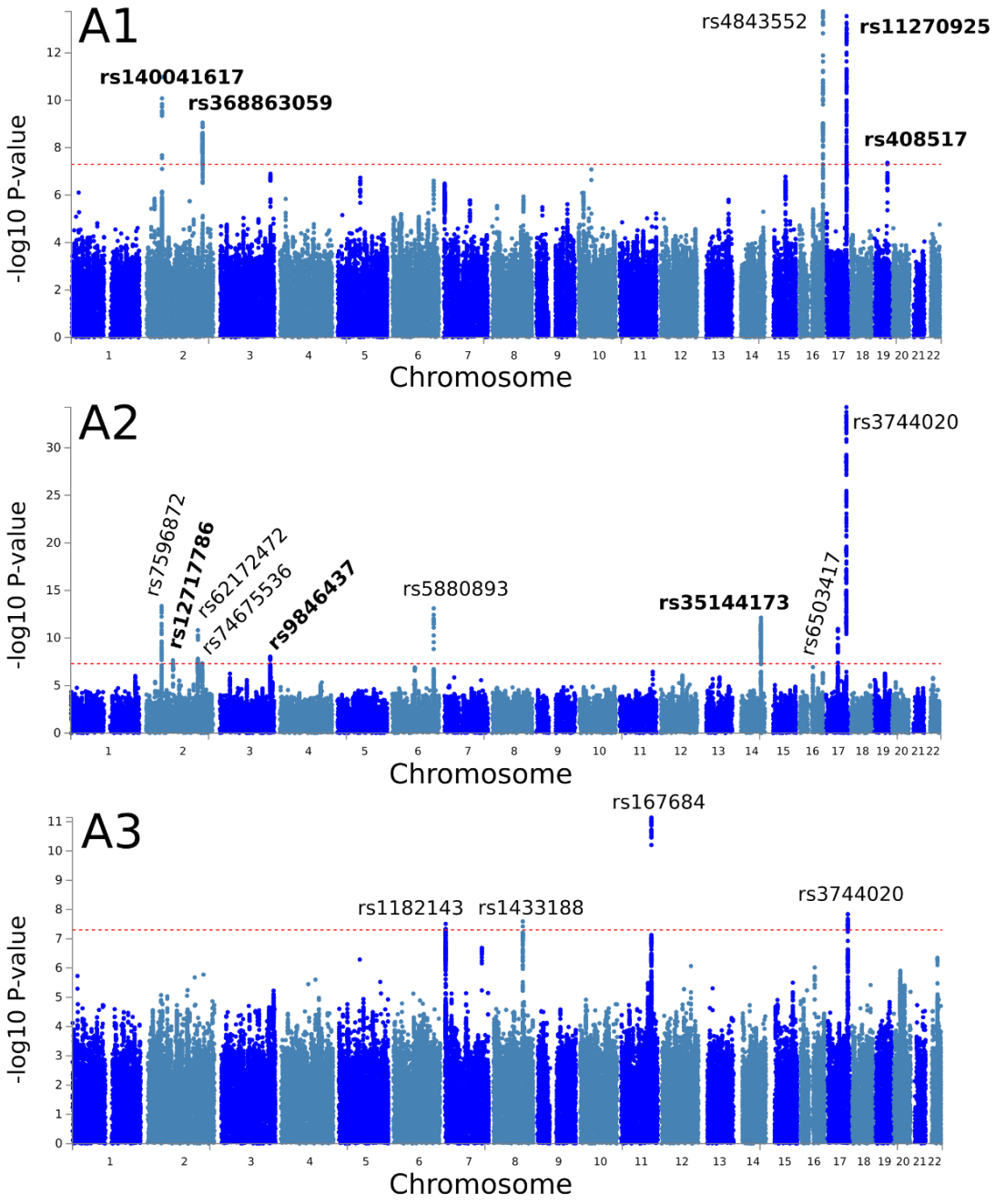


Figure 3.6: Manhattan plots for the GWAS for the Smile-GAN probability scores (A1, A2, and A3).

Moreover, the imaging-derived phenotypes showed highly significant SNP-based heritability estimates (A1: h2=0.44±0.04, A2: h2=0.55±0.04, A3: h2=0.45±0.04; all p-values<E-04). Finally, A3 was significantly associated with both PRS-LLD1 and LLD2 with opposite direction association effects (LLD1: beta=-0.05±0.01; p-value<0.001, LLD2: beta=0.05±0.01; p-value=0.001). A1 was associated with the PRS-LLD1 (beta=0.04±0.02; p-value=0.007), with LLD1 characterized by preserved brain volume (**Table 3.6**).

Table 3.6: Polygenic risk score for the Smile-GAN probability scores (A1, A2, and A3) for Late-Life Depression subtype 1 (LLD1) and 2 (LLD2). *The Polygenic risk score was derived using a Bayesian method (PRS-CS) with the conventional clumping plus thresholding approach in previous work*[137]*. N=7,571. The notation "ns" indicates insignificant associations with p-value>=0.05.*

Smile-GAN probability score	Phenotype	Beta (β)	P-value
A1	LLD1	0.04±0.02	0.007
	LLD2	-0.02±0.02	ns
A2	LLD1	0.01±0.01	ns
	LLD2	-0.02±0.01	ns
A3	LLD1	-0.05±0.01	<0.001
	LLD2	0.05±0.01	0.001

3.3.4 Functional and white matter microstructural associations

Internetwork connectivity analysis revealed that A3 had the most significant differences relative to the reference A0 group. Increased connectivity for several pairs of networks, such as the default mode - motor, somatosensory - occipital visual, and dorsal attention - occipital visual networks, and decreased connectivity for other pairs, such as the default mode - frontotemporal, subcortical - frontotemporal, and occipital visual - fronto-insular-parietal networks are observed (**Figure 3.7**). These results align with the literature showing both increased[19][167] and decreased[16][17] internetwork connectivity, uncovering a complex functional reorganization of the brain with aging.

Regarding FA analysis, consistent with the known associations of WMH and CVD risk factors[168][169][170] with WM integrity, the A2 subgroup showed significant microstructural WM integrity disruption relative to A0 for 41 tracts, with the most prominent disruption observed in posterior thalamic radiation, corona radiata, superior fronto-occipital and longitudinal fasciculus, and anterior limb of the internal capsule (**Figure 3.8**).

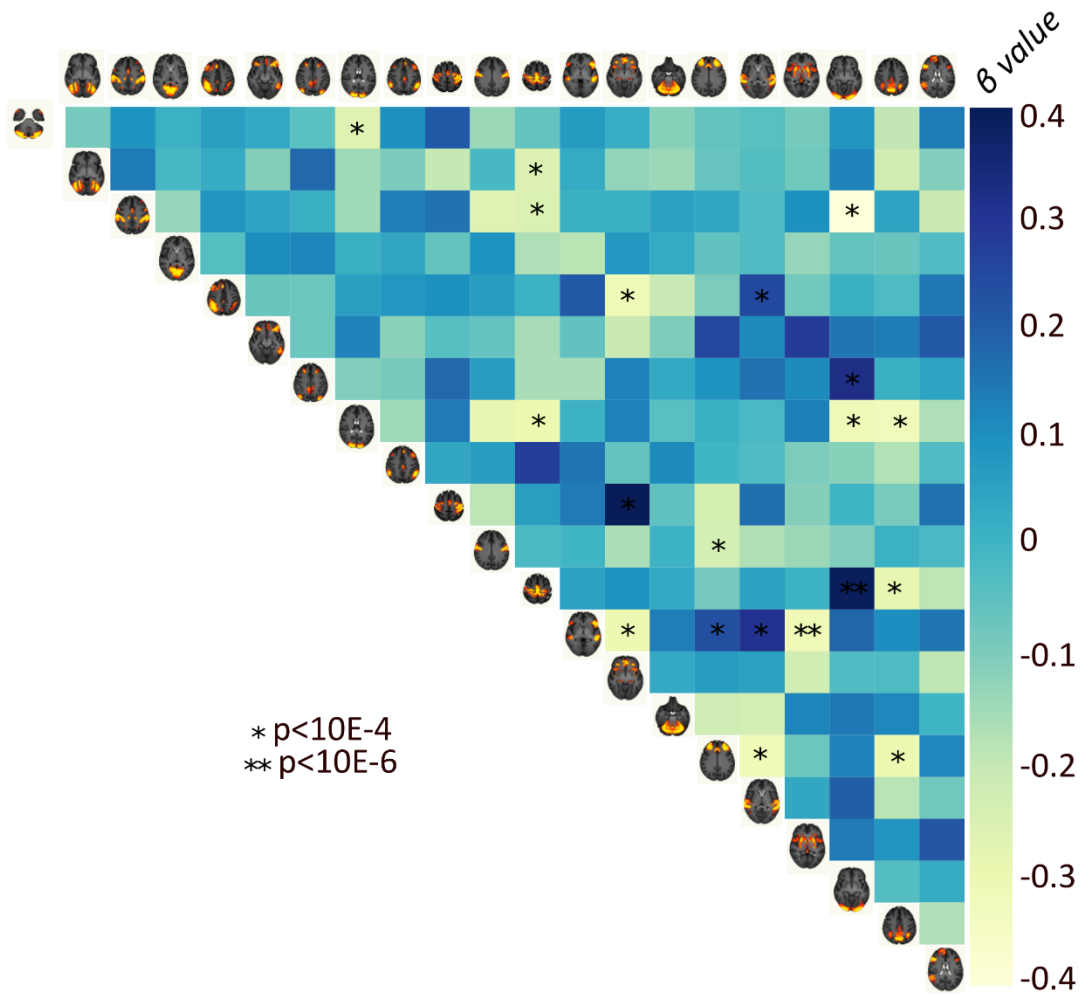


Figure 3.7: Functional connectivity associations of the A3 Smile-GAN subgroup. *Regression coefficients for internetwork connectivity for the A3 subgroup relative to the reference group A0. Bonferroni correction for multiple comparisons adjusted the significance level at 0.05/210.*

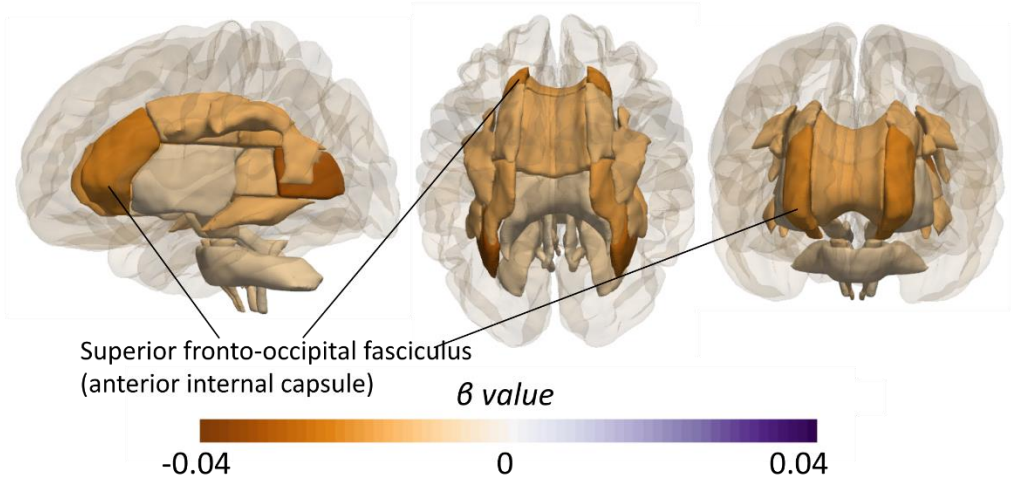


Figure 3.8: White matter microstructural integrity associations of the A2 Smile-GAN subgroup. *Regression coefficients for fractional anisotropy for the A2 vs. A0 group. Bonferroni correction for multiple comparisons adjusted the significance level at 0.05/48.*

3.3.5 Longitudinal outcomes

Across individuals with ≥4 longitudinal MRI scans (N=670, 7.99±4.74 years follow-up, baseline age=69.15±8.93 years), small differences were observed between subgroups in longitudinal atrophy (**Figure 3.9**). Progression of WMH (N=595, 8.08±4.93 years follow-up, baseline age=69.28±8.93) was significantly faster in A2. A2 and A3 subgroups showed the greatest longitudinal cognitive decline (number of individuals with longitudinal cognitive scores was 438 to 1933, mean longitudinal cognitive testing was over 5.04 to 7.99 years with SD 2.60 to 4.74 years across tests, and mean baseline age was over 69.43 to 70.72 years with SD 6.45 to 8.43 years across tests) in agreement with the faster progression from cognitively unimpairment to MCI (**Figure 3.9C**), emphasizing the long-term implications of the baseline MRI subgroups.

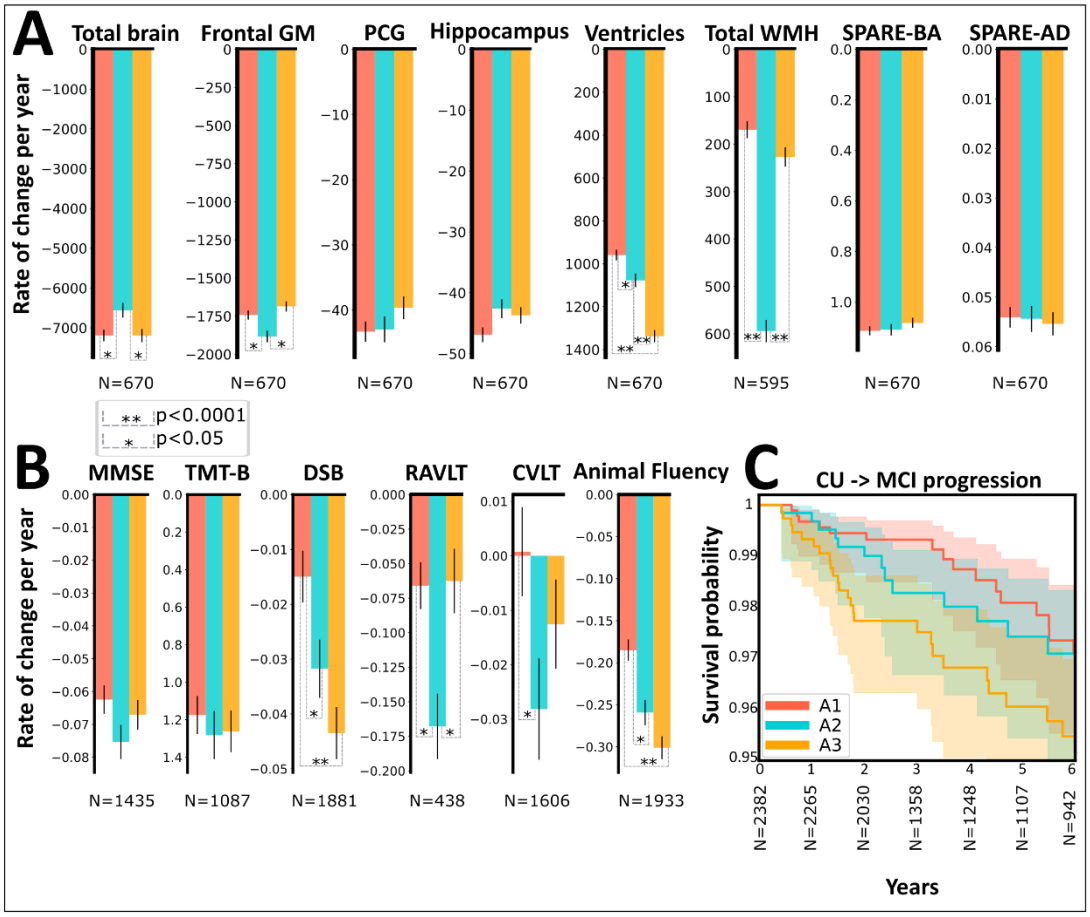


Figure 3.9: Longitudinal outcomes for the Smile-GAN subgroups. *Rate of change per year for A) ROI and WMH volumes (units: mm³), SPARE-BA, SPARE-AD (unitless), and B) cognitive scores calculated using linear mixed-effects models with subject-specific random intercept for the Smile-GAN brain aging subgroups. The models included subgroup indicators, time of visit, and their interaction term while adjusting for baseline age, sex, study, education, and DLICV. Subgroup comparisons of rates of change were conducted using the Wald test. N indicates the number of individuals having ≥4 longitudinal measures for the plotted feature. FDR correction for multiple comparisons was used with a p-value threshold of 0.05. The rate of change for ventricles, SPARE scores, total WMH, and TMT-B is presented with an inverted scale, so faster brain aging (reflected by either more rapid atrophy, lesions accumulation, or cognitive decline) is in the same direction across graphs. C) Kaplan-Meier survival curves show the*

probability of remaining CU and not progressing to MCI for subjects with baseline age within the 65-75 range. N indicates the number of individuals followed up for each time interval. A log-rank test was used to compare the survival curves of the Smile-GAN subgroups. The only significant difference is between A1 and A3 curves (p-value=0.01). Longitudinal results for A0 are not shown because the A0 group was derived using a methodology different from that of the Smile-GAN subgroups. Additionally, the sample size for longitudinal A0 was small, and thus, the results were not robust. Abbreviations: GM, Gray Matter; WMH, white matter hyperintensities; PCG, posterior cingulate gyrus; SPARE-BA, spatial pattern of atrophy for recognition of brain aging; SPARE-AD, spatial pattern of abnormality for recognition of early Alzheimer's disease; MMSE, mini-mental state examination; TMT-B, trail making test B; DSB, digit span backward; RAVLT, Rey auditory verbal learning test; CVLT, California verbal learning test; CU, cognitively unimpaired; MCI, mild cognitive impairment.

3.4 Discussion

Genetics, lifestyle, CVD risk factors, and neuropathologies modify brain aging heterogeneously across individuals even before cognitive symptoms are expressed. Advanced DL methods were applied to a large, diverse, harmonized multi-cohort sample to identify characteristic neuroanatomical subgroups of brain variation. Consistent subgroups of brain aging emerged in each of the decade-long intervals between 45-85 years: A1, or typical aging subgroup with low atrophy and WMH load, and two accelerated aging subgroups, A2 and A3. Heterogeneity in brain aging was observed in the CU population, with stable patterns across decades. These subgroups were detectable from mid-life and associated with cardiometabolic and genetic risk factors, functional connectivity and white matter microstructural measures, and cognition (**Figure 3.10**).

One of the primary findings was the emergence of two accelerated brain aging trajectories, best visualized from the manifold algorithm, which were particularly distinct in ages 65 and older. A2 was associated with hypertension, WMH, disrupted WM microstructural integrity, and vascular disease-associated genetic risk factors, evidenced by the GWAS (**Figure 3.5**), and opposite from the protective effect in A1. This subgroup was also mildly enriched for Aβ+ (ages≥65) and AD-related genetic risk factors, including APOE-ε4. A3 showed widespread GM atrophy and moderate presence of CVD risk factors. Thus, A2 and A3 may have different brain reserve[67], affecting susceptibility to future pathology.

Despite differences in patterns of atrophy, A2 and A3 had comparable poorer cognitive functions than A0. Thus, atrophy and WMH seem to act additively to cause cognitive decline. This may account for lower atrophy in A2 versus A3; further atrophy may predispose to conversion to MCI, resulting in exclusion from this CU cohort. Although underlying pathology related to neuroimaging findings was not explicitly defined, the observed effect is comparable to previous studies demonstrating that the combined involvement of

neurodegenerative and vascular pathology is more pronounced in the earliest stages of cognitive impairment[171][51]. All subgroups had low SPARE-AD scores, indicating no significant AD-related neurodegeneration. Overall, Aβ+ individuals were a minority of cases and relatively evenly distributed across subgroups, suggesting that factors influencing structural brain aging in this CU population may be largely independent of AD before the emergence of symptoms. A3, on the other hand, was not uniquely enriched in a particular studied cardiovascular disease risk factor. It underwent multiple alterations in rsfMRI internetwork connectivity. A3 had the highest prevalence of depression, and the A3 probability score was associated with depression-related PRS. Further investigation of the relationship of A3 to other risk factors is warranted to understand this group better.

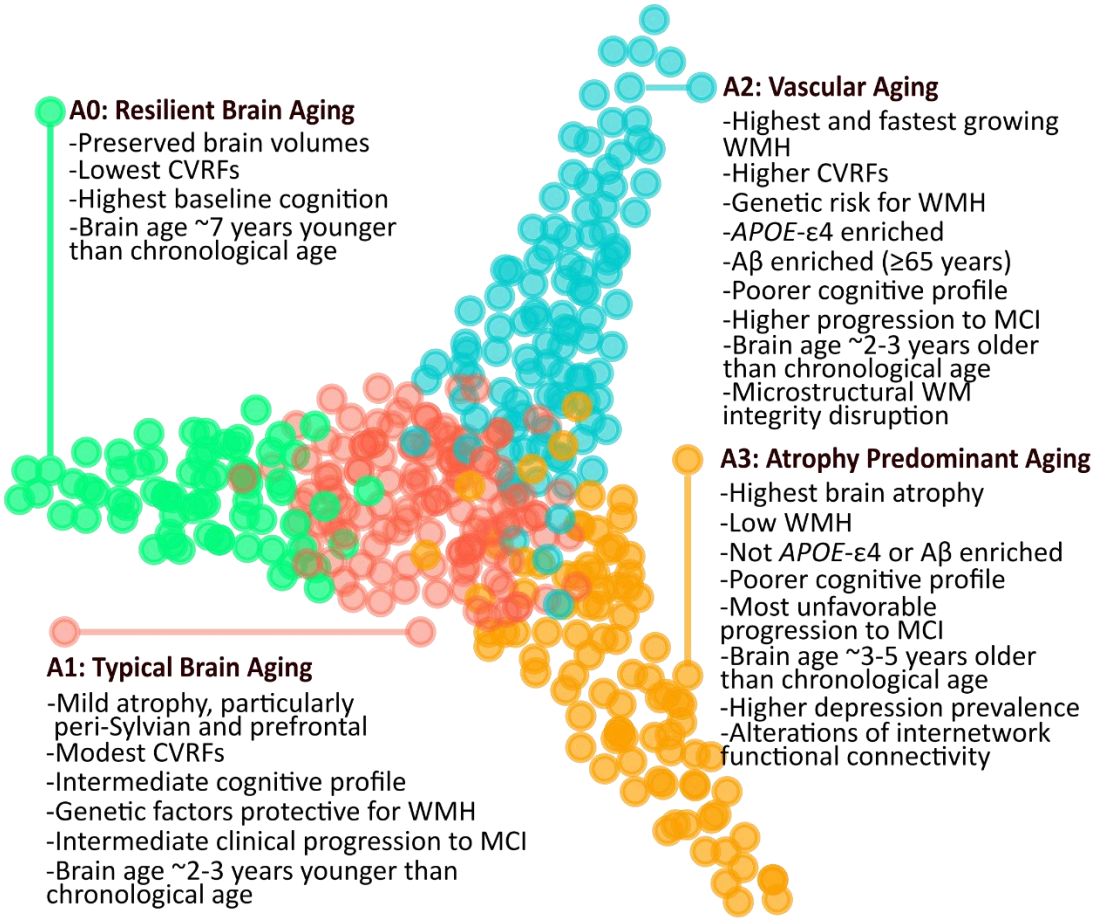


Figure 3.10: Schematic summary of key features of the brain aging subgroups. *Abbreviations: WMH, white matter hyperintensities; WM, white matter; MCI, mild cognitive impairment; APOE, Apolipoprotein E; CVRFs, cardiovascular risk factors; Aβ, Amyloid β.*

This study has several strengths, including the large, diverse, multi-site sample covering a wide age range and the use of advanced harmonization and DL methods. Additionally, identifying multiple SNPs associated with WMH and brain atrophy is consistent with the neuroimaging profile of the subgroups. However, this study also has limitations. (1) The heterogeneity in sampling strategies and data acquisition of each contributing study might impede generalization. (2) There is a low availability of amyloid data and insufficient availability of tau measures and biomarkers related to non-AD neurodegeneration. (3) The lack of long-term follow-up prevents the derivation of robust conclusions regarding the clinical progression and transition to MCI. (4) Regarding sample composition, there is a 'ceiling' effect as people with more severe findings are more likely to be classified as cognitively impaired and thus be excluded from the sample. (5) While morphologic and correlation similarities of subgroups have been observed across decades, equivalence cannot be conclusively established since different models and reference groups (A0) were used per decade, and there was no substantial follow-up across decades. Additionally, A0, despite representing a more resilient population, remains subject to some degree of pathology.

3.5 Conclusion

Consistent and reproducible neuroimaging subgroups defined by regional atrophy and WMH burden were identified across cognitively unimpaired individuals aged 45-85. Two axes of accelerated aging emerged, one showing elevated CVD risk factors, WM integrity disruption, and enrichment of cerebral Aβ and the other with more diffuse and severe atrophy probably driven by lifestyle factors. These subgroups likely reflect differential susceptibility to AD and other neurodegenerative diseases, cognitive decline, and clinical progression.

4 Identification of heterogeneous aging-related brain changes through a coupled cross-sectional and longitudinal non-negative matrix factorization

Currently, heterogeneity techniques have predominantly employed cross-sectional data, thereby overlooking the dynamic progression of the disease patterns. This chapter introduces a novel machine learning framework based on non-negative matrix factorization (NMF) that captures the heterogeneous brain changes associated with aging and related diseases by leveraging cross-sectional and longitudinal data. Unlike traditional methods that rely solely on cross-sectional data and categorize individuals into rigid subtypes, the proposed methodology enables the co-expression of multiple aging patterns driven by integrating population-based brain change maps alongside maps reflecting the dynamic progression of changes observed in individual trajectories. Validated on a semi-synthetic dataset, it is then applied to a large, multi-cohort aging population; distinct components of aging-related atrophy strongly associated with Alzheimer's disease biomarkers, cognitive decline, cardiovascular risk factors, and progression of cognitive impairment are identified. These findings highlight the potential of the proposed methodology to predict clinical progression and customize interventions based on individual neuroanatomical profiles, offering a path toward more personalized therapeutic approaches for neurodegenerative diseases.

4.1 Introduction

So far, semi-supervised clustering techniques[114][108][110] have offered distinct perspectives in addressing the substantial interindividual heterogeneity in brain aging and related disorders by delineating specific patterns or transformations between a reference (e.g., healthy individuals) and a target population (e.g., patients), thus minimizing the influence of disease-unrelated confounders. However, most of these methods model disease heterogeneity as a dichotomous process and extract categorical representation memberships (e.g., subtypes) through hard clustering optimization. Thus, they fail to capture the continuous spectrum along which brain aging and associated pathologies unfold, potentially overlooking the co-occurrence of multiple subtypes within individuals. Recently, a semi-supervised representation learning method via GAN, termed Surreal-GAN (Semi-Supervised Representation Learning via GAN), was developed as an extension of Smile-GAN discussed in the previous chapter,

overcame this limitation by capturing disease-related heterogeneity through continuous low-dimensional representations (R-indices), each reflecting the severity of an independent, relatively homogeneous imaging pattern[172][173]. However, Surreal-GAN uses only cross-sectional data, failing to leverage dynamic observations of the progression of various pathological brain changes.

In recent years, a data-driven approach, NMF[174][175], has gained prominence as a robust technique for analyzing high-dimensional data across several fields. The core concept of NMF is to decompose a given non-negative matrix V into the product of two lower-dimensional, non-negative matrices W and H, such that:

$$V \approx WH$$

where V of size mxn is the original data matrix, W of size mxk is a matrix containing the basis vectors/components/dictionary, H of size kxn is a matrix containing the coefficients that indicate the contribution of each component in approximating the original data, and k is the number of estimated components.

NMF produces a decomposition in which both the components and the associated coefficients are constrained to be non-negative. This decomposition is typically obtained by solving an energy minimization problem:

$$\min_{W,H} \|V - WH\|_F^2, \text{ subject to } W \geq 0, H \geq 0$$

So, the optimal non-negative matrices are those that most accurately reconstruct the data matrix while adhering to the non-negativity constraints. Since the problem is non-convex, leading to multiple local minima, iterative algorithms, such as the multiplicative update rules[176] and the alternating least squares[177], are widely employed to solve it.

The non-negativity constraint differentiates NMF from other matrix factorization methods, such as the PCA and ICA, and produces matrices with distinct properties. First, the non-negativity ensures that the factors produced consist only of positive values, which aligns with many real-world datasets where negative values are not meaningful—such as in images, document-word matrices, or gene expression data. This constraint enables NMF to generate a parts-based, additive representation of the data, meaning that each data point is expressed as a combination of distinct parts, making it highly interpretable. For example, in image processing, NMF might decompose an image into parts

like edges or textures, rather than combinations of negative and positive features, as in other methods like PCA. Another result of the non-negativity constraint is sparsity in the factor matrices. Since NMF restricts the values to be non-negative, many entries in the matrices tend to be zero, leading to sparse representations. This sparsity not only enhances interpretability but also highlights the most relevant features of the data, making NMF particularly useful for feature selection and pattern discovery.

NMF has applications in several fields like natural language processing[178][179], image analysis[180][181][182], bioinformatics[183][184], and, more recently, neuroimaging[185][186][187]. In neuroimaging, NMF has shown an ability to segment the brain into functionally and structurally relevant components. By capturing patterns of covariation across individuals using cross-sectional data, NMF enables the modeling of complex brain changes associated with development, aging, and disease, complementary to methods such as LDA and SuStaIn. Despite these advances, the potential of NMF to capture patterns of pathological brain changes, particularly through the integration of longitudinal studies that track the dynamics of pathological progression, remains largely untapped.

This chapter introduces a novel methodology termed Coupled Cross-sectional and Longitudinal NMF (CCL-NMF). CCL-NMF aims to identify heterogeneous patterns of brain changes simultaneously from cross-sectional and longitudinal data using a joint optimization formulation.

While large cross-sectional datasets are accessible and can capture cumulative brain changes due to aging or disease over long periods, they have well-known limitations, such as secular effects and the absence of personalized baselines for comparison. Consequently, the effects of aging or disease are inferred from broader population-level comparisons. In contrast, longitudinal data provide a direct, individualized perspective on brain changes over time, yielding insights into the dynamic progression of neurobiological processes. However, longitudinal datasets are scarcer. The current study develops a mutually constrained NMF framework to delineate components (i.e. brain aging patterns) that encapsulate distinct brain alteration patterns derived jointly from cross-sectional and longitudinal data, each having independent, different sample sizes. Notably, CCL-NMF avoids rigid classification into mutually exclusive categorical subtypes, allowing individuals to exhibit varying degrees of co-expression across multiple patterns, which is important for capturing co-existing pathologies. The proposed methodology is formulated in a general framework, enabling its application to analyses of heterogeneity of any disease characterized by monotonic brain alterations (e.g., gradual brain atrophy or

white matter hyperintensity accumulation as measured by MRI, or increasing deposition of neuropathologies such as amyloid and tau as measured by positron emission tomography).

The following section introduces a preliminary model, termed Cross-sectional and Longitudinal NMF (CL-NMF), to distinguish it from the final refined CCL-NMF model. CL-NMF is applied to investigate GM atrophy heterogeneity in an aging population from the BLSA study. The model is then refined to its final formulation, CCL-NMF, validated using a semi-synthetic dataset, and applied to a large, multi-cohort aging population from the iSTAGING dataset.

4.2 Preliminary model development and application to the BLSA aging cohort.

Here, a preliminary model termed Cross-sectional and Longitudinal NMF (CL-NMF) is proposed to explore the heterogeneity of brain aging by integrating both cross-sectional and longitudinal information. This model aims to unravel the complex and diverse brain aging patterns by accounting for population-based brain alterations and the dynamic progression of changes captured through individual trajectories. Importantly, this approach enables the assessment of individualized expression levels across these components, thus providing additional valuable tools for personalized patient management and enhancing the potential for more precise stratification in clinical trials. Application of the proposed method to structural MRI data of an aging population from the BLSA study identified succinct, reproducible, and clinically informative brain aging components.

4.2.1 Methods

4.2.1.1 Model

CL-NMF addresses brain aging heterogeneity by decomposing brain changes into distinct components using NMF. These components aim at capturing coordinated brain changes that might be associated with underlying neuropathologic processes. A mutually constrained NMF framework is introduced, integrating two complementary sources of information: maps representing cross-sectional and longitudinal brain changes (C-map and L-map, respectively). The C-map captures aging effects over decades at a population level. The L-map captures the dynamic patterns of brain change over time on an individual basis. The joint NMF approach identifies components shared by cross-sectional and longitudinal maps based on the assumption that an aging or disease effect estimated cross-sectionally at a population level should be compatible with dynamic brain changes captured by longitudinal data. The joint

NMF also estimates corresponding coefficients (loadings), representing the degree of expression of each component from each individual by optimizing the reconstruction of both types of maps, thereby capturing the complex interplay between static and dynamic aspects of brain alterations.

To derive the C-map, a method is used to estimate the normative brain variability, specifically from a middle-aged cohort without reported pathologies. The assumption is that this population is relatively unaffected by neuropathologic processes that typically emerge beyond age 50. Deviations from normality are quantified by projecting an aging population onto the estimated normative space. The aging population is conceptualized as having emerged from a similar healthy middle-aged population but with alterations due to the gradual accrual of brain atrophy linked to normal aging, various neuropathologies, and the cumulative effects of genetic, lifestyle, and environmental factors. These factors induce complex and multifactorial brain changes, modeled as deviations from normality. The heterogeneity of these deviations, emerging from the projection of aging individuals onto the healthy middle-aged population space, is then summarized by the subsequent NMF into dominant patterns of brain alteration. The L-map, which captures longitudinal brain changes, is derived using a statistical model that estimates the rate of atrophy, as detailed in the following sections. The hypothesis is that longitudinal data directly reflects brain changes associated with the underlying neuropathologic processes present on an individual basis. By integrating the cross-sectionally-derived population-based brain changes with the dynamic progression of brain changes captured by individual trajectories, this method provides a novel decomposition that will offer greater power for measuring aging and disease-related effects.

PCA for estimating the cross-sectional deviation map (C-map)

Given **S1**, the reference population, and **S2**, the target population, the C-map includes the deviations of **S2** from the normative space estimated from **S1** using PCA. Projection of the **S2** population in this normative space provides estimates of the deviation from the reference/normative anatomy. So, for subject i, the deviation vector $\mathbf{d}_i$ is:

$$\mathbf{d}_i = \mathbf{x}_i - \int_{j=1}^{J} b_{ji}\, \mathbf{v}_j \quad (1)$$

where $\mathbf{x}_i$ is the actual anatomy vector and $\mathbf{v}_j$ are the principal components spanning the normative brain variability calculated in **S1**. The number of principal components selected is adequate to explain 95% of the variance.

LME for estimating the longitudinal change map (L-map)

Linear mixed-effects model[146] is used to estimate the longitudinal rate of change map (L-map) for individuals with multiple measurements over time. For feature j, subject i and timepoint t, the model is specified as follows:

$$Y_{i,j,t} = \beta_{0_j} + \beta_{1_j}\mathrm{Time}_{i,j,t} + \beta_{2_j}X_{i,j} + \gamma_{0_{i,j}} + \gamma_{1_{i,j}}\mathrm{Time}_{i,j,t} + \varepsilon_{i,j,t} \quad (2)$$

where $X_{i,j}$ is a covariate matrix for the fixed effects (e.g., site, baseline age, etc). β_{0_j} β_{1_j}, and β_{2_j} are shared across all subjects, while the errors $\varepsilon_{i,j,t}$ are independent and identically distributed with a mean of zero. The subject-specific random intercept and slope parameters, $\gamma_{0_{i,j}}$ and $\gamma_{1_{i,j}}$, are assumed to follow a bivariate normal distribution.

The calculation of each subject's rate of change involves two components: the population-average slope β_{1_j} from the fixed-effects term and the subject-specific random slope $\gamma_{1_{i,j}}$ for subject i. So, the final rate of change is given by $\beta_{1_j} + \gamma_{1_{i,j}}$. The resulting L-map is structured with dimensions corresponding to the number of brain regions by the number of individuals with longitudinal measurements, capturing the regional rates of change. Since the focus is on GM atrophy rates, positive values —few in number and small in magnitude— were set to zero. To ensure compatibility with the non-negativity constraints required for input data in NMF, the sign of the L-map was inverted, yielding solely non-negative values.

Joint NMF

After extracting the C-map and L-map, the joint NMF implementation is carried out. For X_C of size DxN_C and X_L of size DxN_L, where D represents the dimensionality of brain features, and N_C (N_L) denotes the number of subjects with cross-sectional deviations (longitudinal change rates), the objective is to extract K components that encapsulate the brain aging patterns using an NMF scheme. This approach operates under the hypothesis that both data types share the same components (dictionary matrix W of size DxK). The shared dictionary ensures that cross-sectional components of brain aging are consistent with the dynamic progression patterns captured by longitudinal measurements. However, cross-sectional and longitudinal measures have distinct loading coefficients: H_C of size KxN_C for cross-sectional data and H_L of size KxN_L for longitudinal data. The model can be expressed as:

$$X_C \approx WH_C\,, X_L \approx WH_L\,, \text{subject to } W > 0, H_C > 0, H_L > 0 \quad (3)$$

The loss function is:

$$L = \|X_C - WH_C\|_F^2 + \|X_L - WH_L\|_F^2 \quad (4)$$

This is a mutually constrained dual factorization of cross-sectional and longitudinal data, optimized using a multiplicative update rule[188],[176].

$$w_{ij} \leftarrow w_{ij} \frac{(\sum_I X_I H_I{}^T)_{ij}}{(\sum_I W X_I H_I{}^T)_{ij}} \quad (5)$$

$$h_{ij}^I \leftarrow h_{ij}^I \frac{(W^T X_I)_{ij}}{(W^T W H_I)_{ij}}\,, (I = C, L) \quad (6)$$

Prior to NMF, (non-zero elements of) X_C and X_L are rescaled using MinMax scaling to the range [0,1] to ensure uniform feature scaling. Initialization of W, $H_C, \text{and } H_L$ is performed with random values sampled from a uniform distribution [0, 0.5]. The initial matrices are normalized using a diagonal matrix S, derived from the l2-norms of the columns of W_{init}:

$$W_{init}{}' = W_{init} S^{-1},\ H_{C_{init}}{}' = S H_{C_{init}},\ H_{L_{init}}{}' = S H_{L_{init}} \quad (7)$$

This normalization step is repeated at each iteration to maintain numerical stability and ensure convergence. Finally, the K selection is made based on the reproducibility index, sparsity, and reconstruction error (Section 4.2.1.3).

4.2.1.2 Aging dataset

Next, the method was applied to parse the heterogeneity of aging-related gray matter atrophy in an aging population. T1-weighted MRI data from the BLSA study were used. All participants gave written informed consent to the study for data acquisition and analyses according to the Declaration of Helsinki. The institutional review board of the University of Pennsylvania approved this project. T1-weighted image intensity inhomogeneity was corrected[125], followed by multi-atlas skull-stripping[126]. 114 GM ROIs were segmented using a multi-atlas, multi-warp label fusion-based method[127] (**Table 4.1**). The **S1** group was defined as CU subjects without comorbidities (e.g., hypertension, diabetes, smoking, and obesity) and younger than 55 years (N=105; 41% males; 44±8 years); all other older or equal to 55 years participants were treated as the **S2** group (N=957; 48% males; 69±12 years; CU:934, MCI:12, AD:8, other neurodegenerative disease:3).

The **S1** and **S2** ROI volumes were residualized to rule out the effect of sex and DLICV estimated in the **S1** group using linear regression. Subsequently, adjusted volumes were standardized with respect to the **S1** group. A total of 54 principal components accounted for 95% of the variance in **S1** gray matter. The Euclidean distance from the center of the PC distribution was calculated to assess the typicality or normality of individuals. Subjects in the **S2** group with an Euclidean distance exceeding the 70th percentile of the **S1** distance distribution, designated as **S3**, were used in the NMF. The normality threshold was empirically set at the 70th percentile of the Euclidean distance distribution to ensure an adequate sample size for **S3** (N=392; 69% males; 72±12 years; CU:373, MCI:10, AD:8, other neurodegenerative disease:1).

Table 4.1: 114 anatomic gray matter (GM) regions of interest (ROIs).

Anatomic GM ROIs	
Accumbens area (R)	Triangular part of the inferior frontal gyrus (R)
Accumbens area (L)	Triangular part of the inferior frontal gyrus (L)
Amygdala (R)	Middle occipital gyrus (R)
Amygdala (L)	Middle occipital gyrus (L)
Caudate (R)	Medial orbital gyrus (R)
Caudate (L)	Medial orbital gyrus (L)
Hippocampus (R)	Postcentral gyrus medial segment (R)
Hippocampus (L)	Postcentral gyrus medial segment (L)
Pallidum (R)	Precentral gyrus medial segment (R)
Pallidum (L)	Precentral gyrus medial segment (L)
Putamen (R)	Superior frontal gyrus medial segment (R)
Putamen (L)	Superior frontal gyrus medial segment (L)
Thalamus (R)	Middle temporal gyrus (R)
Thalamus (L)	Middle temporal gyrus (L)
Basal forebrain (L)	Occipital pole (R)
Basal forebrain (R)	Occipital pole (L)
Anterior cingulate gyrus (R)	Occipital fusiform gyrus (R)
Anterior cingulate gyrus (L)	Occipital fusiform gyrus (L)
Anterior insula (R)	Opercular part of the inferior frontal gyrus (R)
Anterior insula (L)	Opercular part of the inferior frontal gyrus (L)
Anterior orbital gyrus (R)	Orbital part of the inferior frontal gyrus (R)
Anterior orbital gyrus (L)	Orbital part of the inferior frontal gyrus (L)
Angular gyrus (R)	Posterior cingulate gyrus (R)
Angular gyrus (L)	Posterior cingulate gyrus (L)

Calcarine cortex (R)	Precuneus (R)
Calcarine cortex (L)	Precuneus (L)
Central operculum (R)	Parahippocampal gyrus (R)
Central operculum (L)	Parahippocampal gyrus (L)
Cuneus (R)	Posterior insula (R)
Cuneus (L)	Posterior insula (L)
Entorhinal area (R)	Parietal operculum (R)
Entorhinal area (L)	Parietal operculum (L)
Frontal operculum (R)	Postcentral gyrus (R)
Frontal operculum (L)	Postcentral gyrus (L)
Frontal pole (R)	Posterior orbital gyrus (R)
Frontal pole (L)	Posterior orbital gyrus (L)
Fusiform gyrus (R)	Planum polare (R)
Fusiform gyrus (L)	Planum polare (L)
Gyrus rectus (R)	Precentral gyrus (R)
Gyrus rectus (L)	Precentral gyrus (L)
Inferior occipital gyrus (R)	Planum temporale (R)
Inferior occipital gyrus (L)	Planum temporale (L)
Inferior temporal gyrus (R)	Subcallosal area (R)
Inferior temporal gyrus (L)	Subcallosal area (L)
Lingual gyrus (R)	Superior frontal gyrus (R)
Lingual gyrus (L)	Superior frontal gyrus (L)
Superior temporal gyrus (R)	Supplementary motor cortex (R)
Superior temporal gyrus (L)	Supplementary motor cortex (L)
Temporal pole (R)	Supramarginal gyrus (R)
Temporal pole (L)	Supramarginal gyrus (L)
Transverse temporal gyrus (R)	Superior occipital gyrus (R)
Transverse temporal gyrus (L)	Superior occipital gyrus (L)
Lateral orbital gyrus (R)	Superior parietal lobule (R)
Lateral orbital gyrus (L)	Superior parietal lobule (L)
Middle cingulate gyrus (R)	Medial frontal cortex (R)
Middle cingulate gyrus (L)	Medial frontal cortex (L)
Middle frontal gyrus (R)	
Middle frontal gyrus (L)	

Longitudinal trajectories of GM volumetric change were estimated using LME models after accounting for baseline DLICV, baseline age, sex, and site

covariates. 281 subjects (67.6% males; baseline age=73.03±10.23 years; baseline diagnosis: CU:274, MCI:5, AD:2) had longitudinal scans. The follow-up time was 5.79±4.3 years.

4.2.1.3 Metrics for optimal CL-NMF K selection

To determine the optimal number of CL-NMF components, the reproducibility of the solutions was assessed in a split-sample context and evaluated their data-fitting capability. This approach parallels model selection techniques extensively utilized in clustering research. The fundamental premise is that inferring either too few or too many components can lead to instability: excessive components might capture random data variations, whereas insufficient components may amalgamate distinct patterns due to the limited expressiveness of the model.

The reproducibility analysis was conducted by dividing the dataset into two halves with comparable age and sex distributions, examining how the reproducibility of the solution changes with the number of components estimated. The reproducibility was quantified by measuring the overlap between independently estimated components from the two splits, which were matched using the Hungarian algorithm[189]. The overlap, measured by the inner product, termed 'reproducibility index', ranges from 0 to 1, with higher values indicating greater reproducibility.

While achieving a stable solution is crucial, ensuring the solution fits the data well is equally important. Therefore, the reproducibility analysis was complemented by investigating how the component sparsity changes with the number of components. The sparsity of the derived components was evaluated using Hoyer's[190] sparsity measure. Higher sparsity is advantageous as it enhances the model's interpretability and generalizability.

4.2.1.4 Statistical analysis

The associations of the BLSA CL-NMF components with (1) the total WMH volume, (2) the SPARE-BA, (3) the SPARE-AD, and (4) the MMSE were examined. To this end, linear regression modeling was performed, adjusting for age, sex, and DLICV (and education for the MMSE). The most predictive CL-NMF component was progressively added to the model, and the changes in adjusted R^2 were assessed. The adjusted R^2 accounts for the number of predictors in the model and penalizes for excessive variables.

To evaluate the associations between the CL-NMF components and future progression from CU to MCI, a Cox proportional hazards model was employed while adjusting for age, sex, baseline DLICV, and education. The hazard ratio

(HR) was calculated and reported as the effect size measure, indicating the extent to which each CL-NMF component affects the risk of progression of cognitive impairment. The follow-up time was 7.08±4.74 years.

4.2.2 Results

Split-half reproducibility index and sparsity metrics were used to examine the optimal number of CL-NMF components, ranging from 2 to 10 (**Figure 4.1**). The split-half reproducibility analysis demonstrated that reproducibility decreases overall as the number of components increases. However, there is a plateau from K=4 to 5. Since K=5 results in sparser components, it was selected for subsequent analysis

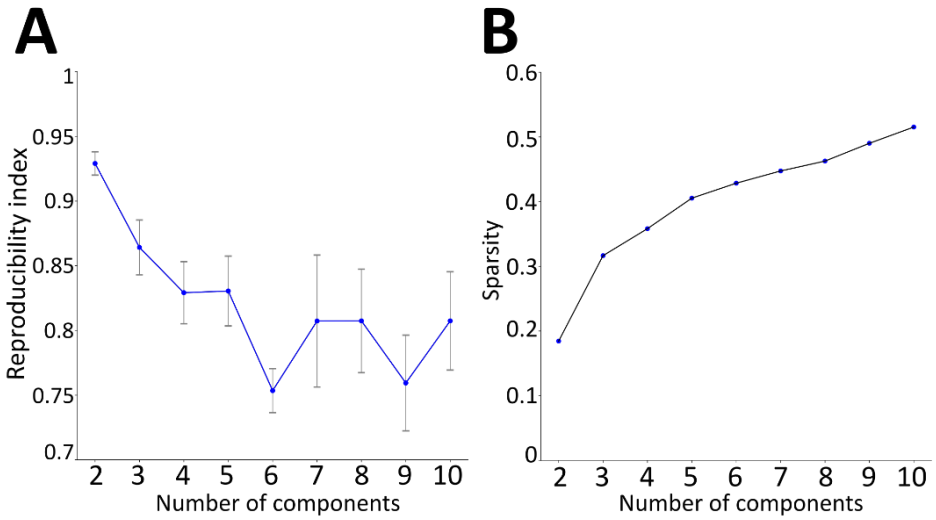


Figure 4.1: A) Split-half reproducibility index and B) sparsity reported as functions of the number of CL-NMF components (N_C=392, N_L=281).

CL-NMF identified five distinct components of aging

Figure 4.2 shows the five identified brain aging components.

-**CL-NMF1** captures diffuse brain atrophy, particularly in peri-Sylvian and prefrontal cortex.

-**CL-NMF2** reflects posterior cortical brain atrophy.

-**CL-NMF3** represents basal ganglia atrophy.

-**CL-NMF4** mainly reflects medial temporal lobe atrophy.

-**CL-NMF5** captures atrophy primarily in orbitofrontal, precentral, and posterior cingulate gyri.

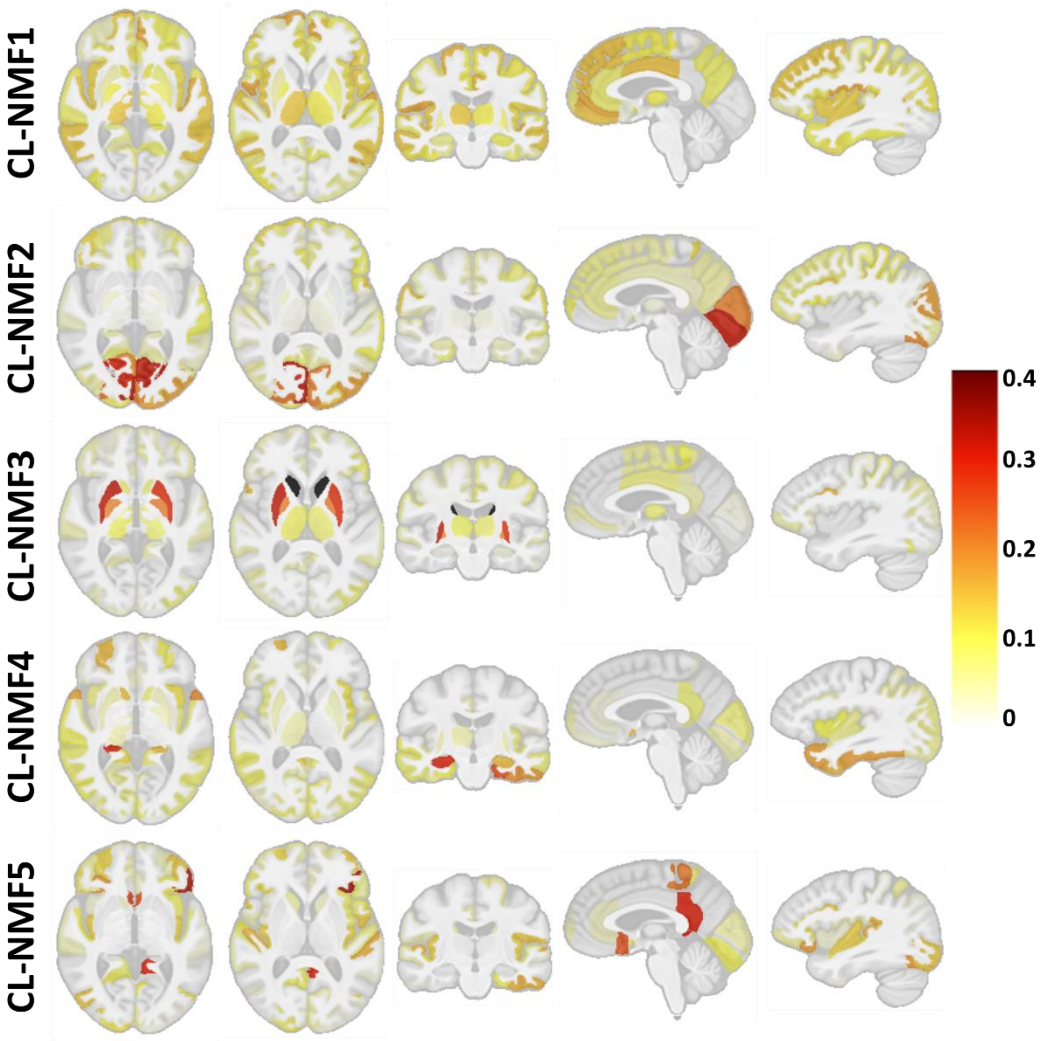


Figure 4.2: CL-NMF dictionary in brain maps format for K=5 (N_C=392, N_L=281). *Red (white) colors indicate higher (lower) contribution of the visualized brain region in the CL-NMF component.*

Associations of aging components with WMH, SPARE scores, MMSE, and future risk of progression to MCI

Moreover, the associations of the BLSA CL-NMF components with the total WMH volume, SPARE scores, and MMSE were examined (**Figure 4.3A**). The analysis specifically investigated whether including CL-NMF components improved the baseline regression model, which included only demographics (age, sex, and DLICV). For total WMH volume and SPARE scores, the CL-NMF components increased the adjusted R^2, indicating that the components explained a substantial amount of variance and improved the models. Specifically, for total WMH volume, the adjusted R^2 rose from 0.29 to 0.39, with the **CL-NMF1** marked by peri-Sylvian atrophy being the most valuable predictor in agreement with previous study[172]. For SPARE-BA, the adjusted R^2 slightly increased from 0.68 to 0.75, and the **CL-NMF1** mainly drove this increase. Chronological age is highly associated with SPARE-BA, thus being the most useful predictor. SPARE-AD adjusted R^2 substantially increased from 0.23

to 0.49; **CL-NMF4**, characterized by medial temporal lobe atrophy, was primarily responsible for this jump. Finally, in the MMSE plot, the adjusted R^2 initially remained stable or decreased with the inclusion of **CL-NMF1**-**3** and **CL-NMF5**, and only including the **CL-NMF4** improved the model by raising the adjusted R^2 from 0.057 (with demographics as predictors) to 0.092 (including demographics and all CL-NMF components). Literature has widely demonstrated the associations between medial temporal lobe atrophy and cognitive decline and Alzheimer's disease[191][192][193][194][195].

Finally, the predictive power of the BLSA CL-NMF components for future progression from CU to MCI was examined. The Cox proportional hazards model was utilized to test the associations between CL-NMF components and the risk of progression while adjusting for age, sex, baseline DLICV, and education. **CL-NMF4** ($p=0.04$, log(HR) (95% CI) = 0.93 (0.02, 1.84)) was the only component significantly associated with the risk of MCI progression in agreement with previous studies[196][197][198] (**Figure 4.3B**).

To sum up, section 4.2 introduced a method for dissecting the heterogeneity observed in brain aging, effectively capturing this variation in low-dimensional components that remain consistent across both cross-sectional and longitudinal trajectories of brain change. The method was applied to analyze GM atrophy heterogeneity in an aging population using as a reference a healthy middle-aged cohort both drawn from the BLSA study. Five distinct and reproducible aging-related pathological components were identified in a population predominantly composed of cognitively normal individuals. **CL-NMF1** showed a significant association with white matter lesion volume, suggesting the potential presence of vascular pathology. **CL-NMF4** exhibited features characteristic of Alzheimer's disease, aligning with the SPARE-AD index and MMSE, and emerged as a key predictor for the clinical progression from CU to MCI. A more concrete exploration of the proposed methodology, as well as its application to a larger and more diverse cohort, are discussed in the next section.

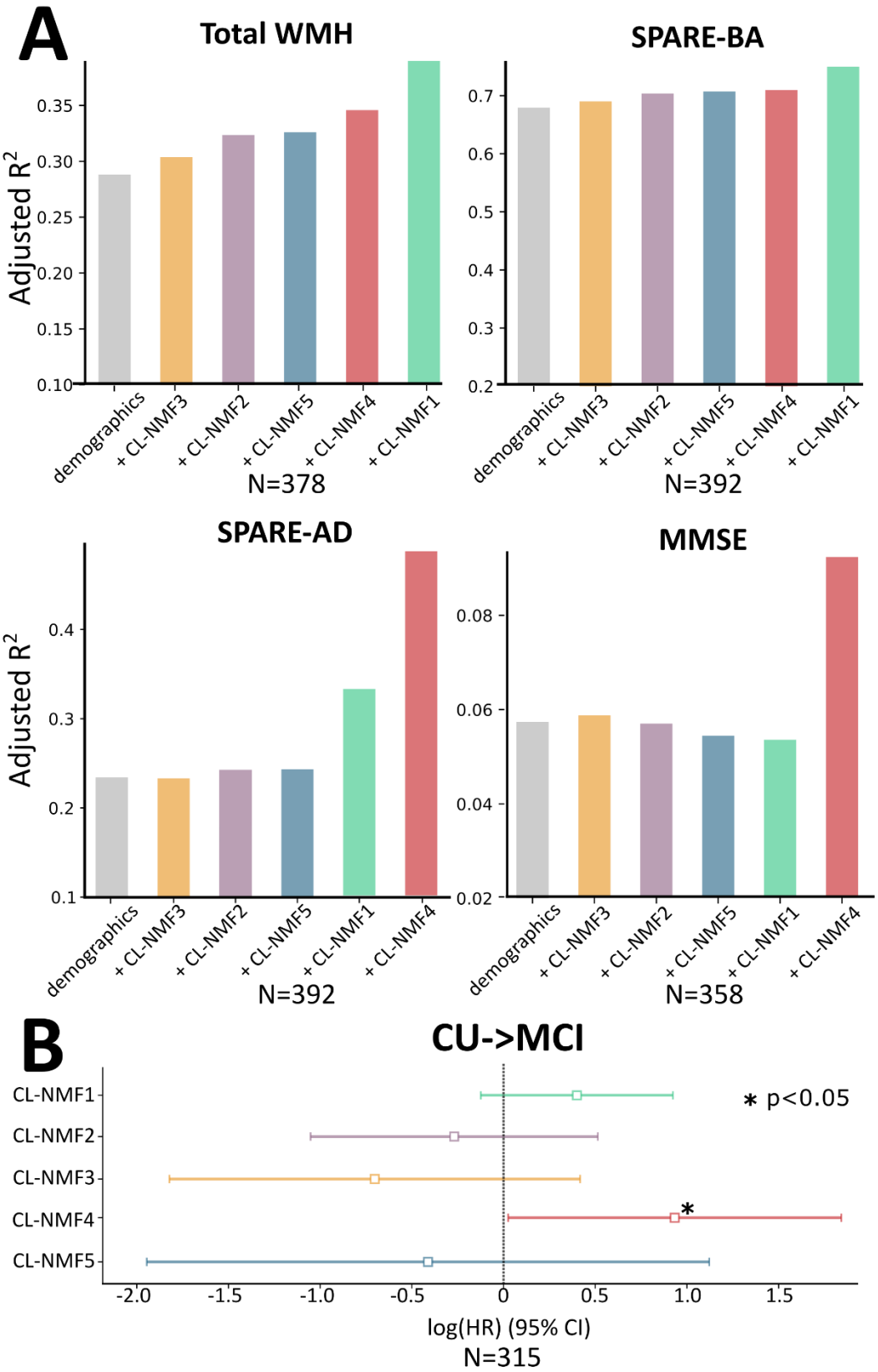


Figure 4.3: Associations of cross-sectional CL-NMF loadings with cognition, biomarkers, clinical features, and future risk of progression to MCI. *A) Adjusted R2 as a function of predictors for total WMH volume, SPARE-BA, SPARE-AD, and MMSE. Predictors start from demographics (age, sex, DLICV for all cases; education was also included in demographics for MMSE) and gradually incorporate the CL-NMF loadings. B) Longitudinal CU to MCI progression. N indicates the sample size for each graph. Abbreviations: WMH, white matter hyperintensities; MMSE, mini-mental state examination.*

4.3 Refined model development, validation with a semi-synthetic dataset, and application to the iSTAGING aging population.

A novel machine learning-based approach designed to disentangle the heterogeneity of brain aging was introduced in the previous section, leveraging both cross-sectional and longitudinal data. In this section, refinements to the model are presented. First, the PCA is replaced with an autoencoder capable of capturing non-linear relationships to model normative brain variability, revealing more intricate latent patterns in the data. Second, in the NMF part, a weighting coefficient is incorporated into the loss function to balance the contribution of C-map and L-map to the dictionary learning process. This adjustment is necessary because, while the number of subjects with cross-sectional and longitudinal data was comparable in the previous application to the BLSA dataset, longitudinal data are typically less abundant than cross-sectional data. Since longitudinal data provide a more direct measurement of brain changes, it is crucial to balance the influence of cross-sectional data in the dictionary estimation. This refined model is referred to as Coupled Cross-sectional and Longitudinal NMF (CCL-NMF) to distinguish it from the earlier version.

Following the presentation of modifications made to the model, semi-synthetic data with predefined (ground truth) disease patterns and severity levels were used to evaluate the proposed methodology. The approach was subsequently applied to a large, multi-cohort aging population from the iSTAGING consortium and identified seven dominant components of brain atrophy that were consistent cross-sectionally and longitudinally. The identified components were correlated with AD biomarkers, cognitive function, CVD risk factors, and progression of cognitive impairment, underscoring the utility of the method in capturing the intricate complexity of aging and disease-related neurodegenerative processes in the human brain. Furthermore, comparisons with a state-of-the-art deep learning model using the same dataset demonstrated that the CCL-NMF components provided improved predictive performance for biomarkers and clinical variables—including amyloid, tau, cognitive worsening, hypertension, obesity, and APOE status—thereby refining our grasp of brain aging pathways. Finally, this approach offered a practical approach for researchers to quantify the expression of these seven atrophy components in their datasets through simplified, readily applicable models.

4.3.1 Methods

4.3.1.1 Model

Normative modeling for estimating the cross-sectional deviation map (C-map)
Denote **S1** to be the reference population and **S2** to be the target population. The deviations of **S2** from the normative space drawn by **S1** are estimated using an adversarial autoencoder (AA)[199][200] model. The core concept of this normative method is that because the AA is exclusively trained on **S1** data, it learns to encode and precisely reconstruct the **S1** data, but it will be less accurate when reconstructing data from **S2**. Specifically, the error between the input and the reconstructed estimate is expected to capture the deviation of **S2** from **S1**.

The AA architecture comprises an encoder (E) with two hidden layers, each containing 110 neurons and a latent space dimension of 10 neurons. The decoder (D) and the discriminator (D_z) are similarly structured, with two hidden layers of 100 neurons each. The latent space is regularized to match a Gaussian distribution. All hidden layers employ a leaky Rectified Linear Unit (ReLU) with non-linearity, while the latent space and the decoder's output layer utilize a linear activation function.

The AA training has two phases:
1) Reconstruction phase: This phase minimizes the reconstruction loss, ensuring the output closely matches the input. The encoder maps data (**x**) to latent space (**z**), and the decoder reconstructs it. The reconstruction loss is:

$$\mathrm{L}_{\mathrm{recon}} = \|\mathbf{x} - \mathrm{D}(\mathrm{E}(\mathbf{x}))\|_2^2 \quad (8)$$

2) Regularization phase: This phase uses adversarial training to enforce the latent space (**z**) to match the prior Gaussian distribution (P(z)). The discriminator distinguishes real samples from prior (P(z)) and fake samples from the encoder. The adversarial loss is:

$$\mathrm{L}_{\mathrm{adv}} = \mathbb{E}[\log(\mathrm{D}_z(\mathbf{z}))] + \mathbb{E}[\log(1 - \mathrm{D}_z(\mathrm{E}(\mathbf{x})))] \quad (9)$$

The encoder minimizes this loss to fool the D_z.

The Adam optimizer is used for 1000 epochs and applies early stopping with 50 epochs of patience. A minibatch approach is implemented within this gradient descent-based optimizer, with a batch size of 200. A cyclical learning rate enhances the training efficiency, facilitating convergence with fewer

epochs. The initial learning rate is 0.0001, with a maximum learning rate of 0.005. The learning rate cycle follows a basic triangular shape with an amplitude decay factor (gamma) of 0.98.

The features were first corrected for sex and DLICV. The linear correction models were trained on **S1** baseline measures and were applied to both **S1** baseline measures and **S2** baseline and longitudinal measures. Before the AA, the features were standardized to z-scores. Again, the z-score models were trained on **S1** baseline measures and applied to both **S1** and **S2** baseline measures, since the AA concerns only cross-sectional data. The **S1** was split into three subsets: $\mathbf{S1_{train}}$, $\mathbf{S1_{val}}$, and $\mathbf{S1_{heldout}}$, with a split ratio of 65%, 15%, and 20%, respectively. The AA was trained on $\mathbf{S1_{train}}$, validated on $\mathbf{S1_{val}}$, and then applied to **S2** baseline measures. The mean squared deviation, $\mathrm{MSD} = \frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}(\mathrm{x_j} - \hat{\mathrm{x}}_\mathrm{j})^2$, where R is the number of brain regions, x is the input and $\hat{\mathrm{x}}$ is the reconstructed output, was used to select individuals with the largest deviations. Here, since the focus was on atrophy-related volume loss, brain regions with negative deviations (i.e., larger reconstructed output than input) were not considered in the MSD calculation. Those **S2** individuals with MSD≥75th percentile MSD of $\mathbf{S1_{heldout}}$, referred to as **S3**, were selected as the ones with large neuropathology deviation from normality and were used in the NMF. C-map with a size equal to the number of brain regions by the number of **S3** individuals includes the deviations. Negative deviations have been suppressed by replacing them with 0.

*Joint NMF (**Figure 4.4**)*

After incorporating the weighting coefficient α, determined based on the application or dataset, that balances the contributions of the C-map and L-map to the dictionary learning process, the loss function becomes:

$$\mathrm{L} = \alpha\|\mathrm{X_C} - \mathrm{WH_C}\|_\mathrm{F}^2 + \|\mathrm{X_L} - \mathrm{WH_L}\|_\mathrm{F}^2 \quad (10)$$

And the multiplicative update rule becomes:

$$\mathrm{w_{ij}} \leftarrow \mathrm{w_{ij}} \frac{(\alpha \mathrm{X_C H_C}^\mathrm{T} + \mathrm{X_L H_L}^\mathrm{T})_\mathrm{ij}}{(\alpha \mathrm{W X_C H_C}^\mathrm{T} + \mathrm{W X_L H_L}^\mathrm{T})_\mathrm{ij}} \quad (11)$$

$$\mathrm{h_{ij}^I} \leftarrow \mathrm{h_{ij}^I} \frac{(\mathrm{W^T X_I})_\mathrm{ij}}{(\mathrm{W^T W H_I})_\mathrm{ij}}, (\mathrm{I} = \mathrm{C}, \mathrm{L}) \quad (12)$$

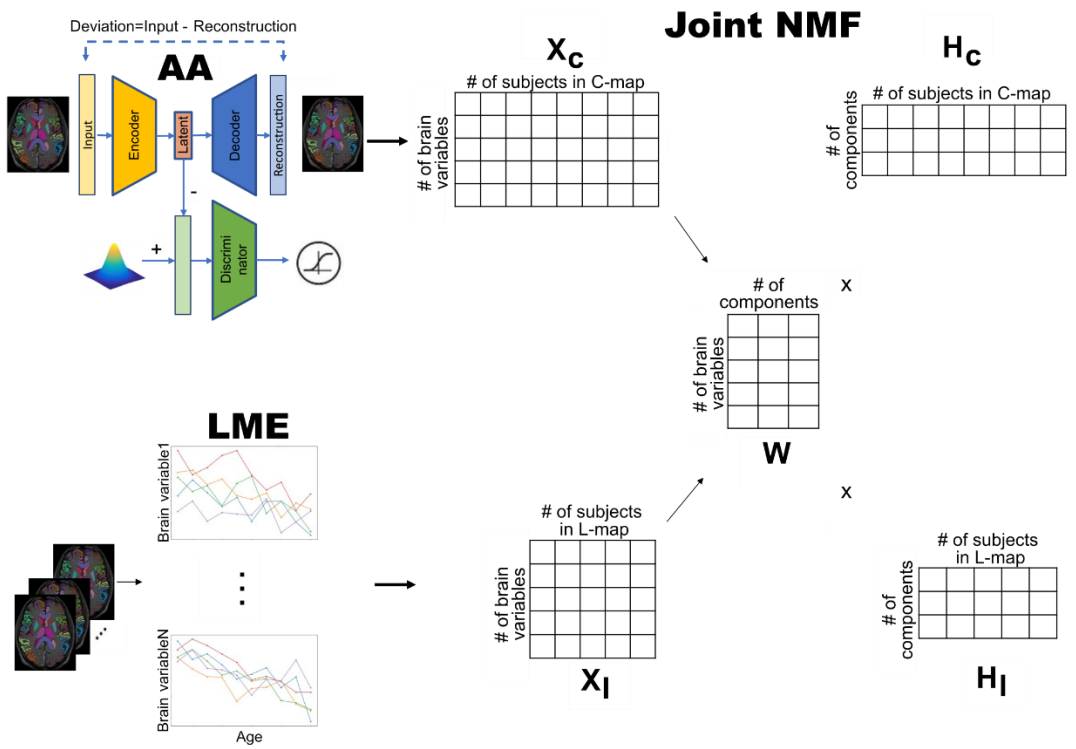


Figure 4.4: Conceptual overview of the CCL-NMF model. *Adversarial autoencoder (AA) estimates the cross-sectional deviation map (C-map) of the target population (here, an aging cohort) from the normative space defined by the reference population (here, a healthy middle-aged cohort). AA is exclusively trained on the reference population, and thus, it learns to encode and precisely reconstruct the reference population data. When reconstructing data from the target population, the AA produces an error that captures the deviation of the target population from the reference cohort on an individual basis. Target individuals with the largest deviations are used in the NMF. Linear mixed-effects models (LME) estimate the longitudinal rate of change map (L-map) for the target individuals with the largest cross-sectional deviations with multiple measurements over time. C-map, X_C, of size DxN_C and L-map, X_L, of size DxN_L, where D represents the dimensionality of brain features, and N_C (N_L) denotes the number of subjects with cross-sectional deviations (longitudinal change rates), are factorized into a shared dictionary W of size DxK and separate loading coefficients (H_C of size KxN_C and H_L of size KxN_L) through a joint NMF scheme: $X_C \simeq WH_C$, $X_L \simeq WH_L$, s.t. $W>0$, $H_C>0$, $H_L>0$. K is the number of components. The shared dictionary ensures that cross-sectional components of brain aging are consistent with dynamic progression patterns captured by longitudinal measurements. 'x' between the W and the H matrices stands for matrix multiplication.*

4.3.1.2 Dataset

Here, CCL-NMF was applied to parse the heterogeneity of aging-related GM changes using 119 GM ROIs (the 114 GM ROIs displayed in **Table 4.1** and five ROIs in the cerebellum: bilateral cerebellum exterior, Cerebellar Vermal Lobules I-V, VI-VII, and VIII-X) extracted from baseline T1-weighted MRI. The T1-w image intensity inhomogeneity was corrected[125], followed by multi-atlas skull-stripping[126], and then the ROIs were segmented using a multi-atlas, multi-warp label fusion-based method[127]. All participants gave written informed consent to the study for data acquisition and analyses according to the Declaration of Helsinki. The institutional review board of the University of Pennsylvania approved this project.

Semi-synthetic dataset

Semi-synthetic data generated by simulating atrophy patterns in cognitively normal individuals from the UK Biobank cohort were employed to validate the methodology. The dataset comprised 4,517 CU (mean age 51.83±2.33 years, 56.61% females), which were divided into two subsets: a 20% subset (**$S1_{syn}$**: N=903; mean age 51.90±2.34 years, 56.59% females) comprised individuals whose data remained unchanged, while the remaining 80% (**$S2_{syn}$**: N=3,614; mean age 51.81±2.33 years, 56.61% females) underwent simulated atrophy. In **$S2_{syn}$**, atrophy was introduced across 40 timepoints, each separated by one year, with varying patterns and onset times randomly selected from a Gaussian distribution (μ=7, σ=3). Within **$S2_{syn}$**, 20% of individuals were given synthetic frontal atrophy, 20% occipital, 20% parietal, 20% subcortical, and 20% temporal atrophy. The ROIs affected in each atrophy pattern are listed in **Table 4.2**. The simulation applied a 1% annual atrophy rate to the ROIs within the pattern and a 0.1% rate to the remaining brain ROIs. For the cross-sectional dataset, a single timepoint was randomly selected for each subject from the 40 available, resulting in a synthetic age distribution of 71.30±11.82 years. The L-map of **$S3_{syn}$** was calculated using an LME model with baseline age and ROI volume as covariates.

Table 4.1: Anatomic gray matter (GM) regions of interest (ROIs) affected in each pattern in the semi-synthetic data.

Frontal	Left anterior orbital gyrus
	Left lateral orbital gyrus
	Left medial orbital gyrus
	Left posterior orbital gyrus
	Right anterior orbital gyrus
	Right lateral orbital gyrus
	Right medial orbital gyrus
	Right posterior orbital gyrus
	Left anterior insula
	Left posterior insula
	Right anterior insula
	Right posterior insula
	Left frontal pole
	Left middle frontal gyrus
	Left opercular part of the inferior frontal gyrus
	Left orbital part of the inferior frontal gyrus
	Left precentral gyrus
	Left superior frontal gyrus
	Left triangular part of the inferior frontal gyrus
	Right frontal pole
	Right middle frontal gyrus
	Right opercular part of the inferior frontal gyrus

	Right orbital part of the inferior frontal gyrus
	Right precentral gyrus
	Right superior frontal gyrus
	Right triangular part of the inferior frontal gyrus
	Left gyrus rectus
	Left medial frontal cortex
	Left precentral gyrus medial segment
	Left superior frontal gyrus medial segment
	Left subcallosal area
	Left supplementary motor cortex
	Right gyrus rectus
	Right medial frontal cortex
	Right precentral gyrus medial segment
	Right superior frontal gyrus medial segment
	Right subcallosal area
	Right supplementary motor cortex
	Left central operculum
	Left frontal operculum
	Left parietal operculum
	Right central operculum
	Right frontal operculum
	Right parietal operculum
Occipital	Left occipital fusiform gyrus
	Right occipital fusiform gyrus
	Left inferior occipital gyrus
	Left middle occipital gyrus
	Left occipital pole
	Left superior occipital gyrus
	Right inferior occipital gyrus
	Right middle occipital gyrus
	Right occipital pole
	Right superior occipital gyrus
	Left calcarine cortex
	Left cuneus
	Left lingual gyrus
	Right calcarine cortex
	Right cuneus
	Right lingual gyrus
Parietal	Left angular gyrus
	Left postcentral gyrus
	Left supramarginal gyrus
	Left superior parietal lobule
	Right angular gyrus
	Right postcentral gyrus
	Right supramarginal gyrus
	Right superior parietal lobule
	Left postcentral gyrus medial segment
	Left precuneus

	Right postcentral gyrus medial segment
	Right precuneus
Temporal	Left fusiform gyrus
	Right fusiform gyrus
	Left inferior temporal gyrus
	Left middle temporal gyrus
	Left superior temporal gyrus
	Left temporal pole
	Right inferior temporal gyrus
	Right middle temporal gyrus
	Right superior temporal gyrus
	Right temporal pole
	Left planum polare
	Left planum temporale
	Left transverse temporal gyrus
	Right planum polare
	Right planum temporale
	Right transverse temporal gyrus
Subcortical	Left Accumbens Area
	Left Caudate
	Left Pallidum
	Left Putamen
	Right Accumbens Area
	Right Caudate
	Right Pallidum
	Right Putamen
	Left Thalamus Proper
	Right Thalamus Proper
	Left Amygdala
	Left Basal Forebrain
	Left Hippocampus
	Right Amygdala
	Right Basal Forebrain
	Right Hippocampus

Before the NMF, gaussian noise specific to each ROI was added to both maps to make the problem more realistic. For the C-map, noise was added for each subject and each ROI from a Gaussian distribution with a mean equal to the mean (across subjects not belonging to the pattern the specific ROI belongs to) C-map value of the ROI, and a standard deviation equal to the standard deviation of the C-map value of the ROI. For the L-map, noise was added for each subject and each ROI from a Gaussian distribution with a mean equal to the absolute value/magnitude of the mean (across subjects) L-map value of the ROI, multiplied by a random floating number between 0 and 2.5, and a standard deviation equal to the absolute value/magnitude of the standard

deviation of the L-map value of the ROI, multiplied by a random floating number between 0 and 2.5.

Three types of NMF experiments were conducted for the five simulated atrophy patterns in the semi-synthetic dataset. First, the NMF was run using solely the C-map. Second, the NMF was run using solely the L-map. Third, the CCL-NMF was run utilizing both cross-sectional (C-map) and longitudinal maps (L-map). In the last case, a sensitivity analysis was performed to assess the impact of the α coefficient from Eq. (10) on the accuracy with which the CCL-NMF dictionary captured the simulated atrophy patterns.

To evaluate the model's ability to identify the simulated atrophy patterns, the inner product matrix was calculated between the l2-normalized matrix representing the ground truth simulated atrophy patterns and the l2-normalized dictionary of the model. The closer the inner product matrix is to the diagonal matrix representing the perfect identification/reconstruction of the ground truth simulated atrophy patterns, the better the dictionary captures the simulated atrophy patterns. To quantify the divergence from perfect ground truth reconstruction, the norm of the difference between the two matrices (termed divergence matrix) was used; a smaller norm of the divergence matrix indicates a more accurate reconstruction of the ground truth.

iSTAGING aging dataset

The dataset for real data experiments was drawn from the iSTAGING consortium. Data from the following studies were included: ADNI, AIBL, BIOCARD, BLSA, CARDIA, HANDLS, OASIS, Penn-PMC, SHIP, UK Biobank, WHIMS, and WRAP. The imaging parameters for each study are presented elsewhere[130]. Interstudy ROI harmonization was conducted using the Neuroharmonize toolbox[78]), based on the ComBat statistical methodology. Clinical data and cognitive status, where available, were provided by the source study.

The reference population **S1** consisted of CU individuals without known CVD risk factors (obesity, hypertension, and diabetes) and age younger or equal to 50 years (N=977; mean age 39.88±8.09 years, 54.86% females, 100% CU). The target group **S2** comprised individuals older than 50 (N=48,949; mean age 65.41±7.92 years, 53.98% females, 94.23% CU). The demographics of **S1** and **S2** populations by origin study are displayed in **Table 4.3**. The L-map was calculated using LME models with the site, baseline age, and ROI volume as covariates. The LME analysis was performed using individuals with three or more longitudinal measures to minimize uncertainty in the rate of change

estimation. Finally, the joint NMF was run to obtain a varying number of components K (K=2, .., 15).

Table 4.2: Demographic summary of reference (S1) and target (S2) population.

	Sample size		Age (years)		Sex (%males)		Diagnosis (%CU)	
	Target	Reference	Target	Reference	Target	Reference	Target	Reference
ADNI	2391	-	73.1±7.2	-	52.4	-	36.4	-
AIBL	922	4	73.1±6.4	45.4±2.5	43.5	25	76	100
BIOCARD	259	-	60.8±8	-	40.9	-	97.7	-
BLSA	916	100	70.1±9.5	40.5±7.4	47.3	42	97.5	100
CARDIA	534	170	53.9±2.3	47.2±2.3	46.4	50.6	1	100
HANDLS	147	33	58.6±5.8	42.9±4.4	44.9	57.6	1	100
OASIS	1097	10	71.8±9.2	47±2	45	20	73.3	100
PENN	959	-	73.9±8	-	43.1	-	20.8	-
SHIP	1810	660	63±7.9	37.6±8.1	48	44.1	100	100
UK BIOBANK	38582	-	64.5±7.3	-	47.1	-	100	-
WHIMS	1080	-	69.6±3.6	-	0	-	100	-
WRAP	252	-	62.1±5.8	-	29	-	99.6	-

4.3.1.3 Statistical analysis

Linear and logistic regression analyses were conducted to investigate the associations between CCL-NMF loading coefficients and AD pathology, cognition, SPARE-AD score, and CVD risk factors. Age, sex, and study (and education for cognitive scores) were used as covariates in the regression models. Cox proportional hazards models, adjusted for age, sex, DLICV, and education, were used to examine the associations between CCL-NMF loadings and the risk of progression from CU to MCI or from MCI to AD. Hazard ratios were calculated to quantify the effect of each CCL-NMF component on the risk of cognitive impairment progression. Bonferroni correction was applied to control for type I errors and account for multiple comparisons. Logistic regression and Cox proportional hazards models with 5-fold stratified cross-validation based on age, sex, and diagnosis were employed to assess the predictive accuracy for binary outcomes and progression from MCI to AD, respectively. Models using demographics alone, demographics plus CCL-NMF loadings, demographics plus R-indices, and models incorporating all previous predictors were compared. To apply the derived CCL-NMF model to external data sets without retraining, a regression model was developed for each CCL-NMF loading, incorporating ROIs, age, sex, and DLICV as predictors. Finally, the component visualization using RAVENS maps was implemented in the

Volume Imaging in Neurological Research, Co-Registration, and ROIs included (VINCI64 v5.03[201]) platform.

4.3.2 Results

4.3.2.1 Validation in the semi-synthetic dataset

Among the 3,614 subjects in the **S2$_{syn}$** group, 1,365 subjects (**S3$_{syn}$**) exceeded the 75th percentile MSD, and thus, their C-map was used in the NMF. A subset of 300 was used to generate the L-map, reflecting the realistic scenario where only a subset of individuals has longitudinal data.

Figure 4.5 presents the results from three different NMF cases:

1) Using solely C-map (N_C=1,365; synthetic mean age 77.11±11.31 years, 38.24% females, patterns: frontal: 319 (23%), occipital: 228 (17%), parietal: 239 (18%), subcortical: 307 (22%), temporal: 272 (20%)) in the NMF – C-NMF –, the norm was 1.67.
2) Using only L-map (NL=300; synthetic mean age 76.82±10.92 years, 40% females, patterns: frontal: 83 (28%), occipital: 49 (16%), parietal: 50 (17%), subcortical: 54 (18%), temporal: 64 (21%)) – L-NMF –, the norm was 1.46.
3) Using C-map (N_C=1,365) and L-map (N_L=300) – CCL-NMF –, the norm was 1.39.

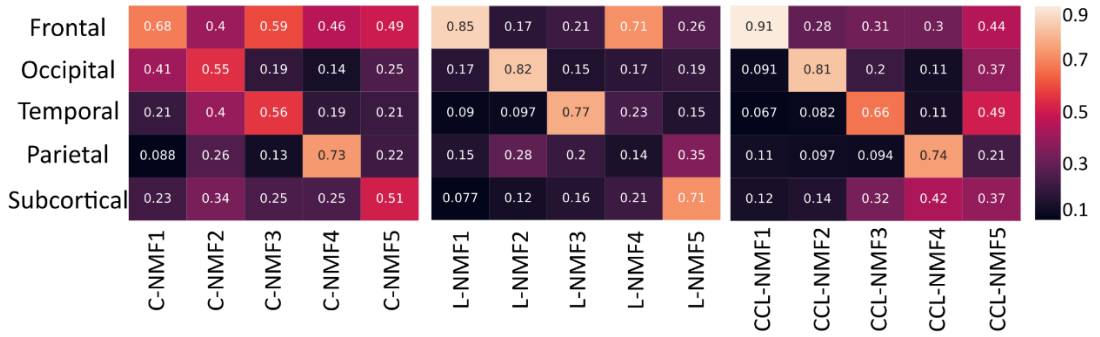


Figure 4.5: Results in semi-synthetic data. *From left to right, matrices show the inner product matrix between the l2-normalized matrix of the ground truth simulated atrophy patterns and the l2-normalized dictionary for C-NMF (only C-map), L-NMF (only L-map), and CCL-NMF using C-map and L-map. A strongly diagonal structure indicates agreement between the simulated and estimated patterns of brain change.*

Fig. 4.5 shows that the model using both C-map and L-map can identify the simulated patterns with higher accuracy. **Figure 4.6** displays the divergence norm as a function of α weighting coefficient where α ranges from 0 (i.e., L-NMF) to 1 (no data type balancing). The most accurate pattern reconstruction was achieved using $\alpha=N_L/N_C=300/1,365=0.22$. Since longitudinal data provide a more direct measure of brain changes but are less readily available than

cross-sectional data, it is essential to balance the contribution of cross-sectional data in dictionary learning. This can be achieved by weighting their influence according to the ratio of sample sizes between the two data types.

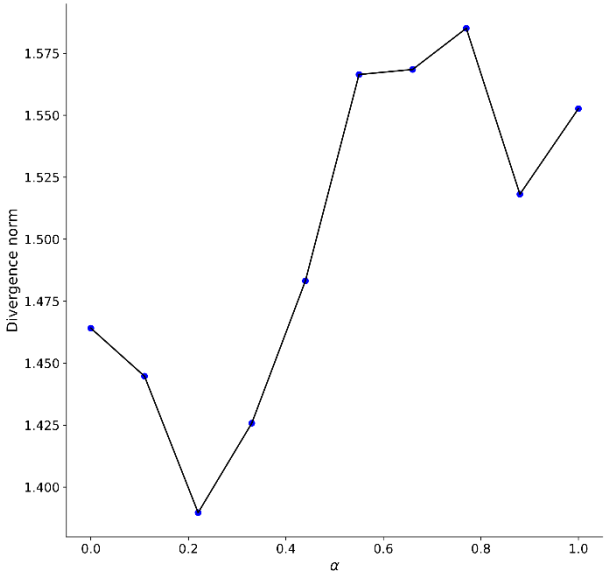


Figure 4.6: Divergence norm as a function of α weighting coefficient. *α ranges from 0 (i.e., L-NMF) to 1 (no balancing). The divergence norm is defined as the norm of the difference between the inner product matrix (between the l2-normalized matrix representing the ground truth simulated atrophy patterns and the l2-normalized dictionary of the model) and the identity matrix; a smaller norm of the divergence matrix indicates a more accurate reconstruction of the ground truth.*

4.3.2.2 Application to the iSTAGING aging dataset

Among the 48,949 subjects in the **S2** group, 13,950 subjects (**S3**: mean age 66.47±8.38 years, 30.82% females, 89.5% CU) exceeded the 75th percentile MSD. Out of those subjects, 1,063 subjects had three or more longitudinal measurements and thus built the L-map. The follow-up time was 4.71±3.79 years, and the baseline age was 72.81±8.08 years. Detailed demographic characteristics of the populations used for constructing the C-map and L-map are provided in **Tables 4.4** and **4.5**.

CCL-NMF identified seven distinct components of brain aging

The CCL-NMF was run from K=2 to 15. Split-half reproducibility index, sparsity, and weighted reconstruction error (as defined by Eq. (10)) were used to determine the optimal number of components (**Figure 4.7**). Sparsity increased with higher K, while the reconstruction error decreased with higher K. The split-half reproducibility analysis revealed a declining trend in reproducibility as the number of components increased, with a peak occurring at K=3 and 7. K=7 was selected for subsequent analyses because of its higher sparsity and lower reconstruction error than the K=3 solution. **Figure 4.8** displays the components for the two halves used to calculate the reproducibility index.

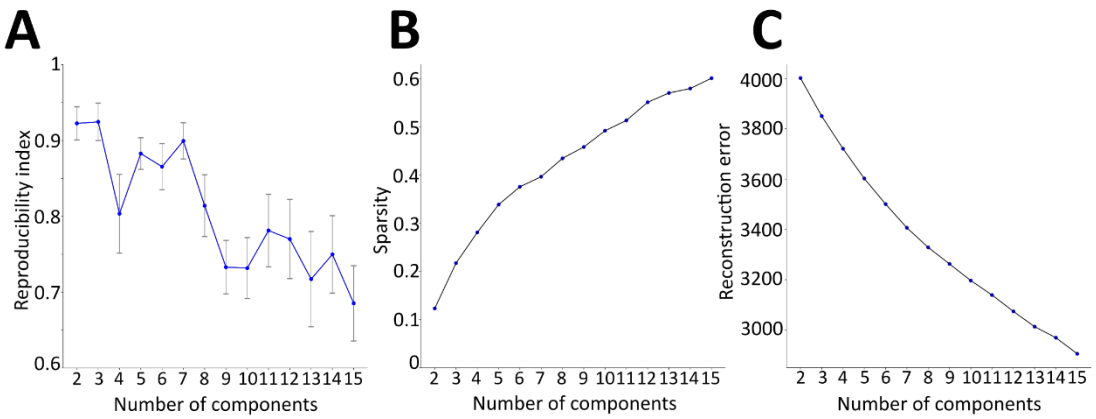


Figure 4.7: A) Split-half reproducibility index, B) sparsity, and C) weighted reconstruction error reported as functions of the number of CCL-NMF components (N_C=13,950, N_L=1,063).

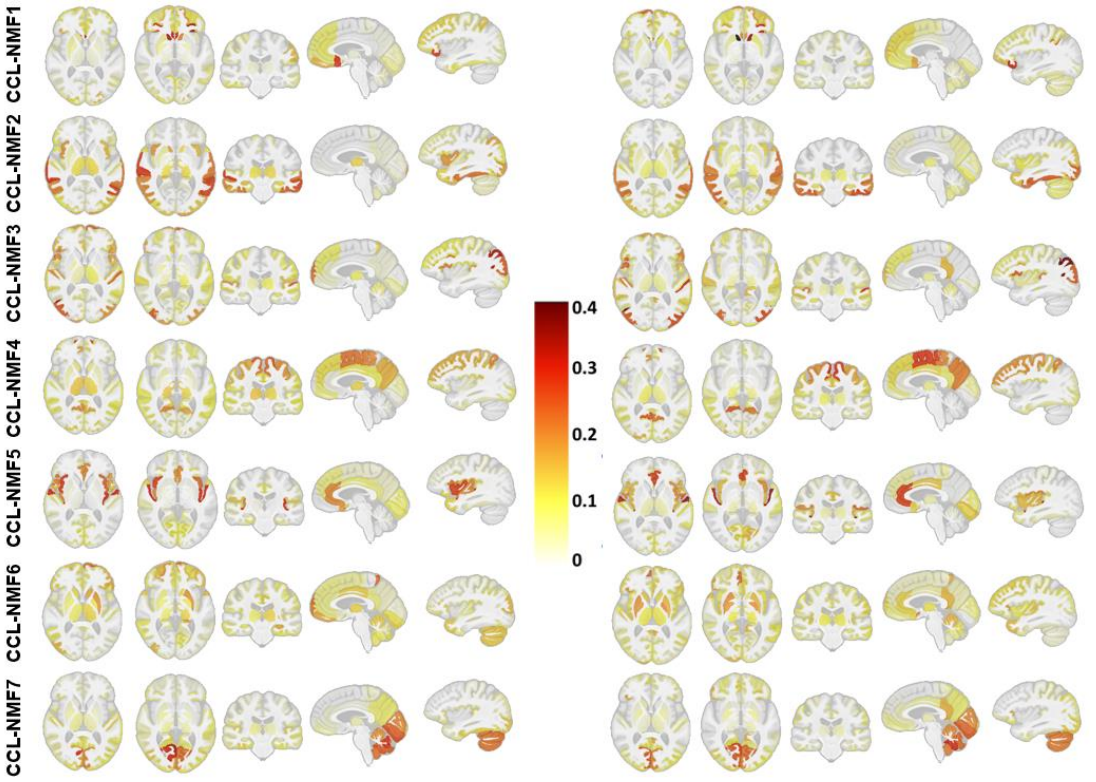


Figure 4.8: Split sample CCL-NMF dictionaries. *Left side: CCL-NMF dictionary obtained for the first half split. Right side: matched CCL-NMF dictionary obtained for the second half split. Red (white) colors indicate higher (lower) contribution of the brain region in the CCL-NMF component.*

Table 4.3: Demographic summary and volumetric measures of individuals included in C-map (N_C=13,950). *Abbreviations: WMH, white matter hyperintensities; Aβ, Amyloid β; APOE, Apolipoprotein E.*

Other races: Hispanic/Latino, Native American, Multiracial, unknown, other; information about races is presented as given in the originating studies.

Other diagnoses: Frontotemporal Dementia, Hydrocephalus, Lewy Body Dementia, Posterior Cortical Atrophy, Parkinson's Disease, Vascular Dementia, Dementia, early MCI, and others; information about diagnoses is presented as given in the originating studies.

Study	Sample size	Sex (% males)	Race	Age (years)	Diagnosis	Aβ (% positive)	APOE-ε4 carriers	Total brain volume (mm^3)	Total WMH volume (mm^3)
ADNI	1010	70.59	Asian: 1.36 Black: 3.09 White: 94.56 Other: 0.99	74.16 (54.27-91.31)	AD: 28.22 CU: 22.77 MCI: 49.01 Other: 0	68.2	48.06	1217432.5 (854135.12-1618505.63)	6053.42 (0-70106.37)
AIBL	385	60.78	Asian: 0 Black: 0 White: 100 Other: 0	74 (55-93)	AD: 15.84 CU: 65.46 MCI: 18.70 Other: 0	30	28.08	1209099.08 (892467.64-1544824.86)	7416.5 (0-78002.12)
BIOCARD	71	74.65	Asian: 1.43 Black: 0 White: 98.57 Other: 0	62.8 (50.27-86.27)	AD: 1.41 CU: 94.36 MCI: 4.23 Other: 0	34.55	29.58	1325000.91 (1040860.59 - 1635686.72)	2667.63 (186.56-9987.89)
BLSA	214	75.23	Asian: 1.87 Black: 19.16 White: 77.10 Other: 1.87	71.87 (51-93)	AD:1.87 CU: 94.86 MCI: 2.8 Other: 0.47	0	21.84	1230488.56 (938491.24-1539092.29)	6897.04 (0-76972.58)
CARDIA	109	68.81	Asian: 0 Black: 25.69 White: 74.31 Other: 0	53.77 (51-61)	AD: 0 CU: 100 MCI: 0 Other: 0	0	29.58	1321156.85 (961006.16-1659822)	2384.61 (181.2-19571)

HANDLS	24	66.67	Asian: 0 Black: 16.67 White: 83.33 Other: 0	60.41 (53-71.7)	AD: 0 CU: 100 MCI: 0 Other: 0	0	0	1223220.35 (952729.24-1431638.46)	1666.45 (16.8-6998.4)
OASIS	399	66.17	Asian: 0.75 Black: 11.31 White: 87.94 Other: 0	73.57 (50.08-99.32)	AD: 34.84 CU: 58.9 MCI: 4.01 Other: 2.25	41.82	47.22	1210603.71 (851367.47-1543633.36)	4695.31 (60-33950.19)
PENN	417	59.71	Asian: 0.96 Black: 13.91 White: 82.49 Other: 2.64	74.97 (51-95)	AD: 40.05 CU: 10.79 MCI: 26.86 Other: 22.3	83.58	41.27	1197960.9 (843412-1742158)	9270.58 (0-61557)
SHIP	366	77.32	Asian: 0 Black: 0 White:100 Other: 0	64.84 (51-86)	AD: 0 CU: 100 MCI: 0 Other: 0	0	0	1322079.46 (966118.33-1629973.71)	4349.45 (0-64119.96)
UK Biobank	10653	71.14	Asian: 0.8 Black: 0.5 White: 97.65 Other: 1.05	64.93 (50-81.77)	AD: 0 CU: 100 MCI: 0 Other: 0	0	28.06	1323894.26 (792061.61-1882503.1)	3306.6 (0-61651.68)
WHIMS	264	0	Asian: 1.89 Black: 4.17 White: 93.56 Other: 0.38	69.72 (64-79)	AD: 0 CU: 100 MCI: 0 Other: 0	0	23.29	1166355.62 (752469.43-1422641.6)	4649.42 (0-52623.63)
WRAP	38	60.53	Asian: 0 Black: 2.63 White: 94.74 Other: 2.63	61.45 (50.3-75.8)	AD: 0 CU: 100 MCI: 0 Other: 0	19.23	50	1282488.11 (1095699-1547776)	1403.97 (0-8402)

Table 4.5: Demographic summary and volumetric measures of individuals included in L-map (N_L=1,063). *Abbreviations: WMH, white matter hyperintensities; Aβ, Amyloid β; APOE, Apolipoprotein E. AD, Alzheimer's disease; CU, cognitive unimpaired; MCI, mild cognitive impairment.*
Other races: Hispanic/Latino, Native American, Multiracial, unknown, other; information about races is presented as given in the originating studies.

Study	Sample size	Sex (% males)	Race	Age (years)	Diagnosis	Aβ (%	APOE-ε4 carriers	Total brain volume (mm^3)	Total WMH volume (mm^3)
ADNI	706	70.82	Asian: 1.47 Black: 2.5 Other: 0.89 White: 95.14	74.32 (55.87-91.31)	AD: 26.91 CU: 18.98 MCI: 54.11	68.02	50.14	1215903.87 (854135.12-1618505.63)	5564.47 (0-70106.37)
AIBL	103	66.02	Asian: 0 Black: 0 Other: 0 White: 100	73.02 (55-85.53)	AD: 6.8 CU: 77.67 MCI: 15.53	25	28.57	1225920.1 (1014697-1544824.86)	6756.61 (104.4-36731)
BIOCARD	39	74.36	Asian: 2.56 Black: 0 Other: 0 White: 97.44	60.3 (50.27-79.61)	AD: 0 CU: 97.44 MCI: 2.56	27.59	25.64	1346877.92 (1071488.67-1635686.72)	2127.87 (186.56-8256.45)
BLSA	110	79.09	Asian: 0 Black: 13.64 Other: 2.72 White: 83.64	72.14 (57-87)	AD: 0 CU: 96.36 MCI: 3.64	0	17.43	1250270.69 (980211.64-1536942.04)	6330.66 (0-76972.58)
OASIS	67	68.66	Asian: 0 Black: 2.99 Other: 0 White: 97.01	67.81 (50.08-89.31)	AD: 4.48 CU: 95.52 MCI: 0	23.26	38.81	1266936.29 (1040760.62-1543633.36)	1585.4 (121-4448)
PENN	18	72.22	Asian: 0 Black: 5.56 Other: 0 White: 94.44	73.27 (59-90.13)	AD: 33.33 CU: 38.89 MCI: 22.22 Hydrocephalus: 5.56	80	53.33	1248390.92 (996562.44-1548755.65)	10290.56 (256.54-45686)
WRAP	20	75	Asian: 0 Black: 0 Other: 0 White: 100	62.59 (50.3-74.1)	AD: 0 CU: 100 MCI: 0	28.57	45	1288512.55 (1149044-1547776)	1214.32 (230-4568)

Figure 4.9 shows the seven identified brain aging components. The components are sparse, orthogonal, and right-left symmetrical.
-**CCL-NMF1** captures atrophy primarily in part of the basal ganglia, including putamen and caudate, as well as atrophy in the orbital gyrus, gyrus rectus, and subcallosal area.
-**CCL-NMF2** indicates atrophy primarily in the medial temporal lobe, temporal pole, temporal gyrus, and fusiform gyrus.
-**CCL-NMF3** represents atrophy in the inferior frontal gyrus, occipital gyrus, and part of the temporal gyrus.
-**CCL-NMF4** exhibits medial frontoparietal atrophy, including the superior and middle frontal gyrus, precuneus, middle and posterior cingulate gyrus, supplementary motor cortex, and superior parietal lobule.
-**CCL-NMF5** is characterized by peri-Sylvian atrophy (insula, frontal and central operculum, and planum polare), as well as anterior cingulate gyrus atrophy.
-**CCL-NMF6** captures atrophy primarily in the basal ganglia, including the accumbens area, pallidum, and thalamus.
-**CCL-NMF7** captures atrophy in the cerebellum and medial occipital lobe, including the cuneus, calcarine cortex, and lingual gyrus.

Associations of aging components with clinical, cognition, and cognitive impairment progression.
The identified components exhibited differential associations with clinical features, cognitive measures, biomarkers, APOE alleles, and disease progression (**Figure 4.10**). **CCL-NMF2,** characterized by medial temporal lobe atrophy, showed the strongest associations with Alzheimer's disease pathology, cognitive decline, and progression from MCI to AD. **CCL-NMF5**, defined by prominent peri-Sylvian atrophy, showed strong links to cardiovascular disease risk factors, particularly WMH burden, obesity, and hypertension. Similarly, **CCL-NMF6** was significantly associated with elevated WMH and obesity.

CCL-NMF3 was closely related to advanced age and showed moderate associations with amyloid positivity and cognitive decline, although less pronounced than **CCL-NMF2**, and with WMH, albeit less strongly than **CCL-NMF5**. This component was more widely expressed across the sample (**Figure 4.11**), likely reflecting general aging effects rather than a specific pathological process. **CCL-NMF4** displayed a unique association with tau pathology and a negative correlation with age but showed no significant relationships with other AD biomarkers, cognitive decline, or CVD risk factors. However, the small tau sample size limits definitive conclusions. Lastly, **CCL-NMF7** was not significantly associated with AD biomarkers, CVD risk factors, or cognitive

impairment progression, potentially reflecting associations with other diseases or exposures for which specific biomarkers were unavailable.

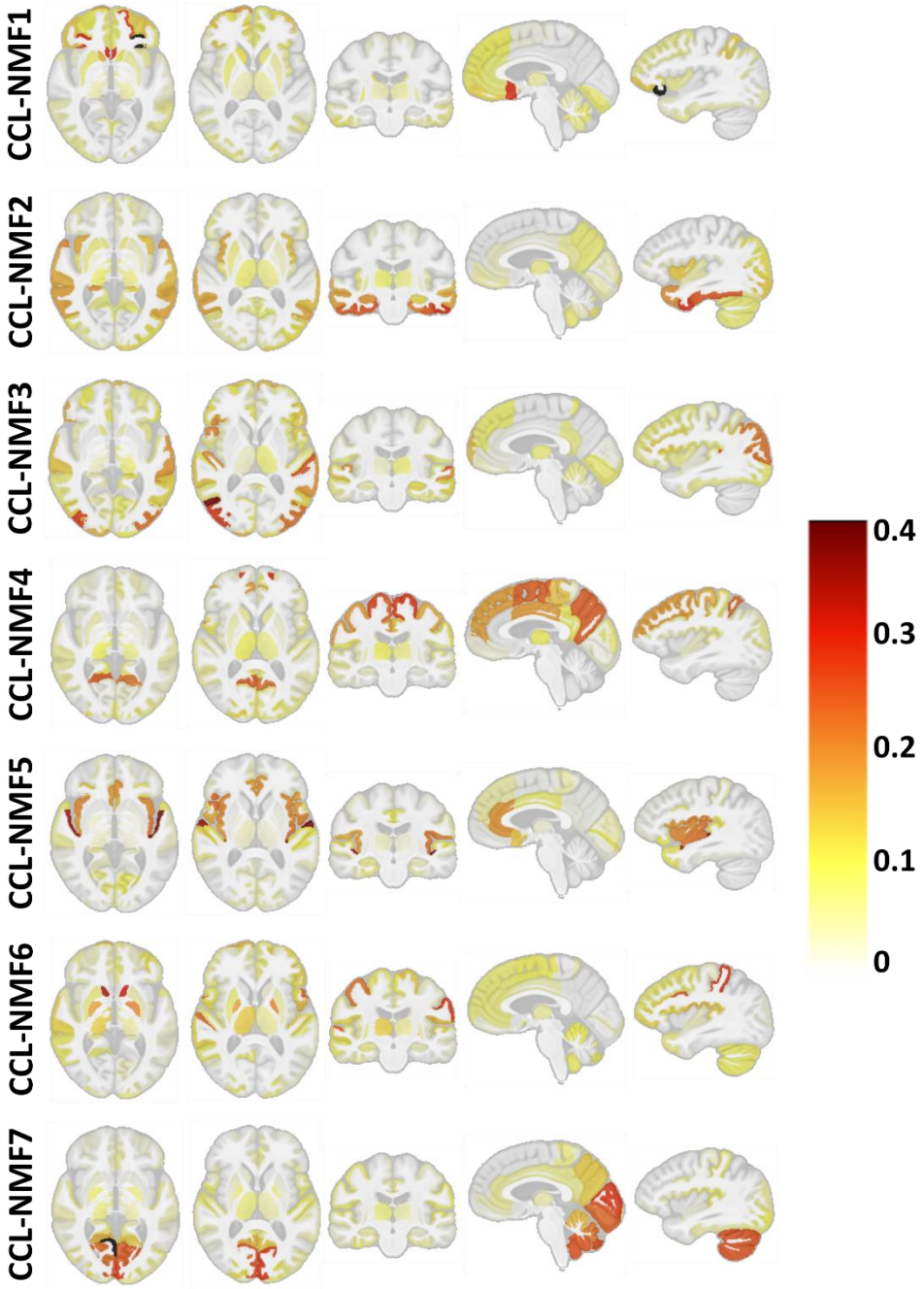


Figure 4.9: CCL-NMF dictionary in brain maps format for K=7 (N_C=13,950, N_L=1,063). *Red (white) colors indicate higher (lower) contribution of the visualized brain region in the CCL-NMF component.*

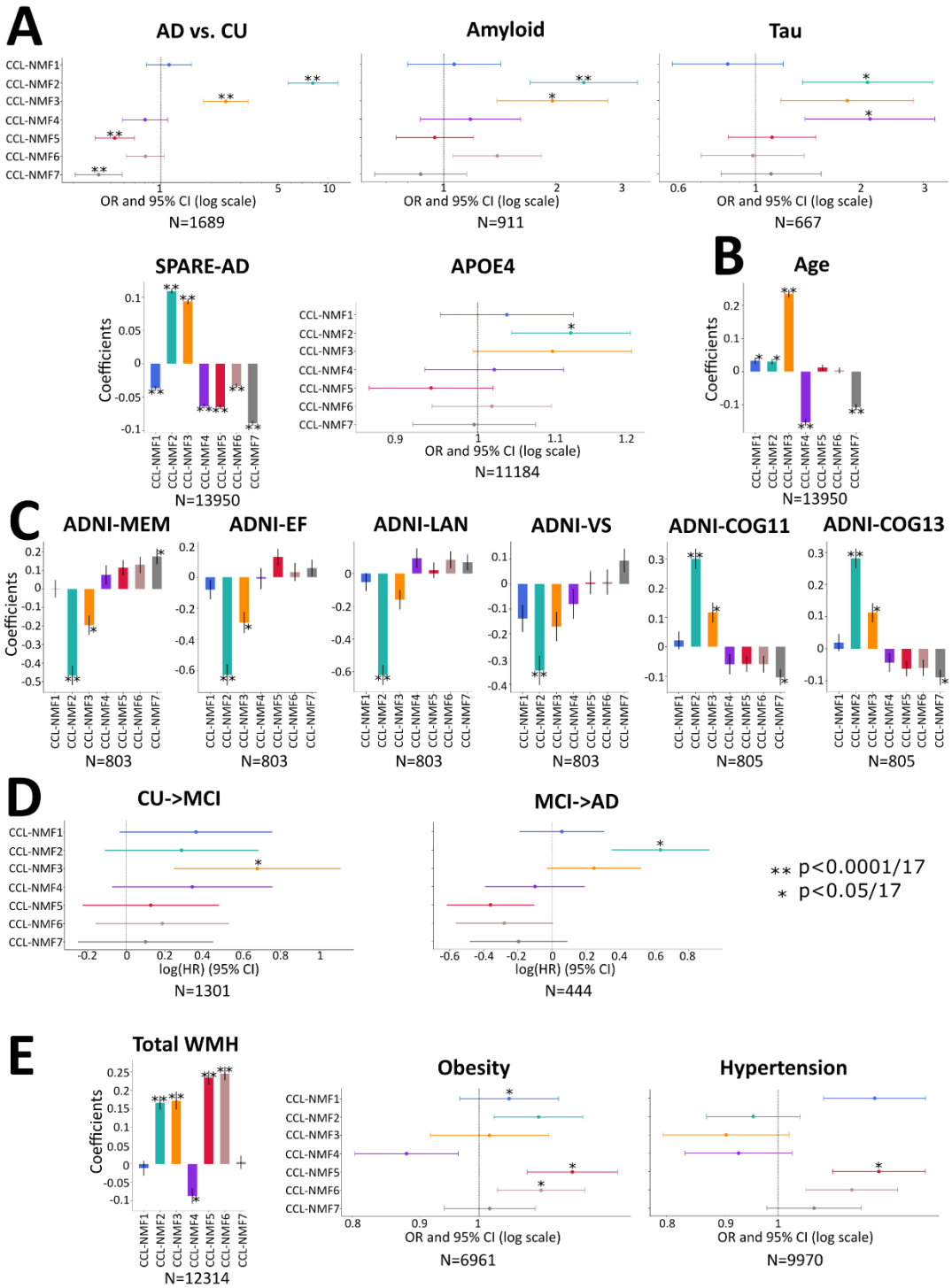


Figure 4.10: Associations of cross-sectional CCL-NMF loadings with A) AD-specific measures, B) age, C) cognition, D) cognitive impairment progression, and E) CVD risk factors. *P-values were adjusted using Bonferroni correction to account for the number of comparisons (N=17), controlling for type I errors. The follow-up time for CU to MCI and MCI to AD conversion was 5.37±4.29 and 2.6±2.65 years, respectively. Age, sex, and study were included as covariates in all models. However, for APOE-ε4 (APOE4) carrier status, age was not included as a covariate, while for ADNI cognitive scores, study was excluded, and education was included instead. Additionally, education was included as a covariate in models assessing cognitive decline progression (Part D). N indicates the sample size for the corresponding analysis shown in the graph. Biomarker and CVD risk factor status is defined as explained in the previous chapter. Abbreviations: SPARE-AD, spatial pattern of abnormality for recognition of early Alzheimer's*

disease; WMH, white matter hyperintensities; CU, cognitively unimpaired; MCI, mild cognitive impairment; ADNI, Alzheimer's Disease Neuroimaging Initiative; ADNI-MEM, ADNI memory composite; ADNI-VS, ADNI visuospatial functioning composite; ADNI-LAN, ADNI language composite; ADNI-EF, ADNI executive function composite; ADAS-COG, Alzheimer's disease assessment scale-cognitive subscale.

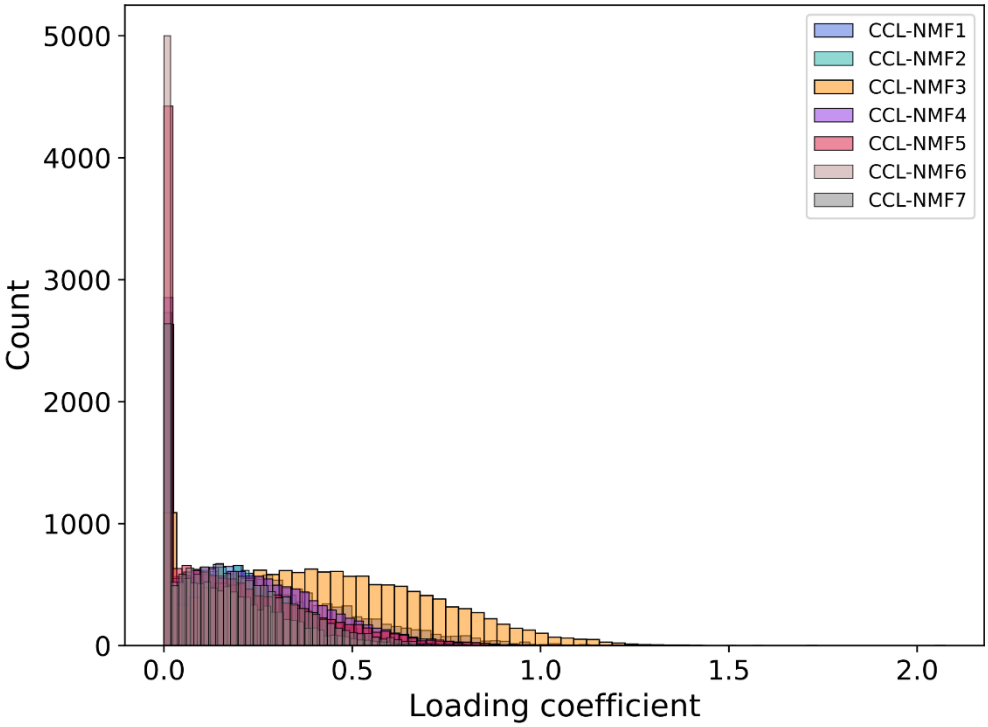


Figure 4.11: Histogram of cross-sectional CCL-NMF loadings.

Comparison with state-of-the-art heterogeneity model

The CCL-NMF components are generally consistent with, yet expand upon, the 5 dimensions of atrophy recently published by Yang et al.[172] with a different method yet on the same data (**Figure 4.12**). Key differences between these representations of brain aging are evident in **CCL-NMF6**, characterized by atrophy in the accumbens area, and **CCL-NMF7**, defined by atrophy in the cerebellum and medial occipital lobe. These patterns were not captured as distinct dimensions in the Surreal-GAN representation.

To evaluate the additional information captured by the CCL-NMF representation compared to Yang et al.'s approach, predictive models were developed using CCL-NMF loading coefficients, Surreal-GAN R-indices, and a combination of both as predictors for various outcomes: APOE-ε4 status (one or two ε4 alleles vs. none), amyloid positivity (positive vs. negative), tau positivity (positive vs. negative), obesity (obese vs. normal weight), and hypertension (hypertensive vs. normotensive). These models were evaluated using 5-fold cross-validation (stratified by age, sex, and diagnosis), with the area under the curve (AUC) as the performance metric. Additionally, the Cox proportional hazards model was

run to predict the progression from MCI to AD, and the concordance index (C-index) was calculated.

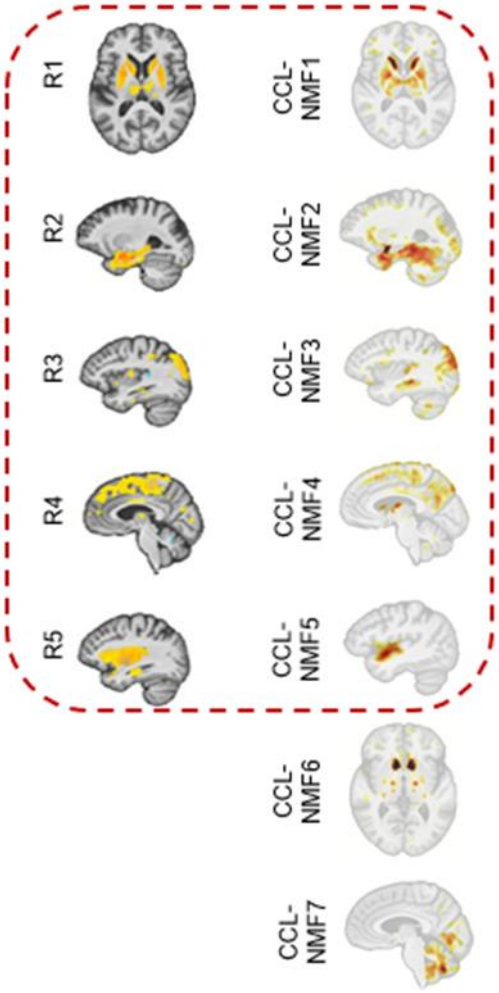


Figure 4.12: Atrophy patterns for CCL-NMF components vs. R-indices shown via voxel-wise t-tests performed for each CCL-NMF component (R-index) while adjusting for age, sex, DLICV, and the remaining CCL-NMF components (R-indices). *False discovery rate correction was conducted to adjust multiple comparisons with a p-value threshold of 0.001. Increasing voxel redness indicates stronger associations with a specific component or index. The first five CCL-NMF components show consistency with the five R-indices.*

As shown in **Figure 4.13**, incorporating either R-indices or CCL-NMF loadings significantly enhanced predictive performance across all features compared to demographic-only models. Notably, models using CCL-NMF loadings consistently outperformed those utilizing R-indices for all features, suggesting that the longitudinal information embedded in CCL-NMF provides more neuropathologically relevant components than the cross-sectional Surreal-GAN approach.

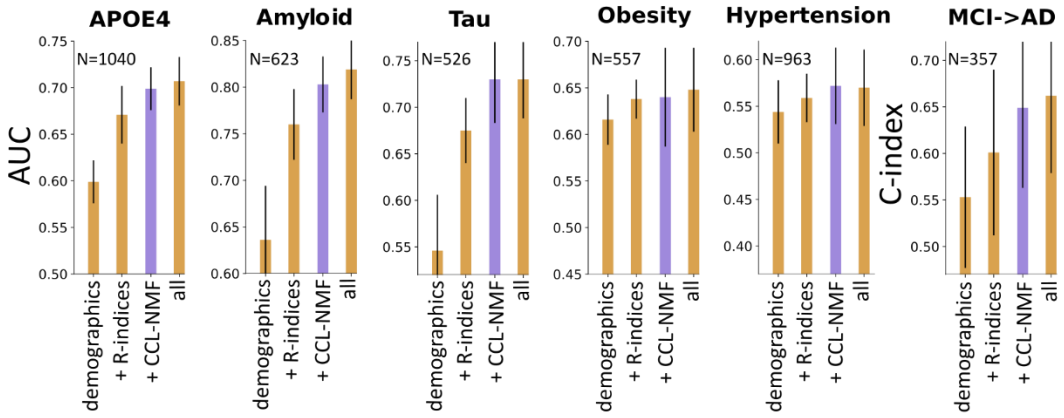


Figure 4.13: Area under the curve (AUC) and concordance index (C-index) for models predicting clinical variables and MCI to AD conversion, respectively. *The predictors can be 1) demographics alone, 2) demographics and R-indices, 3) demographics and CCL-NMF loadings, or 4) all previous combined. Age, sex, and study were included as covariates in all models. For APOE-ε4 (APOE4) carrier status analysis, age was excluded as a covariate, whereas education was additionally included as a covariate in the model evaluating the MCI to AD progression. N indicates the sample size for the corresponding analysis shown in the graph. The status of biomarkers and CVD risk factors is defined as explained in the previous chapter. Abbreviations: AD, Alzheimer's disease; MCI, mild cognitive impairment.*

Easily accessible out-of-sample estimation of CCL-NMF loadings in new datasets

Although the formulation is foundational at the discovery stage, where the CCL-NMF components were established, rederiving components for new datasets would be intensive. Thus, we trained a regression model to estimate CCL-NMF coefficients from ROI data and demographics. This approach facilitates the broader adoption of these brain aging components by the research community. Notably, the ROIs were not harmonized. This circumvents the need for users to acquire reference subjects for harmonization—a step that was essential in the discovery phase to mitigate scanner variability but challenging in new datasets or single-case studies. **Figure 4.14** illustrates the Spearman correlations between original (denoted by uppercase letters) and approximated (denoted by lowercase letters) loading coefficients, estimated using regression models with 5-fold stratified cross-validation based on age, sex, and diagnosis. The correlations between the original and their corresponding approximated loading coefficients are very high, ranging between 0.8 and 0.93 for cross-sectional and between 0.9 and 0.97 for longitudinal loadings. **Figure 4.15** shows the correlations within original and approximated cross-sectional and longitudinal loadings. Importantly, the cross-component correlations within the original and approximated loadings are similar. This indicates that not only are these coefficients estimable, but their covariance structure is preserved using approximated values. The relatively lower correlations between original and approximated cross-sectional loadings (diagonal elements in **Fig. 4.14A**) compared to the longitudinal ones (diagonal elements in **Fig. 4.14B**) may be attributable to the nature of the L-map, which captures individualized brain changes often unaffected by harmonization issues. The lower correlations in

cross-sectional loadings are also expected due to the use of cross-sectional ROI deviations rather than ROI volumes in the CCL-NMF implementation. To evaluate the efficacy of the approximated cross-sectional loadings in capturing associations with Alzheimer's disease, cognition, cardiovascular disease risk factors, and cognitive impairment progression, the associations presented in **Fig. 4.10** were examined using now the approximated loadings (**Figure 4.16**). The loading associations were effectively preserved, indicating that the approximated loadings can reliably be used instead of the original, thereby obviating the need to run an entirely new CCL-NMF. These estimation models have been made available in NiChart (https://cbica.github.io/NiChart_Project/).

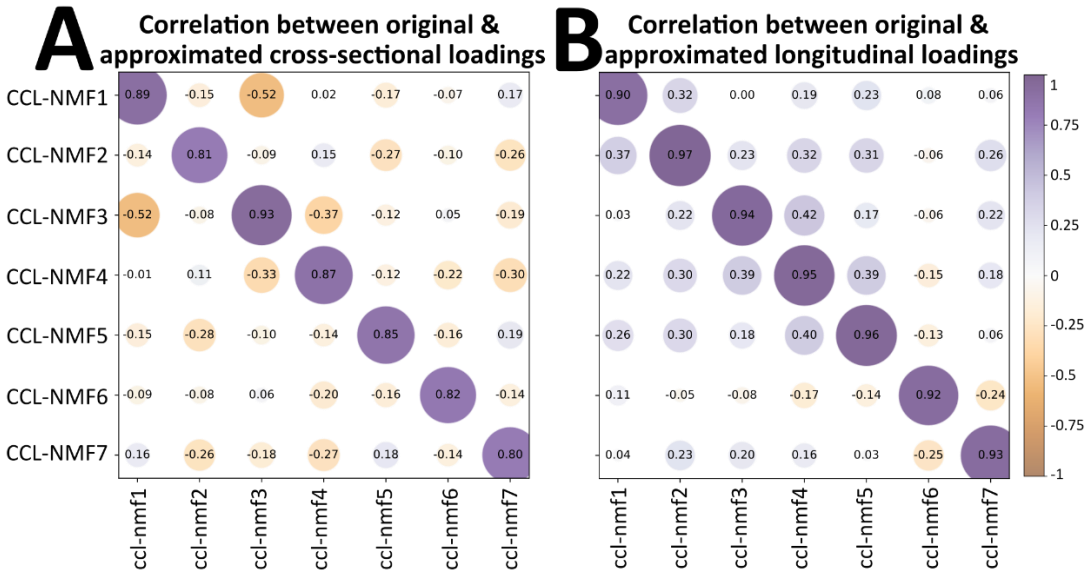


Figure 4.14: Spearman correlations between original and approximated A) cross-sectional and B) longitudinal loading coefficients. *Upper-case (lower-case) letters stand for original (approximated) loading coefficients.*

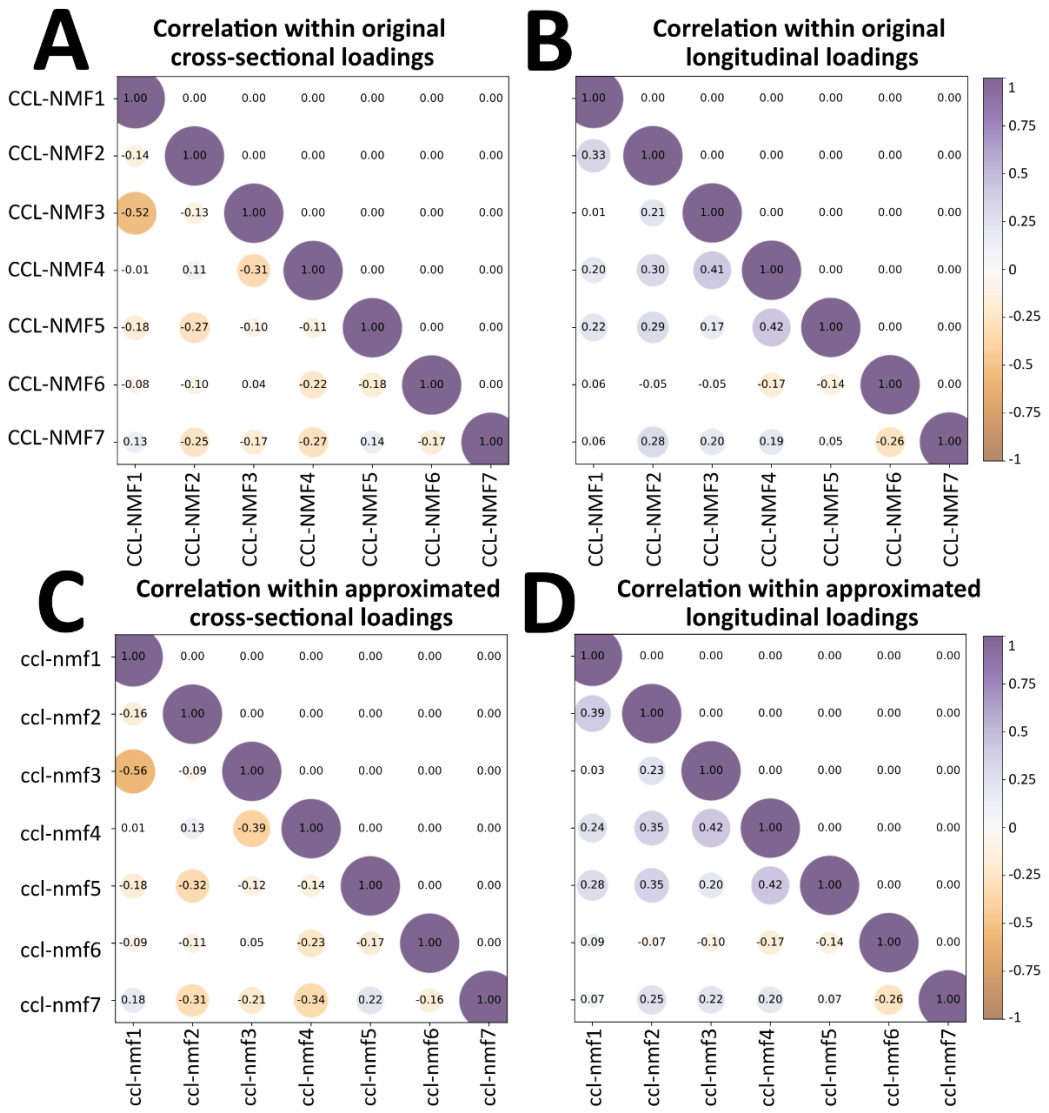


Figure 4.15: Spearman correlations within original (top) and approximated (bottom) cross-sectional (left) and longitudinal (right) loading coefficients. *Upper-case (lower-case) letters stand for original (approximated) loading coefficients.*

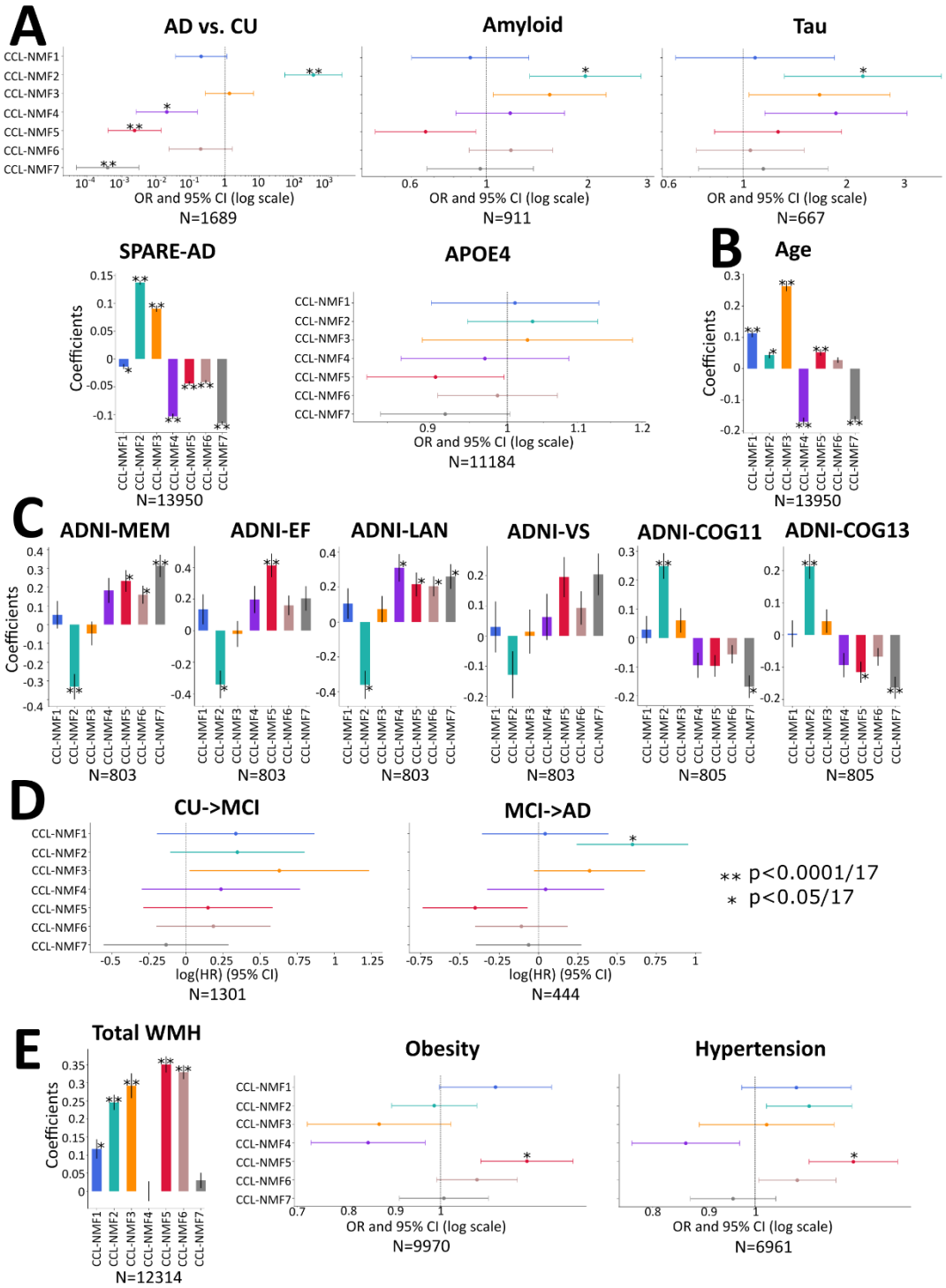


Figure 4.16: Associations of approximated cross-sectional CCL-NMF loadings with A) AD-specific measures, B) age, C) cognition, D) cognitive impairment progression, and E) CVD risk factors. *P-values were adjusted using Bonferroni correction to account for the number of comparisons (N=17), controlling for type I errors. The follow-up time for CU to MCI and MCI to AD conversion was 5.37±4.29 and 2.6±2.65 years, respectively. Age, sex, and study were included as covariates in all models. However, for APOE-ε4 (APOE4) carrier status, age was not included as a covariate, while for ADNI cognitive scores, the study was excluded, and education was included instead. Additionally, education was included as a covariate in models assessing cognitive decline progression (Part D). N indicates the sample size in each graph. Biomarker and CVD risk factor status is defined as explained in the previous chapter. Abbreviations: SPARE-AD, spatial pattern of abnormality for recognition of early Alzheimer's disease; WMH, white matter hyperintensities; CU, cognitively unimpaired; MCI, mild cognitive*

impairment; ADNI, Alzheimer's Disease Neuroimaging Initiative; ADNI-MEM, ADNI memory composite; ADNI-VS, ADNI visuospatial functioning composite; ADNI-LAN, ADNI language composite; ADNI-EF, ADNI executive function composite; ADAS-COG, Alzheimer's disease assessment scale-cognitive subscale.

4.3.3 Discussion

Brain aging exhibits high heterogeneity due to genetic predispositions, lifestyle, environmental influences, and other risk factors that cumulatively contribute to the gradual onset of neuropathologies. MRI, a widely accessible imaging technique, enables detailed assessment of macroscopic neurodegeneration—such as brain atrophy and small vessel ischemic changes. While MRI does not directly capture underlying neuropathologies, the addition of machine learning techniques captures patterns of neurodegeneration associated with such processes[202][203][204][205]. Such patterns, extracted via a coupled cross-sectional and longitudinal NMF scheme, are used to form the dimensions of a succinct yet expressive neuroanatomical coordinate system that encapsulates the heterogeneous changes in the brain with aging and disease-related neurodegenerative processes. By incorporating longitudinal data, this methodology effectively captures trajectories of brain changes over time alongside the cumulative cross-sectional effects. This temporal dimension is essential for identifying predominant patterns that form the foundation of the coordinate system, allowing the model to accurately represent the progressive nature of brain aging and associated neuropathologies. Notably, this approach delineates individual characteristics through multiple latent continuous variables, allowing for the concurrent expression of diverse patterns, thus circumventing the limitations of rigid classification into mutually exclusive categorical subtypes.

Leveraging data from the large, diverse, multi-cohort iSTAGING consortium with a predominately cognitively unimpaired sample and the novel CCL-NMF methodology, the neuroanatomical heterogeneity of aging was explored. Seven distinct and reproducible brain aging components were identified. Among these, **CCL-NMF2**, which primarily captured medial temporal lobe atrophy, displayed characteristics indicative of Alzheimer's disease, such as a strong link with amyloid and tau deposition, APOE4 alleles, cognitive decline, as expected, and can predict clinical progression from MCI to AD. **CCL-NMF5**, marked by perisylvian atrophy, demonstrated strong associations with vascular pathology. **CCL-NMF3**, strongly associated with advanced age, exhibited moderate correlations with amyloid deposition, cognitive decline, and WMH and was more prominently expressed within the aging population, suggesting that it predominantly reflected age-related effects rather than a distinct pathological process. Further investigation into the associations between these components and clinical traits, other aging-related brain disorder markers, and genetic data

will be essential for refining the profiles and implications of the derived components.

The CCL-NMF representation of brain aging was evaluated by comparing it to an alternative machine learning model, Surreal-GAN. These methodologies are fundamentally distinct: Surreal-GAN employs a GAN-based semi-supervised clustering approach to parse heterogeneity, whereas CCL-NMF uses deep-learning-based normative modeling combined with longitudinal brain change maps to identify dimensions of heterogeneity through non-negative matrix factorization. A key methodological difference is that Surreal-GAN relies exclusively on baseline cross-sectional measures, while CCL-NMF integrates cross-sectional and longitudinal measures, allowing the two data types to inform one another. This approach addresses a common limitation of cross-sectional methods by ensuring that patterns of neurodegeneration inferred from cross-sectional data align with those observed in longitudinal progression. Both methods were applied to the same dataset, the iSTAGING discovery dataset, to examine aging-related brain atrophy.

The components derived from CCL-NMF largely corresponded to the five dimensions of atrophy identified using Surreal-GAN, while also offering an expanded representation. Predictive models utilizing CCL-NMF loadings consistently outperformed those using R-indices, indicating the richer representation provided by CCL-NMF, which incorporates both cross-sectional and longitudinal data. By capturing temporal dynamics, CCL-NMF extends beyond traditional heterogeneity models limited to static cross-sectional data, offering a more comprehensive understanding of brain aging and associated pathologies.

In this application, the limited availability of longitudinal data compared to cross-sectional data (with a longitudinal-to-cross-sectional ratio of approximately 1:13) diminished the contribution of longitudinal information to component derivation. Despite using the α coefficient to balance the contributions of the two data types, the CCL-NMF components were predominantly influenced by cross-sectional data. However, in future applications involving datasets with a higher proportion of longitudinal data, the CCL-NMF components are expected to become increasingly refined, further improving their predictive utility and surpassing the patterns identified by Surreal-GAN.

The proposed machine learning framework operates in two primary phases: first, the C-map and L-map are estimated, followed by applying a mutually constrained NMF. This methodology produces a shared dictionary matrix for the

two data types, along with loading coefficient matrices specific to each data type. While the framework incorporates well-established techniques at various stages, implementing the entire process from start to finish can pose challenges for users. To improve accessibility and practical usability, regression models were developed to accurately predict the loading coefficients generated by the CCL-NMF model. This approach enables users to estimate these coefficients directly from their datasets without requiring expertise executing the two-step framework. Notably, the regression models operate on raw data, eliminating the need for users to harmonize their data before input. By addressing the complexity of the model while providing a user-friendly alternative, CCL-NMF supports wider adoption and application of its predictive capabilities across diverse research contexts.

CCL-NMF provides a robust and flexible framework for identifying heterogeneous patterns of brain changes by utilizing both cross-sectional and longitudinal data, surpassing the limitations of traditional heterogeneity models that rely solely on static, cross-sectional data. The model accommodates the co-expression of diverse patterns within the same individual. By deriving individualized expression levels across these patterns, this approach facilitates the development of personalized therapies tailored to specific patients, enabling more targeted and effective treatments. Furthermore, the model demonstrates computational efficiency, ensuring broad accessibility and usability. It does not require a predefined number of components, instead allowing the data to determine the model's complexity, which supports the exploration of higher-dimensional systems. Although the current application focuses on aging-related gray matter atrophy, the generic formulation of the approach enables its adaptability to any brain disorder characterized by monotonic progression over time. Additionally, as NMF has been widely applied across various domains, the CCL-NMF framework benefits from the established familiarity within the research community, facilitating integration and enrichment with insights from other NMF-based models.

The model's primary strength lies in its ability to identify patterns guided by longitudinal information; however, this reliance on temporal data presents limitations in contexts where such data is unavailable. Furthermore, the non-negativity constraint intrinsic to NMF, which facilitates part-based representations, requires single-signed input matrices. In many real-world scenarios, data may inherently include both positive and negative values, or such values may arise due to noise introduced during data acquisition. In the former case, such as functional connectivity data, this requirement may limit the model's applicability. In the latter case, as with noisy longitudinal data that often generates mixed-signed loading maps, it becomes necessary to eliminate

elements with opposite signs, resulting in input matrices containing numerous zero entries. This transformation can lead to a loss of valuable information and poses challenges for accurately representing brain regions with low or sparse signal intensities, potentially compromising the robustness and interpretability of the derived components.

Future research should focus on applying the approach to voxel-wise data to address the limitations of the ROI-based analysis utilized in the current study. Transitioning to voxel-wise analysis would enable the capture of more localized and granular patterns of brain activity, which are often obscured in ROI-wise data due to spatial averaging. This refinement would facilitate a more detailed examination of the spatial heterogeneity existing in neuroimaging data, offering new insights into the complex processes associated with brain aging. Moreover, the model could be extended to examine heterogeneity in amyloid and tau deposition, key biomarkers of neurodegenerative processes, to elucidate their distinct and overlapping contributions to brain aging.

4.4 Conclusion

This chapter introduced a novel machine learning-based approach designed to disentangle the heterogeneity of brain aging, leveraging both cross-sectional and longitudinal data. Applied to structural MRI data from a large aging cohort, the method identified distinct, reproducible, and clinically relevant components associated with brain aging. This approach surpasses traditional cross-sectional methods by integrating temporal dynamics, enabling more nuanced insights into the progressive nature of complex biological processes underlying aging. The findings suggest this approach could help tailor interventions based on individual profiles, advancing personalized therapeutic strategies.

5 Conclusions and future directions

The global increase in life expectancy has led to a significant rise in the older population, projected to reach 1.5 billion by 2050, representing a substantial demographic shift[209]. Concurrently, the incidence of neurodegenerative diseases, particularly Alzheimer's disease, is escalating, imposing a high socioeconomic burden on healthcare systems and families due to the need for long-term care, medical interventions, and caregiver support. According to the World Health Organization, dementia affects approximately 50 million people worldwide, with Alzheimer's disease as the leading cause, accounting for 60–70% of cases[210]. This number is projected to nearly triple by 2050. This demographic shift underscores the urgent need for strategies aimed at delaying brain aging and promoting healthy cognitive aging.

Brain aging is a highly complex process influenced by many factors that cause heterogeneous brain changes. Lifestyle factors, such as diet, physical activity, social engagement, and cognitive stimulation, significantly impact brain health throughout a person's lifetime. For instance, regular exercise and a balanced diet have been shown to protect against brain atrophy[211], while chronic stress and lack of mental stimulation can contribute to accelerated cognitive decline[212]. Environmental factors, including exposure to pollutants and toxins, education, and socioeconomic status, also influence brain health[213], either protecting against or accelerating brain aging.

Additionally, genetics play a crucial role in brain aging, with certain genetic profiles linked to a higher susceptibility to neurodegenerative diseases such as Alzheimer's disease. For example, the APOE ε4 allele is associated with an increased risk for AD[214]. This genetic diversity means that some individuals may be more resilient to age-related brain changes, while others experience accelerated decline, adding another layer of complexity. Finally, co-existing conditions such as cardiovascular disease, diabetes, and cerebrovascular disease interact with the aging process and can exacerbate brain aging[215]. The presence of amyloid plaques, tau tangles, and white matter lesions further complicates aging patterns, as these pathologies can overlap with typical aging processes, making it challenging to distinguish between normal aging and early signs of neurodegenerative disease.

Due to the unique interplay of various factors in each individual, brain aging exhibits significant heterogeneity, underscoring the need for personalized approaches to studying and treating age-related neurodegeneration and cognitive decline. Understanding this variability allows researchers to identify

distinct aging trajectories and develop targeted therapies tailored to each individual's influences. This thesis addresses this challenge by investigating the complex brain changes associated with aging that contribute to cognitive decline and AD. By advancing state-of-the-art machine learning techniques and leveraging large-scale datasets, it identifies distinct imaging patterns linked to different brain aging trajectories. Specifically, it aspires to disentangle the neuroanatomical heterogeneity across the spectrum of brain aging, examining variability driven by AD-related degeneration, coexisting pathologies, and lifestyle, environmental, and genetic risk factors. Additionally, it utilizes the identified dimensions of brain changes to predict future cognitive decline and clinical progression, offering insights that may ultimately enhance early diagnosis, risk stratification, and intervention strategies in aging and neurodegenerative diseases.

The contributions of this thesis are twofold:

- It investigates the heterogeneity of neuroanatomical brain changes at early, asymptomatic stages, offering insights into early indicators of neurodegeneration that may inform preventive strategies and personalized patient management.
- It advances existing methods for analyzing heterogeneity by developing a novel model that integrates cross-sectional and longitudinal data, allowing for a more accurate characterization of age-related brain changes.

Revealing Neuroanatomical Heterogeneity in Preclinical Aging through Advanced AI for Improved Early Diagnosis and Personalized Intervention.

The goal of this research is to address the limited understanding of neuroanatomical heterogeneity in the early phases of brain aging, particularly before the onset of clinical symptoms. While most existing studies focus on the heterogeneity of diagnosed neurodegenerative diseases[114][84][216][79][110], such as AD, this thesis shifts the emphasis to preclinical variability, aiming to uncover the diverse neuroanatomical trajectories that characterize aging in cognitively unimpaired individuals. This work fills a significant knowledge gap, as few studies have explored how individuals transition from normal aging to potential pathologic manifestation at the neural level—a gap largely attributable to the lack of sufficiently large-scale neuroimaging datasets and advanced modeling tools.

This thesis leverages artificial intelligence and advanced harmonization methods, along with large, diverse, multi-cohort datasets, to investigate neuroanatomical heterogeneity at early asymptomatic stages. Specifically, it

employs Smile-GAN to analyze large-scale cross-sectional T1- and T2-weighted MRI data from CU middle-to-late-aged individuals, aggregated by the iSTAGING consortium. Heterogeneity is investigated separately within four decade-long age intervals spanning from 45 to 85 years. The analysis identifies three distinct neuroanatomical subgroups that remain consistent across age decades: a typical aging group with low atrophy and white matter lesions and two accelerated aging subgroups—one associated with high cardiovascular disease risk, white matter disruption, and cerebral amyloid β deposition, and the other characterized by diffuse atrophy likely driven by lifestyle factors. These findings offer insights into differential susceptibilities to AD and other neurodegenerative conditions and underscore the significance of early intervention and tailored preventive strategies.

Developing an Innovative NMF-Based Framework Integrating Longitudinal and Cross-Sectional Data for Uncovering Brain Aging Heterogeneity

Moreover, this thesis enhances our understanding of the heterogeneous neurobiological processes implied in brain aging by developing a novel methodology. Current approaches primarily rely on cross-sectional data, which limits their ability to capture the dynamic nature of brain aging. This thesis addresses this gap by introducing CCL-NMF, which integrates population-based brain change maps alongside maps reflecting the dynamic progression of changes, allowing for a more nuanced view of brain aging. The proposed approach utilizes a mutually constrained NMF framework to delineate components that encapsulate distinct patterns of brain alterations derived from the combined analysis of cross-sectional and longitudinal data.

The proposed methodology broadly applies to conditions characterized by monotonically progressive brain changes. Here, it is specifically utilized to model aging-related atrophy heterogeneity using T1-weighted MRI data from a multi-cohort aging population having as reference a healthy middle-aged cohort both drawn from the iSTAGING consortium. This analysis reveals distinct neuroanatomical components associated with AD biomarkers, cognitive performance, cardiovascular risk, and cognitive impairment progression. Ultimately, by providing individualized expression levels across components through the CCL-NMF loadings, this approach differentiates from previous approaches, such as Smile-GAN categorizing individuals into rigid subtypes, thereby offering additional tools for personalized patient management and clinical trial stratification.

A comparison between the CCL-NMF and the state-of-the-art Surreal-GAN model demonstrates the advantages of CCL-NMF's integration of longitudinal

data, providing dynamic insights that complement the static nature of cross-sectional methods. Models utilizing CCL-NMF loadings exhibit superior predictive accuracy for AD and vascular markers than those using Surreal-GAN's R-indices. Additionally, to improve accessibility, regression models were developed to predict CCL-NMF loadings, allowing for application without the need to rerun the full model. This user-friendly feature promotes broader adoption across research contexts.

Future directions

Future steps will involve a comprehensive exploration of the proposed CCL-NMF methodology. This includes experimenting with integrating regularization terms that promote component sparsity, reduced overlap, and spatial contiguity. Additionally, techniques will be explored to address mixed-sign input data, aiming to mitigate the risk of zeroing out important features and ensuring a more robust analysis of the underlying patterns in the dataset. The model will also be adapted for application to voxel-wise data, enabling the detection of more detailed and localized brain patterns often obscured in ROI-based approaches due to spatial averaging. This voxel-wise analysis will facilitate the identification of distinct patterns of brain change that do not conform to traditional ROI boundaries, thereby revealing novel regions or networks involved in the aging process.

Another step will be integrating genetic information into these analyses by leveraging genetically informed voxel-based atlases, such as the Allen Brain Atlases[217]. Mapping gene expression directly onto voxel-based anatomic data provides a powerful framework for investigating how genetic factors interact with brain anatomy, driving structural or functional changes associated with aging and neurodegeneration and contributing to individual variability in aging trajectories. By uncovering genetically informed neuroanatomical patterns, this analysis could provide valuable insights into the genetic underpinnings of resilience vs. vulnerability to neurodegenerative diseases and differential cognitive decline, paving the way for more precise predictive models and personalized interventions in aging populations.

Finally, while current work focuses primarily on structural MRI data, future work can incorporate additional imaging modalities, such as diffusion tensor imaging for white matter integrity, functional MRI for brain activity, and positron emission tomography for amyloid or tau deposition. This multimodal integration will allow CCL-NMF to capture a more comprehensive and holistic view of brain aging processes, disentangling structural, functional, and molecular changes.

Last but not least, to enhance the characterization of the identified components, future research could explore the links between these components and biomarkers for neurodegenerative diseases beyond Alzheimer's, such as Parkinson's disease, Lewy body dementia, and frontotemporal dementia. Expanding biomarker analysis to include α-synuclein, tauopathies, and TDP-43 pathology could reveal shared or unique brain patterns associated with these diseases. This broader scope will help identify overlapping mechanisms across neurodegenerative conditions, uncover subgroups at risk for mixed pathologies, and enhance understanding of how concurrent neurodegenerative processes interact, impacting brain aging and cognitive decline in complex ways. Finally, investigating the significant role of brain resilience - the ability to maintain cognitive function despite age-related neuroanatomical changes or pathology – in brain aging heterogeneity will improve early intervention strategies and provide insights into mechanisms that protect against neurodegeneration.

References

[1] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging.," *Cell*, vol. 153, no. 6, pp. 1194–1217, Jun. 2013.

[2] A. F. Fotenos, A. Z. Snyder, L. E. Girton, J. C. Morris, and R. L. Buckner, "Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD," *Neurology*, vol. 64, no. 6, pp. 1032 LP – 1039, Mar. 2005.

[3] A. B. Storsve *et al.*, "Differential Longitudinal Changes in Cortical Thickness, Surface Area and Volume across the Adult Life Span: Regions of Accelerating and Decelerating Change," *J. Neurosci.*, vol. 34, no. 25, pp. 8488 LP – 8498, Jun. 2014.

[4] L. A. Grajauskas, W. Siu, G. Medvedev, H. Guo, R. C. N. D'Arcy, and X. Song, "MRI-based evaluation of structural degeneration in the ageing brain: Pathophysiology and assessment," *Ageing Res. Rev.*, vol. 49, pp. 67–82, 2019.

[5] I. J. Bennett and D. J. Madden, "Disconnected aging: cerebral white matter integrity and age-related differences in cognition.," *Neuroscience*, vol. 276, pp. 187–205, Sep. 2014.

[6] H. Liu *et al.*, "Aging of cerebral white matter.," *Ageing Res. Rev.*, vol. 34, pp. 64–76, Mar. 2017.

[7] E. Varangis, C. G. Habeck, Q. R. Razlighi, and Y. Stern, "The Effect of Aging on Resting State Connectivity of Predefined Networks in the Brain," *Front. Aging Neurosci.*, vol. 11, 2019.

[8] L. Farras-Permanyer *et al.*, "Age-related changes in resting-state functional connectivity in older adults.," *Neural Regen. Res.*, vol. 14, no. 9, pp. 1544–1555, Sep. 2019.

[9] D. L. Dickstein, D. Kabaso, A. B. Rocher, J. I. Luebke, S. L. Wearne, and P. R. Hof, "Changes in the structural complexity of the aged brain.," *Aging Cell*, vol. 6, no. 3, pp. 275–284, Jun. 2007.

[10] L. Svennerholm, K. Boström, and B. Jungbjer, "Changes in weight and compositions of major membrane components of human brain during the span of adult human life of Swedes.," *Acta Neuropathol.*, vol. 94, no. 4, pp. 345–352, Oct. 1997.

[11] R. I. Scahill, C. Frost, R. Jenkins, J. L. Whitwell, M. N. Rossor, and N. C. Fox, "A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging.," *Arch. Neurol.*, vol. 60, no. 7, pp. 989–994, Jul. 2003.

[12] D. Bartrés-Faz, I. C. Clemente, and C. Junqué, "[White matter changes and cognitive performance in aging].," *Rev. Neurol.*, vol. 33, no. 4, pp. 347–353, Aug. 2001.

[13] M. Nedergaard and S. A. Goldman, "Glymphatic failure as a final common pathway to dementia.," *Science*, vol. 370, no. 6512, pp. 50–56, Oct. 2020.

[14] E. L. Dennis and P. M. Thompson, "Functional Brain Connectivity Using

fMRI in Aging and Alzheimer's Disease," *Neuropsychol. Rev.*, vol. 24, no. 1, pp. 49–62, 2014.
[15] D. Tomasi and N. D. Volkow, "Aging and functional brain networks," *Mol. Psychiatry*, vol. 17, no. 5, pp. 549–558, 2012.
[16] K. Onoda, M. Ishihara, and S. Yamaguchi, "Decreased functional connectivity by aging is associated with cognitive decline.," *J. Cogn. Neurosci.*, vol. 24, no. 11, pp. 2186–2198, Nov. 2012.
[17] C.-C. Huang *et al.*, "Age-related changes in resting-state networks of a large sample size of healthy elderly.," *CNS Neurosci. Ther.*, vol. 21, no. 10, pp. 817–825, Oct. 2015.
[18] L. Geerligs, R. J. Renken, E. Saliasi, N. M. Maurits, and M. M. Lorist, "A Brain-Wide Study of Age-Related Changes in Functional Connectivity.," *Cereb. Cortex*, vol. 25, no. 7, pp. 1987–1999, Jul. 2015.
[19] R. F. Betzel, L. Byrge, Y. He, J. Goñi, X.-N. Zuo, and O. Sporns, "Changes in structural and functional connectivity among resting-state networks across the human lifespan.," *Neuroimage*, vol. 102 Pt 2, pp. 345–357, Nov. 2014.
[20] L. Geerligs, N. M. Maurits, R. J. Renken, and M. M. Lorist, "Reduced specificity of functional connectivity in the aging brain during task performance.," *Hum. Brain Mapp.*, vol. 35, no. 1, pp. 319–330, Jan. 2014.
[21] E. Varangis, Q. Razlighi, C. G. Habeck, Z. Fisher, and Y. Stern, "Between-network Functional Connectivity Is Modified by Age and Cognitive Task Domain," *J. Cogn. Neurosci.*, vol. 31, no. 4, pp. 607–622, Apr. 2019.
[22] T. Wyss-Coray, "Ageing, neurodegeneration and brain rejuvenation.," *Nature*, vol. 539, no. 7628, pp. 180–186, Nov. 2016.
[23] Y. Hou *et al.*, "Ageing as a risk factor for neurodegenerative disease," *Nat. Rev. Neurol.*, vol. 15, no. 10, pp. 565–581, 2019.
[24] R. C. Petersen, "Aging, Mild Cognitive Impairment, and Alzheimer's Disease," *Dementia*, vol. 18, no. November, pp. 789–805, 2000.
[25] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity.," *J. Intern. Med.*, vol. 256, no. 3, pp. 183–194, Sep. 2004.
[26] B. Winblad *et al.*, "Mild cognitive impairment--beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment.," *J. Intern. Med.*, vol. 256, no. 3, pp. 240–246, Sep. 2004.
[27] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.," *Alzheimers. Dement.*, vol. 7, no. 3, pp. 270–279, May 2011.
[28] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, "Mild cognitive impairment: a concept in evolution.," *J. Intern. Med.*, vol. 275, no. 3, pp. 214–228, Mar. 2014.
[29] E. M. Reiman *et al.*, "Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study.," *Nat. Commun.*, vol. 11, no. 1, p. 667, Feb.

2020.
[30] "2023 Alzheimer's disease facts and figures.," *Alzheimers. Dement.*, vol. 19, no. 4, pp. 1598–1695, Apr. 2023.
[31] R. Brookmeyer, S. Gray, and C. Kawas, "Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset.," *Am. J. Public Health*, vol. 88, no. 9, pp. 1337–1342, Sep. 1998.
[32] J. Dumurgier and S. Sabia, "[Epidemiology of Alzheimer's disease: latest trends].," *Rev. Prat.*, vol. 70, no. 2, pp. 149–151, Feb. 2020.
[33] A. Schäfer, P. Chaggar, T. B. Thompson, A. Goriely, and E. Kuhl, "Predicting brain atrophy from tau pathology: a summary of clinical findings and their translation into personalized models," *Brain Multiphysics*, vol. 2, p. 100039, 2021.
[34] R. Cacace, K. Sleegers, and C. Van Broeckhoven, "Molecular genetics of early-onset Alzheimer's disease revisited.," *Alzheimers. Dement.*, vol. 12, no. 6, pp. 733–748, Jun. 2016.
[35] A. Ward *et al.*, "Prevalence of apolipoprotein E4 genotype and homozygotes (APOE e4/4) among patients diagnosed with Alzheimer's disease: a systematic review and meta-analysis.," *Neuroepidemiology*, vol. 38, no. 1, pp. 1–17, 2012.
[36] W. J. Strittmatter *et al.*, "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 5, pp. 1977–1981, Mar. 1993.
[37] S. J. Andrews, A. E. Renton, B. Fulton-Howard, A. Podlesny-Drabiniok, E. Marcora, and A. M. Goate, "The complex genetic architecture of Alzheimer's disease: novel insights and future directions.," *EBioMedicine*, vol. 90, p. 104511, Apr. 2023.
[38] L. M. Bekris, C.-E. Yu, T. D. Bird, and D. W. Tsuang, "Genetics of Alzheimer disease.," *J. Geriatr. Psychiatry Neurol.*, vol. 23, no. 4, pp. 213–227, Dec. 2010.
[39] J. A. Hardy and G. A. Higgins, "Alzheimer's disease: the amyloid cascade hypothesis.," *Science*, vol. 256, no. 5054, pp. 184–185, Apr. 1992.
[40] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics.," *Science*, vol. 297, no. 5580, pp. 353–356, Jul. 2002.
[41] P. S. Aisen *et al.*, "The Future of Anti-Amyloid Trials.," *J. Prev. Alzheimer's Dis.*, vol. 7, no. 3, pp. 146–151, 2020.
[42] J. Sevigny *et al.*, "The antibody aducanumab reduces Aβ plaques in Alzheimer's disease," *Nature*, vol. 537, no. 7618, pp. 50–56, 2016.
[43] M. A. Mintun *et al.*, "Donanemab in Early Alzheimer's Disease.," *N. Engl. J. Med.*, vol. 384, no. 18, pp. 1691–1704, May 2021.
[44] F. Kametani and M. Hasegawa, "Reconsideration of Amyloid Hypothesis and Tau Hypothesis in Alzheimer's Disease.," *Front. Neurosci.*, vol. 12, p. 25, 2018.
[45] R. A. Sperling *et al.*, "The impact of amyloid-beta and tau on prospective cognitive decline in older individuals.," *Ann. Neurol.*, vol.

85, no. 2, pp. 181–193, Feb. 2019.
[46] T. J. Betthauser *et al.*, "Amyloid and tau imaging biomarkers explain cognitive decline from late middle-age.," *Brain*, vol. 143, no. 1, pp. 320–335, Jan. 2020.
[47] H. Braak, D. R. Thal, E. Ghebremedhin, and K. Del Tredici, "Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years.," *J. Neuropathol. Exp. Neurol.*, vol. 70, no. 11, pp. 960–969, Nov. 2011.
[48] W. R. Bell *et al.*, "Neuropathologic, genetic, and longitudinal cognitive profiles in primary age-related tauopathy (PART) and Alzheimer's disease.," *Alzheimers. Dement.*, vol. 15, no. 1, pp. 8–16, Jan. 2019.
[49] D. Strozyk *et al.*, "Contribution of vascular pathology to the clinical expression of dementia.," *Neurobiol. Aging*, vol. 31, no. 10, pp. 1710–1720, Oct. 2010.
[50] R. Y. Lo and W. J. Jagust, "Vascular burden and Alzheimer disease pathologic progression.," *Neurology*, vol. 79, no. 13, pp. 1349–1355, Sep. 2012.
[51] J. Attems and K. A. Jellinger, "The overlap between vascular disease and Alzheimer's disease - lessons from pathology," *BMC Med.*, vol. 12, no. 1, pp. 1–12, 2014.
[52] A. Soldan *et al.*, "White matter hyperintensities and CSF Alzheimer disease biomarkers in preclinical Alzheimer disease.," *Neurology*, vol. 94, no. 9, pp. e950–e960, Mar. 2020.
[53] X.-F. Meng *et al.*, "Midlife vascular risk factors and the risk of Alzheimer's disease: a systematic review and meta-analysis.," *J. Alzheimers. Dis.*, vol. 42, no. 4, pp. 1295–1310, 2014.
[54] C. DeCarli *et al.*, "Vascular Burden Score Impacts Cognition Independent of Amyloid PET and MRI Measures of Alzheimer's Disease and Vascular Brain Injury.," *J. Alzheimers. Dis.*, vol. 68, no. 1, pp. 187–196, 2019.
[55] J. L. Robinson *et al.*, "Non-Alzheimer's contributions to dementia and cognitive resilience in The 90+ Study.," *Acta Neuropathol.*, vol. 136, no. 3, pp. 377–388, Sep. 2018.
[56] A. Serrano-Pozo *et al.*, "Reactive glia not only associates with plaques but also parallels tangles in Alzheimer's disease.," *Am. J. Pathol.*, vol. 179, no. 3, pp. 1373–1384, Sep. 2011.
[57] D. Soulet and S. Rivest, "Microglia," *Curr. Biol.*, vol. 18, no. 12, pp. 506–508, 2008.
[58] M. Lyman, D. G. Lloyd, X. Ji, M. P. Vizcaychipi, and D. Ma, "Neuroinflammation: the role and consequences.," *Neurosci. Res.*, vol. 79, pp. 1–12, Feb. 2014.
[59] M. T. Heneka *et al.*, "Neuroinflammation in Alzheimer's disease," *Lancet Neurol.*, vol. 14, no. 4, pp. 388–405, 2015.
[60] Y. Yoshiyama *et al.*, "Synapse loss and microglial activation precede tangles in a P301S tauopathy mouse model.," *Neuron*, vol. 53, no. 3, pp. 337–351, Feb. 2007.
[61] N. Maphis *et al.*, "Reactive microglia drive tau pathology and contribute to the spreading of pathological tau in the brain.," *Brain*, vol. 138, no.

Pt 6, pp. 1738–1755, Jun. 2015.
[62] C. M. Karch and A. M. Goate, "Alzheimer's disease risk genes and mechanisms of disease pathogenesis.," *Biol. Psychiatry*, vol. 77, no. 1, pp. 43–51, Jan. 2015.
[63] H. Scheiblich, M. Trombly, A. Ramirez, and M. T. Heneka, "Neuroimmune Connections in Aging and Neurodegenerative Diseases," *Trends Immunol.*, vol. 41, no. 4, pp. 300–312, 2020.
[64] Q. Qin, Z. Teng, C. Liu, Q. Li, Y. Yin, and Y. Tang, "TREM2, microglia, and Alzheimer's disease.," *Mech. Ageing Dev.*, vol. 195, p. 111438, Apr. 2021.
[65] G. Blessed, B. E. Tomlinson, and M. Roth, "The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects.," *Br. J. Psychiatry*, vol. 114, no. 512, pp. 797–811, Jul. 1968.
[66] E. M. Arenaza-Urquijo and P. Vemuri, "Resistance vs resilience to Alzheimer disease: Clarifying terminology for preclinical studies.," *Neurology*, vol. 90, no. 15, pp. 695–703, Apr. 2018.
[67] Y. Stern, C. A. Barnes, C. Grady, R. N. Jones, and N. Raz, "Brain reserve, cognitive reserve, compensation, and maintenance: operationalization, validity, and mechanisms of cognitive resilience.," *Neurobiol. Aging*, vol. 83, pp. 124–129, Nov. 2019.
[68] E. H. Corder *et al.*, "Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease.," *Nat. Genet.*, vol. 7, no. 2, pp. 180–184, Jun. 1994.
[69] P. Vemuri *et al.*, "Cognitive reserve and Alzheimer's disease biomarkers are independent determinants of cognition.," *Brain*, vol. 134, no. Pt 5, pp. 1479–1492, May 2011.
[70] E. M. Arenaza-Urquijo, M. Wirth, and G. Chételat, "Cognitive reserve and lifestyle: moving towards preclinical Alzheimer's disease.," *Front. Aging Neurosci.*, vol. 7, p. 134, 2015.
[71] B. A. Mander, J. R. Winer, W. J. Jagust, and M. P. Walker, "Sleep: A Novel Mechanistic Pathway, Biomarker, and Treatment Target in the Pathology of Alzheimer's Disease?," *Trends Neurosci.*, vol. 39, no. 8, pp. 552–566, Aug. 2016.
[72] J. J. Yang *et al.*, "Association of Healthy Lifestyles With Risk of Alzheimer Disease and Related Dementias in Low-Income Black and White Americans," *Neurology*, vol. 99, no. 9, p. e944 LP-e953, Aug. 2022.
[73] K. Ning, L. Zhao, W. Matloff, F. Sun, and A. W. Toga, "Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants," *Sci. Rep.*, vol. 10, no. 1, p. 10, 2020.
[74] W. J. Jagust, C. E. Teunissen, and C. DeCarli, "The complex pathway between amyloid β and cognition: implications for therapy," *Lancet Neurol.*, vol. 22, no. 9, pp. 847–857, 2023.
[75] L. Ferrucci, "The Baltimore Longitudinal Study of Aging (BLSA): A 50-Year-Long Journey and Plans for the Future," *Journals Gerontol. Ser. A*, vol. 63, no. 12, pp. 1416–1419, Dec. 2008.
[76] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative

(ADNI) Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201 LP – 209, Jan. 2010.
[77] K. L. Miller *et al.*, "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat. Neurosci.*, vol. 19, no. 11, pp. 1523–1536, 2016.
[78] R. Pomponio *et al.*, "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," *Neuroimage*, vol. 208, p. 116450, 2020.
[79] X. Zhang, E. C. Mormino, N. Sun, R. A. Sperling, M. R. Sabuncu, and B. T. T. Yeo, "Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 42, pp. E6535–E6544, 2016.
[80] A. L. Young *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," *Nat. Commun.*, vol. 9, no. 1, p. 4273, 2018.
[81] C. C. Price, J. J. Tanner, I. M. Schmalfuss, B. Brumback, K. M. Heilman, and D. J. Libon, "Dissociating Statistically-Determined Alzheimer's Disease/Vascular Dementia Neuropsychological Syndromes Using White and Gray Neuroradiological Parameters," *J. Alzheimer's Dis.*, vol. 48, pp. 833–847, 2015.
[82] G. Tosto, S. E. Monsell, S. E. Hawes, G. Bruno, and R. Mayeux, "Progression of Extrapyramidal Signs in Alzheimer's Disease: Clinical and Neuropathological Correlates," *J. Alzheimer's Dis.*, vol. 49, pp. 1085–1093, 2016.
[83] J. Nettiksimmons, C. DeCarli, S. Landau, and L. Beckett, "Biological heterogeneity in ADNI amnestic mild cognitive impairment," *Alzheimer's Dement.*, vol. 10, no. 5, pp. 511-521.e1, 2014.
[84] K. Poulakis *et al.*, "Heterogeneous patterns of brain atrophy in Alzheimer's disease.," *Neurobiol. Aging*, vol. 65, pp. 98–108, May 2018.
[85] S. Jeon *et al.*, "Topographical Heterogeneity of Alzheimer's Disease Based on MR Imaging, Tau PET, and Amyloid PET.," *Front. Aging Neurosci.*, vol. 11, p. 211, 2019.
[86] J. Wen *et al.*, "Subtyping Brain Diseases from Imaging Data," in *Machine Learning for Brain Disorders*, O. Colliot, Ed. New York, NY: Springer US, 2023, pp. 491–510.
[87] A. F. Marquand, T. Wolfers, M. Mennes, J. Buitelaar, and C. F. Beckmann, "Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders," *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 1, no. 5, pp. 433–447, 2016.
[88] A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann, "Conceptualizing mental disorders as deviations from normative functioning," *Mol. Psychiatry*, vol. 24, no. 10, pp. 1415–1424, 2019.
[89] S. Rutherford *et al.*, "The normative modeling framework for computational psychiatry," *Nat. Protoc.*, vol. 17, no. 7, pp. 1711–1734, 2022.
[90] S. Rutherford *et al.*, "Evidence for embracing normative modeling," *Elife*, vol. 12, pp. 1–24, 2023.

[91] T. J. Cole, "The development of growth references and growth charts.," *Ann. Hum. Biol.*, vol. 39, no. 5, pp. 382–394, Sep. 2012.
[92] R. Ge *et al.*, "Normative modelling of brain morphometry across the lifespan with CentileBrain: algorithm benchmarking and model optimisation," *Lancet Digit. Heal.*, vol. 6, no. 3, pp. e211–e221, 2024.
[93] A. F. Marquand, I. Rezek, J. Buitelaar, and C. F. Beckmann, "Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies," *Biol. Psychiatry*, vol. 80, no. 7, pp. 552–561, 2016.
[94] R. Dinga, C. J. Fraza, J. M. M. Bayer, S. M. Kia, C. F. Beckmann, and A. F. Marquand, "Normative modeling of neuroimaging data using generalized additive models of location scale and shape," *bioRxiv*, 2021.
[95] T. R. Insel, "Mental Disorders in Childhood: Shifting the Focus From Behavioral Symptoms to Neurodevelopmental Trajectories," *JAMA*, vol. 311, no. 17, pp. 1727–1728, May 2014.
[96] G. B. Karas *et al.*, "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease," *Neuroimage*, vol. 23, no. 2, pp. 708–716, 2004.
[97] S. Verdi *et al.*, "Revealing Individual Neuroanatomical Heterogeneity in Alzheimer Disease Using Neuroanatomical Normative Modeling," *Neurology*, vol. 100, no. 24, pp. E2442–E2453, 2023.
[98] X. Wang, R. Zhou, K. Zhao, A. Leow, Y. Zhang, and L. He, "Normative Modeling Via Conditional Variational Autoencoder and Adversarial Learning to Identify Brain Dysfunction in Alzheimer's Disease," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2023-April, 2023.
[99] R. Dimitrova *et al.*, "Heterogeneity in Brain Microstructural Development Following Preterm Birth," *Cereb. Cortex*, vol. 30, no. 9, pp. 4800–4810, 2020.
[100] R. C. Gur *et al.*, "Neurocognitive Growth Charting in Psychosis Spectrum Youths," *JAMA Psychiatry*, vol. 71, no. 4, pp. 366–374, Apr. 2014.
[101] T. Wolfers *et al.*, "Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models," *JAMA Psychiatry*, vol. 75, no. 11, pp. 1146–1155, Nov. 2018.
[102] T. Wolfers *et al.*, "Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder.," *Hum. Brain Mapp.*, vol. 42, no. 8, pp. 2546–2555, Jun. 2021.
[103] D. Kessler, M. Angstadt, and C. Sripada, "Growth Charting of Brain Connectivity Networks and the Identification of Attention Impairment in Youth," *JAMA Psychiatry*, vol. 73, no. 5, pp. 481–489, 2016.
[104] T. Wolfers *et al.*, "Refinement by integration: aggregated effects of multimodal imaging markers on adult ADHD.," *J. Psychiatry Neurosci.*, vol. 42, no. 6, pp. 386–394, Nov. 2017.
[105] T. Wolfers, C. F. Beckmann, M. Hoogman, J. K. Buitelaar, B. Franke, and A. F. Marquand, "Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models.," *Psychol. Med.*, vol. 50, no. 2, pp. 314–323, Jan. 2020.
[106] M. Zabihi *et al.*, "Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models.," *Biol. psychiatry.*

Cogn. Neurosci. neuroimaging, vol. 4, no. 6, pp. 567–578, Jun. 2019.
[107] M. Zabihi *et al.*, "Fractionating autism based on neuroanatomical normative modeling.," *Transl. Psychiatry*, vol. 10, no. 1, p. 384, Nov. 2020.
[108] E. Varol, A. Sotiras, and C. Davatzikos, "HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework," *Neuroimage*, vol. 145, pp. 346–364, 2017.
[109] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.
[110] A. Dong, N. Honnorat, B. Gaonkar, and C. Davatzikos, "CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns.," *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 612–621, Feb. 2016.
[111] A. Myronenko and X. Song, "Point set registration: coherent point drift.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.
[112] D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA: Springer US, 2009, pp. 659–663.
[113] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.
[114] Z. Yang *et al.*, "A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure," *Nat. Commun.*, vol. 12, no. 1, p. 7065, 2021.
[115] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014, vol. 27.
[116] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
[117] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.
[118] P. Baldi, "Autoencoders," in *Deep Learning in Science*, Cambridge University Press, 2021, pp. 71–98.
[119] A. Rokem, "Detecting brain anomalies with autoencoders," *Nat. Comput. Sci.*, vol. 1, no. 9, pp. 569–570, 2021.
[120] O. Trofimova *et al.*, "Brain tissue properties link cardio-vascular risk factors, mood and cognitive performance in the CoLaus|PsyCoLaus epidemiological cohort," *Neurobiol. Aging*, vol. 102, pp. 50–63, 2021.
[121] B. Dubois *et al.*, *Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria*, vol. 12, no. 3. 2016.
[122] S.-H. Wan, M. W. Vogel, and H. H. Chen, "Pre-clinical diastolic dysfunction.," *J. Am. Coll. Cardiol.*, vol. 63, no. 5, pp. 407–416, Feb. 2014.
[123] M. Habes *et al.*, "The Brain Chart of Aging : Machine-learning analytics reveals links between brain aging , white matter disease , amyloid burden , and cognition in the iSTAGING consortium of 10 , 216 harmonized MR scans," *Alzheimer's Dement. J. Alzheimer's Assoc.*, vol.

17, no. 1, pp. 89–102, 2021.
[124] G. Hwang *et al.*, "Disentangling Alzheimer's disease neurodegeneration from typical brain aging using MRI and machine learning," *Alzheimer's Dement.*, vol. 17, no. S4, 2021.
[125] N. J. Tustison, P. A. Cook, and J. C. Gee, "N4Itk: Improved N3 Bias Correction," vol. 29, no. 6, pp. 1310–1320, 2011.
[126] J. Doshi, G. Erus, O. Yangming, B. Gaonkar, and C. Davatzikos, "Multi-Atlas Skull- Stripping.," *Acad. Radiol.*, vol. 20, no. 12, pp. 1566–1576, 2013.
[127] J. Doshi *et al.*, "MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection," *Neuroimage*, vol. 127, pp. 186–195, 2016.
[128] J. Doshi, G. Erus, M. Habes, and C. Davatzikos, "DeepMRSeg: A convolutional deep neural network for anatomy and abnormality segmentation on MR images."
[129] J. Wen *et al.*, "Characterizing Heterogeneity in Neuroimaging, Cognition, Clinical Symptoms, and Genetics Among Patients With Late-Life Depression.," *JAMA psychiatry*, vol. 79, no. 5, pp. 464–474, May 2022.
[130] I. Skampardoni *et al.*, "Genetic and Clinical Correlates of AI-Based Brain Aging Patterns in Cognitively Unimpaired Individuals," *JAMA Psychiatry*, vol. 81, no. 5, pp. 456–467, May 2024.
[131] Y. Noh *et al.*, "Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs.," *Neurology*, vol. 83, no. 21, pp. 1936–1944, Nov. 2014.
[132] M. Ten Kate *et al.*, "Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline.," *Brain*, vol. 141, no. 12, pp. 3443–3456, Dec. 2018.
[133] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses.," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007.
[134] K. Watanabe, E. Taskesen, A. Van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat. Commun.*, vol. 8, no. 1, pp. 1–10, 2017.
[135] A. Buniello *et al.*, "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019.
[136] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: A tool for genome-wide complex trait analysis," *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, 2011.
[137] J. Wen *et al.*, "Neuroimaging-AI Endophenotypes of Brain Diseases in the General Population: Towards a Dimensional System of Vulnerability," *medRxiv*, p. 2023.08.16.23294179, Jan. 2023.
[138] L. Mori, S., Wakana, S., van Zijl, P., and Nagae-Poetscher, "MRI Atlas of Human White Matter.," *AJNR: American Journal of Neuroradiology*, vol. 27, no. 6. pp. 1384–1385, Jun-2006.
[139] J. Ashburner and K. J. Friston, "Voxel-Based Morphometry—The Methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[140] I. C. Wright *et al.*, "A Voxel-Based Method for the Statistical Analysis of Gray and White Matter Density Applied to Schizophrenia," *Neuroimage*, vol. 2, no. 4, pp. 244–252, 1995.
[141] C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick, "Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy," *Neuroimage*, vol. 14, no. 6, pp. 1361–1369, 2001.
[142] D. Ostwald, S. Schneider, R. Bruckner, and L. Horvath, "Random field theory-based p-values: A review of the SPM implementation," 2018.
[143] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding.," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[144] Z. Zhang and J. Wang, "MLLE: Modified Locally Linear Embedding Using Multiple Weights," in *Advances in Neural Information Processing Systems*, 2006, vol. 19.
[145] G. Muniz-Terrera, A. Van Den Hout, R. A. Rigby, and D. M. Stasinopoulos, "Analysing cognitive test data: Distributions and non-parametric random effects," *Stat. Methods Med. Res.*, vol. 25, no. 2, pp. 741–753, 2016.
[146] X. Liu, "Chapter 3 - Linear mixed-effects models," in *Methods and Applications of Longitudinal Data Analysis*, X. Liu, Ed. Oxford: Academic Press, 2016, pp. 61–94.
[147] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, May 1995.
[148] C. Davatzikos, F. Xu, Y. An, Y. Fan, and S. M. Resnick, "Longitudinal progression of Alzheimers-like patterns of atrophy in normal older adults: The SPARE-AD index," *Brain*, vol. 132, no. 8, pp. 2026–2035, 2009.
[149] M. Habes *et al.*, "Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns.," *Transl. Psychiatry*, vol. 6, no. 4, p. e775, Apr. 2016.
[150] H. Eavani *et al.*, "Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods," *Neurobiol. Aging*, vol. 71, pp. 41–50, 2018.
[151] M. Habes *et al.*, "White matter hyperintensities and imaging patterns of brain ageing in the general population," *Brain*, vol. 139, no. 4, pp. 1164–1179, 2016.
[152] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Dec. 1958.
[153] J. M. Bland and D. G. Altman, "The logrank test," *Bmj*, vol. 328, no. 7447, p. 1073, 2004.
[154] L. M. Shaw *et al.*, "Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects.," *Ann. Neurol.*, vol. 65, no. 4, pp. 403–413, Apr. 2009.
[155] I. Lopes Alves *et al.*, "Strategies to reduce sample sizes in Alzheimer's disease primary and secondary prevention trials using longitudinal

amyloid PET imaging," *Alzheimers. Res. Ther.*, vol. 13, no. 1, p. 82, 2021.
[156] S. M. Landau *et al.*, "Amyloid deposition, hypometabolism, and longitudinal cognitive decline.," *Ann. Neurol.*, vol. 72, no. 4, pp. 578–586, Oct. 2012.
[157] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Jan. 1971.
[158] L. Hubert and P. Arabie, "Comparing partitions," *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985.
[159] B. Zhao *et al.*, "Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706).," *Mol. Psychiatry*, vol. 26, no. 8, pp. 3943–3955, Aug. 2021.
[160] D. van der Meer *et al.*, "Understanding the genetic determinants of the brain with MOSTest.," *Nat. Commun.*, vol. 11, no. 1, p. 3512, Jul. 2020.
[161] E. Persyn, K. B. Hanscombe, J. M. M. Howson, C. M. Lewis, M. Traylor, and H. S. Markus, "Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants.," *Nat. Commun.*, vol. 11, no. 1, p. 2175, May 2020.
[162] T. J. Hoffmann *et al.*, "Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation.," *Nat. Genet.*, vol. 49, no. 1, pp. 54–64, Jan. 2017.
[163] "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease.," *Nat. Genet.*, vol. 43, no. 4, pp. 339–344, Mar. 2011.
[164] C. Gouveia, E. Gibbons, N. Dehghani, J. Eapen, R. Guerreiro, and J. Bras, "Genome-wide association of polygenic risk extremes for Alzheimer's disease in the UK Biobank.," *Sci. Rep.*, vol. 12, no. 1, p. 8404, May 2022.
[165] S. E. Graham *et al.*, "The power of genetic diversity in genome-wide association studies of lipids.," *Nature*, vol. 600, no. 7890, pp. 675–679, Dec. 2021.
[166] T. G. Richardson *et al.*, "Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis.," *PLoS Med.*, vol. 17, no. 3, p. e1003062, Mar. 2020.
[167] C. Grady, S. Sarraf, C. Saverino, and K. Campbell, "Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks.," *Neurobiol. Aging*, vol. 41, pp. 159–172, May 2016.
[168] M. C. Power *et al.*, "Midlife and late-life vascular risk factors and white matter microstructural integrity: The atherosclerosis risk in communities neurocognitive study," *J. Am. Heart Assoc.*, vol. 6, no. 5, 2017.
[169] Y. Hannawi *et al.*, "Hypertension Is Associated with White Matter Disruption in Apparently Healthy Middle-Aged Individuals," *Am. J. Neuroradiol.*, Nov. 2018.
[170] T. M. Wassenaar, K. Yaffe, Y. D. van der Werf, and C. E. Sexton, "Associations between modifiable risk factors and white matter of the

aging brain: insights from diffusion tensor imaging studies," *Neurobiol. Aging*, vol. 80, pp. 56–70, 2019.
[171] M. Muller, A. P. A. Appelman, Y. van der Graaf, K. L. Vincken, W. P. T. M. Mali, and M. I. Geerlings, "Brain atrophy and cognition: Interaction with cerebrovascular pathology?," *Neurobiol. Aging*, vol. 32, no. 5, pp. 885–893, 2011.
[172] Z. Yang *et al.*, "Brain aging patterns in a large and diverse cohort of 49,482 individuals," *Nat. Med.*, vol. 30, no. 10, pp. 3015–3026, 2024.
[173] Z. Yang, J. Wen, and C. Davatzikos, "Surreal-GAN:Semi-Supervised Representation Learning via GAN for uncovering heterogeneous disease-related imaging patterns," in *International Conference on Learning Representations*, 2022.
[174] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values†," *Environmetrics*, vol. 5, pp. 111–126, 1994.
[175] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[176] D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems*, 2000, vol. 13.
[177] H. Kim and H. Park, "Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008.
[178] H. Van Hamme, "An On-Line NMF Model for Temporal Pattern Learning: Theory with Application to Automatic Speech Recognition," in *Latent Variable Analysis and Signal Separation*, 2012, pp. 306–313.
[179] D. Tsarev, M. Petrovskiy, and I. Mashechkin, "Using NMF-based text summarization to improve supervised and unsupervised classification," in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 185–189.
[180] D. Klötzl *et al.*, "NMF-Based Analysis of Mobile Eye-Tracking Data," in *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, 2024.
[181] N. Li, S. Wang, H. Li, and Z. Li, "SAC-NMF-Driven Graphical Feature Analysis and Applications," *Mach. Learn. Knowl. Extr.*, vol. 2, no. 4, pp. 630–646, 2020.
[182] E. Cao, K. Cao, K. Feng, and J. Wang, "NMF based image sequence analysis and its application in gait recognition," *CCF Trans. Pervasive Comput. Interact.*, vol. 2, no. 2, pp. 86–96, 2020.
[183] T. Liefeld *et al.*, "NMFClustering: Accessible NMF-based clustering utilizing GPU acceleration.," *bioRxiv : the preprint server for biology*. United States, Jun-2023.
[184] E. Mejía-Roa, D. Tabas-Madrid, J. Setoain, C. García, F. Tirado, and A. Pascual-Montano, "NMF-mGPU: non-negative matrix factorization on multi-GPU systems.," *BMC Bioinformatics*, vol. 16, p. 43, Feb. 2015.
[185] A. Sotiras, S. M. Resnick, and C. Davatzikos, "Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization,"

Neuroimage, vol. 108, pp. 1–16, 2015.
[186] A. Sotiras, J. B. Toledo, R. E. Gur, R. C. Gur, T. D. Satterthwaite, and C. Davatzikos, "Patterns of coordinated cortical remodeling during adolescence: associations with functional specialization and evolutionary expansion," *PNAS*, vol. 30, no. 20, pp. 1–6, 2016.
[187] J. Wen *et al.*, "Genomic loci influence patterns of structural covariance in the human brain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 52, p. e2300842120, Dec. 2023.
[188] L. Zhang and S. Zhang, "A General Joint Matrix Factorization Framework for Data Integration and Its Systematic Algorithmic Exploration," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 9, pp. 1971–1983, 2020.
[189] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist. Q.*, vol. 2, no. 1–2, pp. 83–97, 1955.
[190] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.
[191] N. A. Susianti *et al.*, "The impact of medial temporal and parietal atrophy on cognitive function in dementia," *Sci. Rep.*, vol. 14, no. 1, p. 5281, 2024.
[192] S. Kaushik, K. Vani, S. Chumber, K. S. Anand, and R. K. Dhamija, "Evaluation of MR Visual Rating Scales in Major Forms of Dementia.," *J. Neurosci. Rural Pract.*, vol. 12, no. 1, pp. 16–23, Jan. 2021.
[193] D. Ferreira *et al.*, "Practical cut-offs for visual rating scales of medial temporal, frontal and posterior atrophy in Alzheimer's disease and mild cognitive impairment.," *J. Intern. Med.*, vol. 278, no. 3, pp. 277–290, Sep. 2015.
[194] M. Wei *et al.*, "A new age-related cutoff of medial temporal atrophy scale on MRI improving the diagnostic accuracy of neurodegeneration due to Alzheimer's disease in a Chinese population.," *BMC Geriatr.*, vol. 19, no. 1, p. 59, Feb. 2019.
[195] G. S. Choi *et al.*, "Age-Specific Cutoff Scores on a T1-Weighted Axial Medial Temporal-Lobe Atrophy Visual Rating Scale in Alzheimer's Disease Using Clinical Research Center for Dementia of South Korea Data.," *J. Clin. Neurol.*, vol. 14, no. 3, pp. 275–282, Jul. 2018.
[196] K. Brueggen *et al.*, "Basal Forebrain and Hippocampus as Predictors of Conversion to Alzheimer's Disease in Patients with Mild Cognitive Impairment - A Multicenter DTI and Volumetry Study.," *J. Alzheimers. Dis.*, vol. 48, no. 1, pp. 197–204, 2015.
[197] D. P. Devanand *et al.*, "Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease.," *Neurology*, vol. 68, no. 11, pp. 828–836, Mar. 2007.
[198] C. Pennanen *et al.*, "Hippocampus and entorhinal cortex in mild cognitive impairment and early AD," *Neurobiol. Aging*, vol. 25, no. 3, pp. 303–310, 2004.
[199] W. H. L. Pinaya *et al.*, "Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.
[200] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial

Autoencoders," *ArXiv*, vol. abs/1511.0, 2015.
[201] K. H. S. Vollmar, J. Cizek, M. Sué, J. Klein, A. H. Jacobs, "VINCI - Volume Imaging in Neurological Research, Co-Registration and ROIs included," in *Forschung und wissenschaftliches Rechnen 2003 (Kremer K, Macho V, eds)*, Göttingen: GWDG, 2004, pp. 115–131.
[202] D. Tosun *et al.*, "Identifying individuals with non-Alzheimer's disease co-pathologies: A precision medicine approach to clinical trials in sporadic Alzheimer's disease.," *Alzheimers. Dement.*, vol. 20, no. 1, pp. 421–436, Jan. 2024.
[203] B. Hou *et al.*, "Interpretable deep clustering survival machines for Alzheimer's disease subtype discovery," *Med. Image Anal.*, vol. 97, p. 103231, 2024.
[204] J. Wen *et al.*, "Dimensional Neuroimaging Endophenotypes: Neurobiological Representations of Disease Heterogeneity Through Machine Learning," *Biol. Psychiatry*, vol. 96, no. 7, pp. 564–584, 2024.
[205] X. Feng, F. A. Provenzano, and S. A. Small, "A deep learning MRI approach outperforms other biomarkers of prodromal Alzheimer's disease.," *Alzheimers. Res. Ther.*, vol. 14, no. 1, p. 45, Mar. 2022.
[206] C. Pettigrew *et al.*, "Progressive medial temporal lobe atrophy during preclinical Alzheimer's disease.," *NeuroImage. Clin.*, vol. 16, pp. 439–446, 2017.
[207] L. Pini *et al.*, "Brain atrophy in Alzheimer's Disease and aging.," *Ageing Res. Rev.*, vol. 30, pp. 25–48, Sep. 2016.
[208] J. N. Fink, M. H. Selim, S. Kumar, B. Voetsch, W. C. Fong, and L. R. Caplan, "Insular cortex infarction in acute middle cerebral artery territory stroke: predictor of stroke severity and vascular lesion.," *Arch. Neurol.*, vol. 62, no. 7, pp. 1081–1085, Jul. 2005.
[209] United Nations, "Leaving no one behind in an ageing world. Word social Report 2023," *Report*, pp. 1–161, 2023.
[210] "2024 Alzheimer's disease facts and figures.," *Alzheimers. Dement.*, vol. 20, no. 5, pp. 3708–3821, May 2024.
[211] A. Kaplan *et al.*, "The effect of a high-polyphenol Mediterranean diet (Green-MED) combined with physical activity on age-related brain atrophy: the Dietary Intervention Randomized Controlled Trial Polyphenols Unprocessed Study (DIRECT PLUS)," *Am. J. Clin. Nutr.*, vol. 115, no. 5, pp. 1270–1281, 2022.
[212] M. Girotti, S. E. Bulin, and F. R. Carreno, "Effects of chronic stress on cognitive function – From neurobiology to intervention," *Neurobiol. Stress*, vol. 33, p. 100670, 2024.
[213] N. N. Kumar, Y. L. Chan, H. Chen, and B. G. Oliver, "Editorial: Effects of environmental toxins on brain health and development," *Front. Mol. Neurosci.*, vol. 16, 2023.
[214] C.-C. Liu, C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy.," *Nature reviews. Neurology*, vol. 9, no. 2. England, pp. 106–118, Feb-2013.
[215] R. Song *et al.*, "Associations Between Cardiovascular Risk, Structural Brain Changes, and Cognitive Decline.," *J. Am. Coll. Cardiol.*, vol. 75, no. 20, pp. 2525–2534, May 2020.

[216] A. L. Young *et al.*, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference," *Nat. Commun.*, vol. 9, no. 1, pp. 1–16, 2018.
[217] M. Hawrylycz, L. Ng, D. Feng, S. Sunkin, A. Szafer, and C. Dang, "The Allen Brain Atlas," in *Springer Handbook of Bio-/Neuroinformatics*, N. Kasabov, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 1111–1126.

Glossary

Adversarial autoencoder, AA: Αντιπαραθετικός αυτοκωδικοποιητής

Alzheimer's disease, AD: Νόσος Αλτσχάιμερ

Cardiovascular disease, CVD: Καρδιαγγειακή νόσος

Clustering: Συσταδοποίηση

Cognitive decline: Γνωστική έκπτωση

Cognitively unimpaired, CU: Γνωσιακά Υγιής

Comorbidity: Συννοσηρότητα

Dementia: Άνοια

Generative adversarial network, GAN: Παραγωγικό αντιπαραθετικό δίκτυο

Genome wide association studies, GWAS: Μελέτες συσχέτισης εύρους γονιδιώματος

Heterogeneity: Ετερογένεια

Intracranial volume, ICV: Ενδοκρανιακός όγκος

Linear mixed-effects model, LME: Γραμμικό μοντέλο μικτών επιδράσεων

Machine learning: Μηχανική μάθηση

Magnetic resonance imaging, MRI: Απεικόνιση μαγνητικού συντονισμού/μαγνητική τομογραφία

Mild cognitive impairment, MCI: Ήπια γνωστική διαταραχή

Neurodegenerative disease: Νευροεκφυλιστική νόσος

Neuroimaging: Νευροαπεικόνιση

Non-negative matrix factorization, NMF: Μη-αρνητική παραγοντοποίηση πίνακα

Normative modelling: Κανονιστική μοντελοποίηση

Personalized medicine: Εξατομικευμένη ιατρική

Principal component analysis, PCA: Ανάλυση κύριων συνιστωσών

Reproducibility index: Δείκτης αναπαραγωγιμότητας

Semi-supervised method: Ημι-επιβλεπόμενη μέθοδος

Single nucleotide polymorphism, SNP: Μονονουκλεοτιδικός πολυμορφισμός

Sparsity: Αραιότητα

Survival analysis: Ανάλυση επιβίωσης

White matter lesions/hyperintensities, WMH: Βλάβες λευκής ουσίας