Aus der Klinik für Psychosomatische Medizin und Psychotherapie

Direktor: Prof. Dr. med. Carsten Spitzer

Sektion für Gerotopsychosomatik und dementielle Erkrankungen

Leiter: Prof. Dr. med. Stefan Teipel

In Kooperation mit dem Deutschen Zentrum für Neurodegenerative Erkrankungen e.V.

Verfahren der künstlichen Intelligenz zur automatisierten Erkennung von Demenzerkrankungen

Kumulative Habilitationsschrift

zur

Erlangung des akademischen Grades

Doctor rerum humanarum habilitatus (Dr. rer. hum. habil.)

der Universitätsmedizin Rostock

vorgelegt von: Dr. rer. hum. Martin Dyrba

geboren am: 5. August 1985 in Zwenkau

wohnhaft in: Schwaan

Rostock, den 17. Juni 2023

https://doi.org/10.18453/rosdok_id00004737

Gutachter:

Prof. Dr. med. Stefan Teipel,
Klinik und Poliklinik für Psychosomatik und Psychotherapeutische Medizin,
Sektion für Gerontopsychosomatik und demenzielle Erkrankungen, Universitätsmedizin Rostock
und
Deutsches Zentrum für Neurodegenerative Erkrankungen e.V., Standort Rostock/Greifswald

Prof. Dr. rer. nat. Christian Wachinger,
Institut für diagnostische und interventionelle Radiologie, Technische Universität München

Prof. Dr. med. Joachim L. Schultze,
Life & Medical Sciences Institute (LIMES), Universität Bonn
und
Deutsches Zentrum für Neurodegenerative Erkrankungen e.V., Standort Bonn

Jahr der Einreichung: 2023

Jahr der Verteidigung und Probevorlesung: 2024

Inhaltsverzeichnis

Abbildungsverzeichnis

Danksagung

Eine wissenschaftliche Arbeit ist nie das Werk einer einzelnen Person. Deshalb möchte ich mich an dieser Stelle bei vielen Menschen für ihre Unterstützung bedanken.

Zuerst gilt mein Dank meinen Mentoren,
Herrn Prof. Dr. med. Stefan Teipel,
Herrn Prof. Dr.-Ing. Thomas Kirste und
Prof. Dr. rer. med. Jochen René Thyrian,
für ihre sachkundige und wertvolle Unterstützung meiner Arbeit, ihre kritischen Hinweise und die konstruktiven Diskussionen.

Weiterhin gilt mein Dank meinen Kollegen und Freunden
von der Universitätsmedizin Rostock,
aus dem Deutschen Zentrum für Neurodegenerative Erkrankungen (DZNE) und
dem Institut für Visual & Analytic Computing,
ohne deren Rat und Tat, Motivation und Hilfestellung diese Arbeit nicht möglich gewesen wäre.

Zuletzt gilt mein größter Dank meiner Familie,
meiner Frau Anica und
meinen Kindern Elisabeth, Josef, Johanna und Sarah,
für ihre Unterstützung und ihren Beistand, ihre Geduld, Liebe und Fürsorge in allen Höhen und Tiefen.

1 Einleitung

1.1 Klinische und pathologische Charakterisierung von Demenzerkrankungen

Der Begriff Demenz beschreibt syndromal die Verschlechterung kognitiver Funktionen, vor allem in den Bereichen Gedächtnis, Aufmerksamkeit, Sprache, Orientierung oder Sozialverhalten [1]. Hierbei werden klinisch drei Stadien der kognitiven Beeinträchtigung unterschieden [2]. Bei der subjektiven kognitiven Beeinträchtigung (engl. *subjective cognitive decline, SCD*) berichten Betroffene von der subjektiv wahrgenommenen Verschlechterung der kognitiven Leistung im Vergleich zur früheren Leistungsfähigkeit, jedoch ist die objektiv messbare Leistung noch im Normalbereich bezogen auf Referenzwerte kognitiv gesunder Personen. Bei der leichten kognitiven Störung (LKS, engl. *mild cognitive impairment, MCI*) ist die kognitive Leistungsfähigkeit der Betroffenen eingeschränkt und schlechter als alters-, geschlechts- und bildungsbezogene Referenzwerte, jedoch ist die Alltagskompetenz der Betroffenen noch gegeben. Bei stärker ausgeprägten kognitiven Beeinträchtigungen, die die Alltagskompetenz der Betroffenen maßgeblich einschränken, spricht man von einer Demenz.

Für die Demenz sind verschiedene Ursachen bekannt. Grundsätzlich unterscheidet man primäre Formen aufgrund bestimmter Erkrankungen, die häufig irreversibel verlaufen und das Gehirn nachhaltig schädigen, und sekundäre Formen als Folgeerscheinungen anderer Erkrankungen (z.B. Stoffwechselerkrankungen, Vitaminmangelzustände oder Depression), deren Behandlung die Demenzsymptome teilweise abmildern oder sogar rückbilden kann. Bei den primären Formen der Demenz ist die Alzheimer-Krankheit als häufigste Ursache zu nennen (60-65% der Fälle) [3, 4]. Weiter treten die vaskuläre Demenz (20-25%), die frontotemporale Demenz (10-15%) oder Mischformen auf [3-6]. Zur Abgrenzung und rechtzeitigen Behandlung dieser Demenzerkrankungen ist eine frühzeitige Diagnose besonders wichtig. Jedoch haben Eichler et al. 2015 [7] anhand einer primärärztlichen Kohorte der DelpHi-MV Studie gezeigt, dass gegenwärtig im deutschen Gesundheitssystem weniger als die Hälfte der Demenzbetroffenen in der primärärztlichen Versorgung auch tatsächlich formal eine Diagnose erhalten. Dies deckt sich mit Zahlen internationaler Studien.

Im Wesentlichen sind die drei genannten Primärformen der Demenz altersassoziierte Erkrankungen. Davon abzugrenzen sind die seltenen genetischen Varianten der Alzheimer-Erkrankung, die autosomal-dominant vererbt werden. Betroffene Personen mit Mutationen im Präsenilin 1- (PSEN1), Präsenilin 2- (PSEN2) und Amyloid-Vorläufer-Protein-Gen (APP) erkranken meist schon im Lebensalter zwischen 30 und 60 Jahren. Dagegen tritt die häufige sporadische Form der Alzheimer-Demenz bei Betroffenen ab einem Alter von etwa 65 Jahren auf, wobei das Risiko mit steigendem Alter annähernd linear ansteigt [8]. Die Alzheimer-Krankheit ist durch sogenannte „senile Plaques“ charakterisiert, d.h.

extrazelluläre Ablagerungen von aggregierten Amyloid-β-Peptiden [1, 9], sowie durch intraneuronale fibrilläre Aggregate des τ-Proteins, die als „neurofibrilläre Bündel“ bezeichnet werden [1, 9]. Die Ätiologie der Erkrankung ist derzeit noch nicht abschließend geklärt [1]. Im Verlauf der Erkrankung kommt es zu einer Schädigung der Mitochondrien [10], zur Störung des Zellmetabolismus und final zum Zelltod der Neurone [9]. Dies manifestiert sich früh vor allem als Atrophie der Hippocampusformation und des posterioren Gyrus cinguli, später auch im gesamten Temporallappen sowie in weitreichenden kortikalen Arealen des Frontal- und Parietallappens. Bei der vaskulären Demenz kommt es infolge von Durchblutungsstörungen im Gehirn zum Absterben von Nervengewebe. Als Hauptursachen gelten altersassoziierte Faktoren, die generell das Risiko von Gefäßerkrankungen erhöhen, wie Bluthochdruck, Herzerkrankungen oder Diabetes mellitus. Bei der frontotemporalen Demenz treten die Symptome häufig bereits im Alter zwischen 45 und 60 Jahren auf [6]. Der Oberbegriff frontotemporale Lobärdegeneration (FTLD) beschreibt hierbei drei verschiedene klinische Varianten [6], die anhand der frühen bzw. vorherrschenden Symptome unterschieden werden: die behaviorale frontotemporale Demenz (bvFTD), die semantische Demenz (SD) und die progressive nicht-flüssige Aphasie (PNFA). Die Ätiologie dieser Varianten ist derzeit noch nicht abschließend geklärt [6]. Teilweise überschneiden sich die histologischen Befunde, so dass für alle drei Varianten Proteinopathien des τ-Proteins, des „TAR-DNA-bindenden Protein 43“ (TDP-43) oder des „fused in sarcoma“ (FUS)-Proteins möglich sind [6]. In diesem Zusammenhang ist auch die Amyotrophe Lateralsklerose (ALS) zu nennen, deren histologischen Befunde und Symptome sich teilweise mit der bvFTD überschneiden [11]. Die ALS ist eine fortschreitende Erkrankung der Motoneurone, die zu Muskellähmung führt. Betroffene sterben meist innerhalb weniger Jahre. Die ALS kann dabei isoliert ohne kognitive Beeinträchtigungen auftreten, mit exekutiven und/oder kognitiven Beeinträchtigungen sowie als ALS-FTD. Man spricht hier deshalb von einem ALS-FTD-Spektrum [11].

Der demographische Wandel in Deutschland sowie in allen westlichen Industrienationen und Schwellenländern stellt uns vor große Herausforderungen. Hochrechnungen erwarten eine Prävalenz der Demenz von 2,7 Millionen Betroffenen in Deutschland für das Jahr 2050 (Abbildung 1).

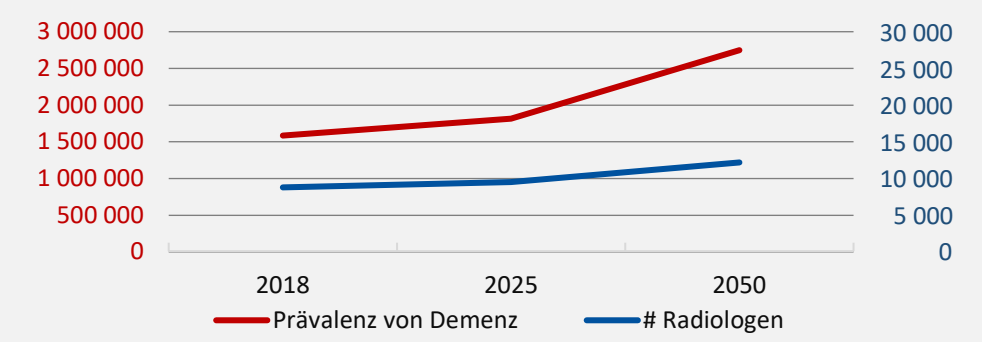


Abbildung 1. Prävalenz der Demenz bis 2050 im Vergleich zur Verfügbarkeit von diagnostischen Ressourcen, hier am Bsp. der Anzahl von Radiologen.
Quellen: Dementia in Europe Yearbook 2019, Statistiken d. Bundesärztekammer.

Dagegen ist die Zahl der Versorgungszentren wie Gedächtnisambulanzen und Fachärzten begrenzt (siehe Abbildung 1 als Bsp. die Anzahl von Radiologen), so dass wir mit einer stärker werdenden Beanspruchung oder gar Überlastung von diagnostischen Ressourcen rechnen müssen. In diesem Dilemma könnten KI-basierte Experten- und Assistenzsysteme den Diagnoseprozess unterstützen und beschleunigen, in dem zeitaufwändige Routineaufgaben automatisiert werden.

1.2 Bildgebung bei Demenzerkrankungen

Die bildgebenden Verfahren, vor allem die Magnetresonanztomographie (MRT) und Positronen-Emissions-Tomographie (PET), haben für die Differentialdiagnose von Demenzerkrankungen eine besondere Bedeutung. Die T_1-gewichtete MRT ermöglicht es, die Hirnstruktur bzw. eine Reduktion des Hirnvolumens (Atrophie) in vivo zu beurteilen [12]. Die Diffusions-Tensor-Bildgebung (DTI) misst die Bewegungsfreiheit (=Diffusion) von Wassermolekülen im Gewebe, die von der Mikrostruktur des Gewebes abhängig ist, z.B. von Zellmembranen oder Mikrotubuli, so dass sich Rückschlüsse auf die Integrität der Fasertrakte der weißen Substanz ziehen lassen [13]. Die Fluorodesoxyglukose-Positronen-Emissions-Tomographie (FDG-PET) bemisst den Glukosemetabolismus. Patienten mit einer Alzheimer-Erkrankung zeigen einen regional spezifischen Hypometabolismus im posterioren Gyrus cinguli [14, 15]. Weiterhin wurden für Demenzerkrankungen spezielle Tracer-Substanzen entwickelt, die mittels PET das Vorhandensein und die lokale Ausbreitung von Amyloid-β- oder τ-Proteinen abbilden können [16]. Eine weitere Bildgebungstechnik ist die funktionelle Magnetresonanz-tomographie (fMRT), bei der die unterschiedlichen magnetischen Eigenschaften von oxygeniertem und desoxygeniertem Blut genutzt werden, um über den Mechanismus der neurovaskulären Kopplung Rückschlüsse auf die neuronale Aktivität in den verschiedenen Hirnregionen zu ziehen [17, 18]. Bei verschiedenen kognitiven Aufgaben zeigen sich verteilt spezifische Hirnregionen aktiv, die ein synchrones Zeitsignal aufweisen. Diese werden deshalb als *intrinsische funktionelle Konnektivitäts-netzwerke* bezeichnet [19, 20]. Die bei der funktionellen Magnetresonanztomographie im Ruhezustand des Probanden (Ruhe-fMRT) synchron aktiven Regionen, darunter Bereiche im posterioren Gyrus cinguli und Precuneus, im medialen Frontallappen und beidseitig in den inferioren Parietallappen, werden *Default Mode Network* (DMN) genannt [21, 22]. Im Verlauf der Alzheimer-Krankheit verändert sich die Synchronizität bzw. funktionelle Konnektivität dieses Netzwerks [23-30].

1.3 Maschinelle Lernverfahren zur Analyse von Bildgebungsdaten

Verfahren der künstlichen Intelligenz, fachsprachlich meist als maschinelle Lernverfahren bezeichnet, sind in der Lage, anhand von Trainingsdaten automatisiert Muster zu identifizieren und in Form einer mathematischen Abbildungsfunktion $y = f(x)$ verschiedenen Ausgabewerten zuzuordnen. Besonders die multivariaten Verfahren, bei denen zahlreiche Eingabemerkmale kombiniert ausgewertet werden, zeigen sich sensibel für verteilte Gewebe- bzw. Intensitätsveränderungen, wie

sie im Verlauf von Demenzerkrankungen auftreten. Beim Gehirn wirken viele Prozesse (z.B. Entwicklung, Alterung, Krankheit) auf die gesamte Struktur, so dass sich bei der Auswertung von mehreren Hirnregionen häufig das **Problem der Kollinearität** ergibt, d.h. Messwerte ähneln sich bzw. teilen ihre Varianz. Dies ist in der empirischen Forschung und insbesondere für datengetriebene maschinelle/statistische Lernverfahren eine große Herausforderung, da die geteilte Varianz zu störenden Korrelationen führt und insbesondere klassische statistische Modelle wie die linearen/logistischen Regressionsanalysen instabil werden lässt [31], d.h. kleine Änderungen an der Zusammensetzung der Stichprobe haben teilweise große Änderungen der geschätzten Modellparameter zur Folge. Deshalb wurden in der Literatur Regularisierungsverfahren etabliert, die die Modellschätzung durch die Einführung von „Straftermen" stabilisieren soll. Die sogenannte Elastic-Net Regularisierung kombiniert dazu die Verfahren LASSO (*least absolute shrinkage and selection operator*) und Ridge Regression, die Strafterme basierend auf der sogenannten L1- bzw. L2-Norm des Parametervektors β verwenden. Diese Normen definieren allgemein eine Metrik zur Bestimmung der Größe oder Länge eines Vektors. Hierbei wird das lineare Optimierungsproblem zur Schätzung der Regressionskoeffizienten β_i dahingehend modifiziert, dass deren Betragssummen (L1-Norm) und Quadratsummen (L2-Norm) Berücksichtigung finden, wie im Folgenden detaillierter ausgeführt wird.

Für die lineare Regression $y = \sum_{i=0}^{p} \beta_i x_i + \varepsilon$ mit p unabhängigen Variablen (Prädiktoren) x_i werden i.d.R. die Parameter β_i anhand des Kriteriums des *kleinsten quadratischen Fehlers* bestimmt (d.h. $\operatorname{argmin} \sum_{j=1}^{n} \varepsilon_j^2$). Für einen Trainingsdatensatz x_{ij} mit $j = 1, \dots, n$ Beobachtungen (Fällen) und $i = 1, \dots, p$ unabhängigen Variablen ergibt sich daraus:

$$\hat{\beta}_i \coloneqq \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{n} \left(y_j - \sum_{i=0}^{p} \beta_i x_{ij} \right)^2.$$

Bei der Elastic-Net Regularisierung wird diese Optimierungsfunktion erweitert zu:

$$\hat{\beta}_i \coloneqq \underset{\beta}{\operatorname{argmin}} \left(\sum_{j=1}^{n} \left(y_j - \sum_{i=0}^{p} \beta_i x_{ij} \right)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sqrt{\sum_{i=1}^{p} \beta_i^2} \right), \text{ mit } \lambda_1 + \lambda_2 = 1.$$

Der L1-Norm LASSO Term $\sum_{i=1}^{p} |\beta_i|$ bestraft dabei die Anzahl an $\beta_i \neq 0$ und stellt somit einen Mechanismus zur Variablenselektion dar. Der L2-Norm Ridge Regression Term $\sqrt{\sum_{i=1}^{p} \beta_i^2}$ reduziert effektiv den Betrag der Regressionskoeffizienten β_i, welche sonst insbesondere für hoch korrelierte Prädiktoren ohne diesen Strafterm beliebig groß und divergierend werden könnten [31].

1.4 Tiefe neuronale Faltungsnetze zur Erkennung komplexer Muster in hochdimensionalen Bilddaten

Über die oben genannten Ansätze hinaus, ermöglichen hochdimensionale nichtlineare Verfahren, wie die tiefen neuronalen Netze, komplexe Zusammenhänge in den Eingabedaten zu erkennen und

miteinander zu kombinieren. Dazu soll im Folgenden ein Überblick über verschiedene neuronale Lernverfahren gegeben und diese konzeptionell erklärt werden. Abbildung 2 illustriert die „Evolution“ von künstlichen neuronalen Modellen.

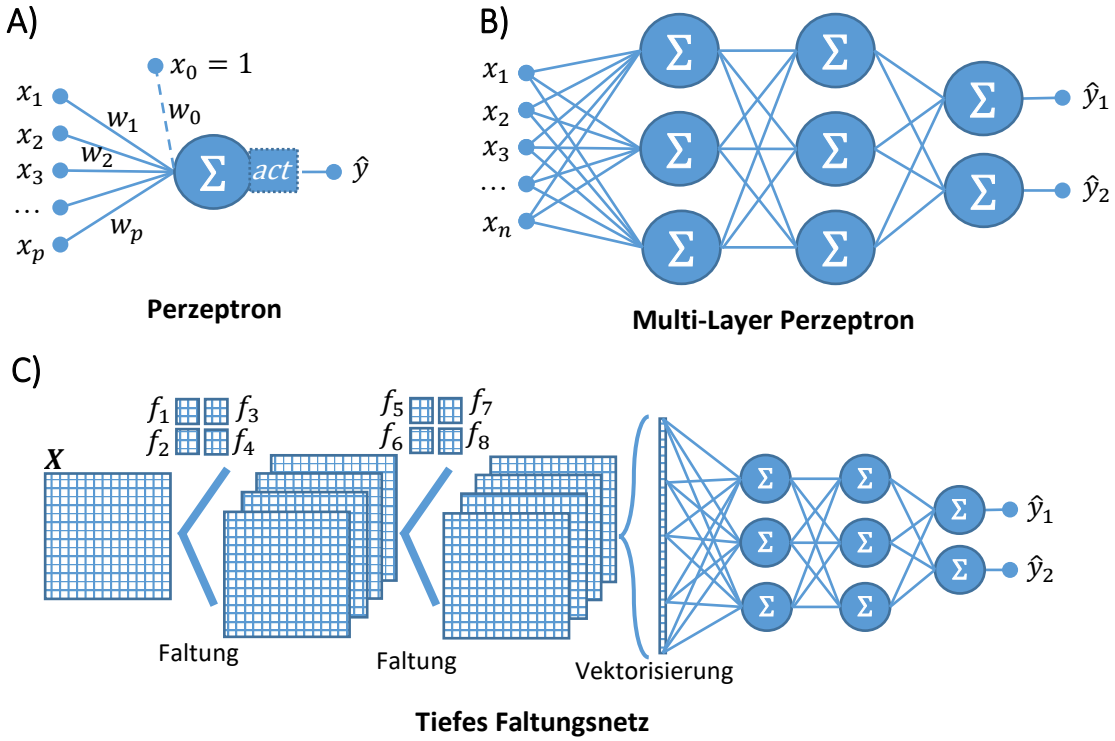


Abbildung 2. Evolution der künstlichen neuronalen Netze für die Bilderkennung.
Eigene Abbildung des Autors.

Abbildung 2 **A)** Künstliches Neuron, häufig auch als **Perzeptron** bezeichnet. Für jeden der p numerischen Eingabewerte x_i wird eine Wichtung w_i bestimmt. Hierbei ist der sogenannte Bias-Term $x_0 = 1$ als Konstante festgelegt. Somit entspricht der vordere Teil des Perzeptrons der linearen Regression und berechnet sich aus $f(x) = \sum_0^p w_i x_i$. Häufig wird anschließend eine sogenannte Aktivierungsfunktion *act* nachgeschaltet, so dass sich die Vorhersage aus $\hat{y} = act(f(x))$ ergibt. Hierzu wird im einfachsten Fall die Identitätsfunktion $act(\dot{x}) = \dot{x}$ (entspricht der linearen Regression) oder die Sigmoid-Funktion $act(\dot{x}) = 1/(1 + e^{-\dot{x}})$ (entspricht der logistischen Regression) verwendet[1]. Aufgrund der geringeren numerischen Komplexität bzw. des schnelleren Trainings wird bei tiefen Faltungsnetzten häufig die Gleichrichtfunktion (*Rectified Linear Unit, ReLU*) als Aktivierungsfunktion genutzt, mit $act(\dot{x}) = \max(0, \dot{x})$, d.h. negative Eingabewerte werden auf 0 gesetzt.

Abbildung 2 **B)** Schichtung und Verkettung von mehreren Perzeptronen als **Multi-Layer Perzeptron,** zur besseren Übersicht hier vereinfacht dargestellt ohne Bias und Aktivierungsfunktion. Häufig wird die erste Schicht als Eingabeschicht bezeichnet, gefolgt von einem oder mehreren „versteckten“

[1] Formal soll hier $\dot{x}$ ein Zwischenergebnis repräsentieren, das sich z.B. aus $\dot{x} = f(x)$ ergibt. Das Symbol x soll zur besseren Übersicht die Eingabewerte der Eingabeschicht der neuronalen Modelle repräsentieren.

Schichten und abschließend der Ausgabeschicht. In solchen vollverbundenen Netzen wird für jede Verbindung zwischen den Neuronen ein eigenes Gewicht w_{ij} bestimmt, weshalb die mathematische Komplexität sehr groß und damit das Training sehr aufwändig ist.

Abbildung 2 **C)** Verkettung von Faltungsschichten (Convolutional Layers) und vollverbundenen Schichten zu einem **tiefen Faltungsnetz**. Die Faltungsoperation mit einem Faltungsfilter f_i ermöglicht eine besondere Merkmalsextraktion, da hierüber spezifische Formen oder Texturen aus den Eingabedaten extrahiert und in die mittleren bzw. hinteren Schichten propagiert werden können. Dabei dienen alle Ausgaben einer Faltungsschicht als Eingabe für die darauffolgende Faltungsschicht und können somit dort neu kombiniert werden. Die finale Ausgabe der Faltungsschichten wird in eine eindimensionale Vektorform überführt und anschließend über ein oder mehrere vollvernetzte Schichten ausgewertet. Der vordere Teil mit den Faltungsschichten ermöglicht die Merkmalsextraktion, die hinteren vollvernetzen Schichten die Klassifikation bzw. Regression.

In aktuellen tiefen Faltungsnetzen werden zwischen den Faltungsschichten zusätzlich sogenannte Pooling-Layer eingefügt, die das Maximum bzw. den Mittelwert von zwei benachbarten Pixeln/Voxeln berechnen und damit die Auflösung je Dimension halbieren. Ebenso haben moderne Netze sogenannte „Skip-“ oder „Residual-Connections“, bei denen Faltungsschichten ihre Eingabe von mehreren vorgelagerten Schichten erhalten. Durch die Hintereinanderschaltung verschiedener Schichten als „Blöcke“ und Verknüpfung dieser lassen sich komplexe hierarchische Netztopologien aufbauen, die in der Lage sind, Merkmale der Eingabedaten auf lokaler (mikroskopisch), mittlerer (mesoskopisch) und globaler Ebene (makroskopisch) zu integrieren und zu abstrakten Konzepten zusammenzufassen. Beispielsweise können bei der Gesichtserkennung mittlere Schichten abstrakte Konzepte wie Augenbrauen, Nase oder Lippen erkennen.

1.5 Erklärbarkeit und Nachvollziehbarkeit von maschinellen Lernverfahren

Eine besondere klinische wie auch methodische Herausforderung ist die Erklärbarkeit und Nachvollziehbarkeit von maschinellen Lernverfahren und deren Ausgaben, die in der Folge auch die Anwendbarkeit und Nützlichkeit algorithmischer Entscheidungsunterstützung im klinischen Kontext stark limitiert. In der Stellungnahme der Zentralen Ethikkommission der Bundesärztekammer zur Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz fordert die Kommission, dass Ärzte und Ärztinnen in die Lage versetzt werden müssen, automatische Entscheidungsempfehlungen auf Plausibilität und Richtigkeit zu prüfen [32]. In der Literatur werden maschinelle Lernverfahren bezüglich ihrer Transparenz und Komplexität in „interpretierbare“ und „opaque“ („Black-Box“) Modelle eingeteilt [33, 34]. Die interpretierbaren Modelle, beispielsweise lineare oder logistische Regressionsmodelle, regelbasierte Verfahren oder Entscheidungsbäume, bilden üblicherweise eine mathematisch weniger komplexe Funktion ab und sind bezüglich des

Zusammenhangs von Eingabe und Ausgabe für den Menschen eher leicht nachvollziehbar. Opaque Modelle dagegen weisen eine hohe mathematische Komplexität und einen nichtlinearen Zusammenhang von Eingabe und Ausgabe auf, so dass die Funktionsweise für Menschen kaum nachvollzogen werden kann. Insbesondere sind bei datengetriebenen und trainierbaren Verfahren, z.B. bei tiefen neuronalen Netzen, die mathematischen Operationen innerhalb des Entscheidungsprozesses solcher Modelle selbst für deren Entwickler größtenteils unklar, da sich die Modellparameter bei jedem Trainingsschritt ändern. In den vergangenen Jahren wurden für opaque Modelle daher post-hoc Verfahren entwickelt, die aus den komplexen Entscheidungsprozessen zur leichteren Interpretierbarkeit vereinfachte Kennwerte oder approximative lineare Modelle ableiten [33, 34]. Hierbei werden Verfahren zur globalen Interpretierbarkeit des Modells, die versuchen das Gesamtverhalten des Modells zu erklären, und lokale Verfahren, die einzelne Aspekte für ein bestimmtes Eingabedatum aufzeigen, differenziert. Globale Verfahren sind beispielsweise Shapley-Werte oder Sensitivitätsanalysen. Die Shapley-Werte haben ihren Ursprung in der Spieltheorie und quantifizieren den mittleren Beitrag jedes Eingabemerkmals auf die Modellausgabe. Lokale Verfahren sind beispielsweise die Schichtweise Relevanzrückverfolgung (*Layer-wise Relevance Propagation, LRP*), in dem die „neuronale Aktivierung" innerhalb von tiefen neuronalen Netzen zurückverfolgt und einzelnen Eingabebereichen zugeordnet werden kann (Abbildung 3) [35].

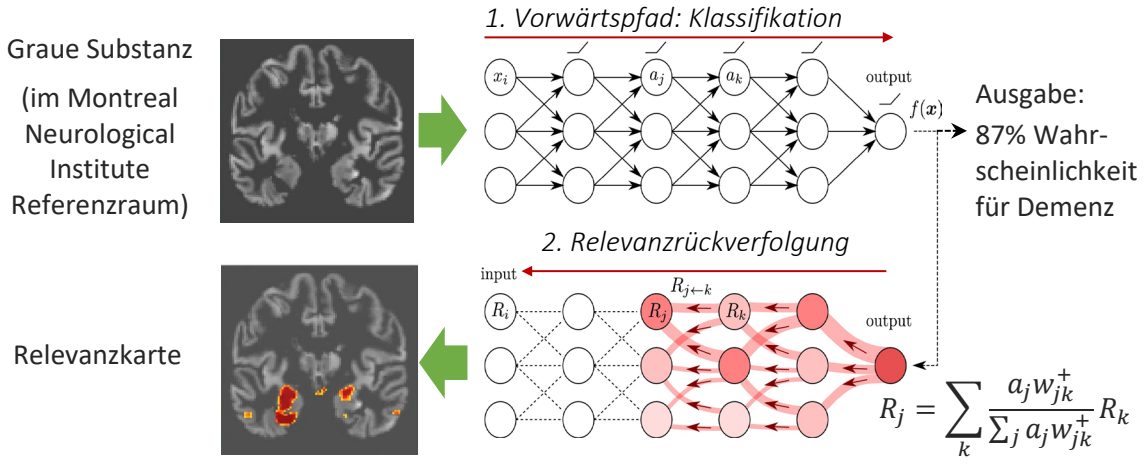


Abbildung 3. Schichtweise Relevanzrückverfolgung in einem tiefen neuronalen Netz.
Adaptiert nach [36].
Symbole: x_i Eingabewert; a_j, a_k Aktivierungen eines Neurons in den Schichten j bzw. k; R_i, R_j, R_k Relevanz in den jeweiligen Schichten; w_{jk}^+ Gewicht der Verbindung zwischen zwei Neuronen.
Die ermittelte Relevanz ergibt sich aus der schrittweisen Relevanzrückverfolgung, hier dargestellt mit der $\alpha\beta$-Propagationsregel und $\alpha = 1, \beta = 0$, bei der nur Pfade mit positiven Gewichten bzw. Aktivierungen berücksichtigt ($\alpha = 1$) und negative Aktivierung unterdrückt werden ($\beta = 0$). Die Relevanz R_j für ein Neuron in Schicht j berechnet sich aus der Summe der Relevanz R_k aller verbundenen Neurone der nachfolgenden Schicht k, gewichtet mit der jeweiligen Aktivierungen a_j und den positiven Gewichten w_{jk}^+ der entsprechenden Verbindungen. Diese Propagationsregel lässt sich vereinfacht so interpretieren, dass die Relevanz R_k auf die Neurone der davorliegenden Schicht R_j proportional zu ihrer exzitatorischen Wirkung auf a_k verteilt wird [36].

1.6 Untersuchungsziele und Hypothesen

Für die vorliegende Habilitationsschrift zu Verfahren der künstlichen Intelligenz für die automatisierte Erkennung von Demenzerkrankungen wurde die übergreifende Forschungsfrage untersucht, mit welchen Verfahren sich kollineare Bildgebungsdaten des Gehirns modellieren und auswerten lassen, wie stabil und robust diese sind, und welche der getesteten Verfahren sich folglich für den Einsatz mit empirischen Daten klinischer Studien am besten eignen.

In diesem Kontext wurden drei Schwerpunkte bearbeitet und in verschiedenen Studien untersucht:

I. **Die Erklärbarkeit von Entscheidungen von künstlichen neuronalen Netzen**
   Bis vor wenigen Jahren gab es kaum Erkenntnis darüber, welche Merkmale oder Bildbereiche tiefe neuronale Netze zur Differenzierung von diagnostischen Gruppen heranziehen. Im Rahmen dieser Arbeit sollten post-hoc Verfahren zur Erklärung von Entscheidungsprozessen für medizinische 2D/3D Bildgebungsdaten identifiziert und bezüglich ihrer Eignung und Nützlichkeit für klinische Nutzerinnen und Nutzer evaluiert werden. Als Hypothese wurde formuliert, dass gut geeignete Attributionsverfahren den Schweregrad der Erkrankung bzw. den Umfang der zugrundeliegenden Atrophie in entsprechender Weise darstellen.

II. **Die Evaluation und Charakterisierung neuartiger Bildgebungsmarker**
   Im Rahmen zahlreicher Kollaborationen wurden multizentrische Stichproben mit multimodaler MRT-Bildgebung erhoben, z.B. funktionelle MRT und Diffusions-Tensor-Bildgebung. Hierzu sollte die diagnostische Eignung von neuartigen Bildgebungsverfahren, Auswertungsmethoden und digitalen Bildgebungsmarken evaluiert werden. Als Referenz wurde das Hippocampusvolumen für die Bewertung der Gruppentrennung im Kontinuum der Alzheimer-Erkrankung festgelegt. Zu evaluierende Verfahren wurden bezüglich Erkennungsrate und multizentrischer Variabilität bzw. Stabilität und Kohärenz verglichen. Ebenso wurden Verfahren und Bildgebungsmarker, wie der BrainAge Score, die in der Alzheimer-Forschung bereits gut etabliert sind, bei anderen neurodegenerativen Erkrankungen, wie ALS und ALS-FTD, untersucht.

III. **Die Multivariate Modellierung statistischer Zusammenhänge von Schädigungsmustern**
   Anhand von großen Stichproben mit multimodaler Bildgebung für die Charakterisierung spezifischer Prozesse bei der Alzheimerkrankheit wurden neuartige Verfahren zur Modellierung statistischer Zusammenhänge von Schädigungsmustern erprobt und evaluiert. Das oben genannte Problem der Kollinearität, d.h. die Ähnlichkeit der Messwerte und ihre geteilte Varianz, wird hierbei mathematisch ausgenutzt, um deren wahrscheinlichen Ursprung zu rekonstruieren und eindeutig zuzuordnen. Die Gauß'schen Graphenmodelle stellten hierzu einen innovativen Ansatz zur Erhebung von *partieller* Kovarianz dar. Die resultierenden partiellen Kovarianz-

netzwerke der Gauß'schen Graphenmodelle wurden mit klassischen Kovarianznetzwerken verglichen, die auf der Pearson-Korrelation beruhen. Ebenso wurde die Eignung, Robustheit und die Nützlichkeit dieser Verfahren für die Auswertung empirischer klinischer Daten bewertet.

2 Studien und Ergebnisse

2.1 Erklärbarkeit von Entscheidungen von künstlichen neuronalen Netzen

Für die Erklärung von neuronalen Modellen zur Bilderkennung gibt es in der Literatur zahlreiche Verfahren zur Ableitung von Salienz- oder Relevanzwerten. Mit diesen sollen aus den komplexen neuronalen Schichten und Rechenoperationen vereinfachte Visualisierungen ermittelt werden, die jedem Eingabemerkmal (Bildpunkt) ein Gewicht zuordnen, wie stark dessen Einfluss auf die Modellentscheidung ist. Diese Verfahren sind gegenwärtig immer noch Gegenstand der Grundlagenforschung in der Informatik, da gewisse Annäherungen und Vereinfachungen innerhalb dieser Ansätze einen großen Einfluss auf die ermittelten Relevanzwerte hat. Gemeinhin werden zur Bewertung und zum Vergleich dieser Verfahren standardisierte Trainingsdatensätze verwendet, wie der MNIST-Datensatz [37] mit handschriftlichen Abbildungen der Zahlen 0 bis 9 oder ImageNet [38] und COCO [39] mit annotierten Fotos von mehreren hundert Objektklassen (z.B. Hund, Katze, Flugzeug). Zu Beginn dieser Habilitationsarbeit gab es keine gesicherten Erkenntnisse, ob diese Verfahren bei komplexen Modellen für medizinische Bilddaten und die Krankheitserkennung ebenfalls funktionieren, d.h. insbesondere für klinische Nutzer nutzbringende Informationen ermitteln können.

Im **Konferenzbeitrag KB1** wurden zahlreiche Verfahren zur Ableitung von Relevanzkarten systematisch verglichen, um eine qualitative Einschätzung der Nützlichkeit und Ähnlichkeit der Ergebnisse durchzuführen. Hierzu wurde ein neuronales Netz auf 3D MRT-Daten zur Erkennung der Alzheimer-Krankheit trainiert. Insgesamt zeigten sämtliche getesteten Verfahren eine vergleichbare Charakteristik und ähnliche Hirnareale als besonders informativ. Eine hohe Relevanz wurde von den Verfahren *Guided Backpropagation*, *Deep Taylor Decomposition*, *Input*Gradient*, und *Layer-wise Relevance Propagation (LRP)* im mesialen Temporallappen, vor allem im Hippocampus, attribuiert. Entsprechend der a priori Hypothese waren hohe Relevanzwerte bei einem Probanden mit Alzheimer-Demenz beidseitig und bei einem Probanden mit leichter kognitiver Störung primär rechts und etwas schwächer dargestellt. Dagegen ermittelten die Verfahren bei einer kognitiv gesunden Kontrollperson abgesehen von kleinerem Rauschen keine nennenswerten Hirnareale mit hohen Relevanzwerten. Im Kontrast dazu scheiterte das beliebte Verfahren *Grad-CAM*, welches eine sehr geringe räumliche Auflösung aufwies, so dass einzelne Hirnareale nur unzureichend abgrenzbar waren sowie, entgegen der Hypothese, alle drei diagnostischen Gruppen etwa gleich hohe Relevanzwerte aufwiesen.

Für den **Konferenzbeitrag KB2** wurde auf der Basis dieser Ergebnisse eine Web-Applikation konzipiert und entwickelt, die es ermöglicht, neue MRT-Daten durch das künstliche neuronale Netz auszuwerten, 3D Relevanzkarten zu berechnen und für die Begutachtung interaktiv zu visualisieren. Dazu wurde das

Visualisierungs-Framework *Bokeh*[2] verwendet, das entsprechende Komponenten für die Webseiten-Darstellung (Frontend) und eine Python-Laufzeitumgebung (Backend) bereitstellt. Die Implementierung erfolgte dem Model–View–Controller Programmierparadigma folgend. Die Applikation steht interessierten Nutzern zu Demonstrationszwecken unter https://explaination.net zur Verfügung und der Quellcode wurde auf GitHub veröffentlicht.

In der **Originalarbeit OA1** wurden die oben genannten technischen Verfahren und Konzepte im Hinblick auf die klinische Tauglichkeit systematisch und umfangreich evaluiert. Dazu wurden 1.) die diagnostische Genauigkeit der künstlichen neuronalen Netze im Rahmen einer zehnfachen Kreuzvalidierung sowie anhand von drei unabhängigen Stichproben erhoben, 2.) Relevanzkarten abgeleitet und bezüglich der räumlichen Verteilung von Relevanzwerten beurteilt und 3.) der Umfang der Relevanzwerte für die Hippocampusregion mit dem etablierten radiologischen Bildgebungsmarker Hippocampusvolumen verglichen. Die neuronalen Netze erreichten hierbei für die unabhängigen Stichproben eine balancierte Genauigkeit zwischen 83% und 88% sowie eine AUC zwischen 0.91 und 0.98 für die Differenzierung von Patienten mit Alzheimer-Demenz und kognitiv gesunden Kontrollen. Für die Trennung von Patienten mit leichter kognitiver Störung und gesunden Probanden war die balancierte Genauigkeit substanziell geringer mit 63% bis 75% und einer AUC zwischen 0.68 und 0.84. Die LRP Relevanzkarten bestätigten die oben genannte Hypothese bezüglich des Ausmaßes von hohen Relevanzwerten entsprechend dem Grad der Erkrankung. Bei der Betrachtung von einzelnen Individuen zeigte sich aber eine gewisse Heterogenität bei den Relevanzkarten, die neben dem Hippocampus auch Atrophie in weiteren kortikalen und subkortikalen Hirnregionen markierten. Der Vergleich von Relevanzwerten in der Hippocampusregion mit dem Hippocampusvolumen zeigten einen sehr hohen linearen Zusammenhang mit einer Pearson-Korrelation von $r \approx -0{,}87$ ($p < 0.001$) über alle Stichproben hinweg.

Im **Konferenzbeitrag KB3** wurde in Ergänzung zu den vorangegangenen Arbeiten vier verschiedene Netzarchitekturen für tiefe Faltungsnetze bezüglich der Genauigkeit und Heterogenität der Relevanzkarten verglichen. Dazu wurden für die vier Netzarchitekturen AlexNet, VGG, ResNet und DenseNet Modelle trainiert und verglichen. ResNet erreichte die höchste Genauigkeit mit 81% für die Erkennung von Patienten mit Alzheimer-Demenz, wobei alle Modelle eine hohe Varianz der Erkennungsraten aufwiesen. VGG und DenseNet hatten hierbei die geringste Standardabweichung in der Höhe von 8% für Demenz vs. Kontrollen bzw. 3% für leichte kognitive Störung vs. Kontrollen. Bei DensNet waren die gemittelten Relevanzkarten am stärksten auf den mesialen Temporallappen fokussiert, was am ehesten der a priori Hypothese zu diagnostisch relevanten Hirnregionen entspricht.

[2] http://bokeh.org/

AlexNet und ResNet waren dagegen deutlich heterogener in der räumlichen Verteilung von Relevanzwerten und VGG fokussierte von den anderen Modellen abweichend primär auf inferior-lateral temporale Regionen.

In der **Originalarbeit OA2** wurden selbst-überwachte Lernverfahren experimentell untersucht. Im Gegensatz zu den vorangegangenen Arbeiten, in denen klassische überwachte Lernverfahren verwendet wurden, d.h. Modelltraining mit Bilddaten als Eingabe und der Diagnose als Ausgabe, werden beim selbst-überwachtem Lernen künstlich die Eingabedaten modifiziert und entsprechende synthetische Ausgabeklassen erzeugt. Ziel dieser Ansätze ist es, repräsentative Merkmale aus den Trainingsdaten zu lernen. In dieser Studie wurden dazu einzelne koronare Schichten des 3D Bildes als Eingabe verwendet, zusammen mit den Operationen 1.) Bildspiegelung, 2.) Vergrößerung und Beschneidung sowie 3.) das zufällige Löschen von einem rechteckigen Bildbereich. Als Ausgabe bzw. Zielfunktion (Loss-Funktion) diente die Information, ob je zwei Bilder ursprünglich das gleiche waren oder verschiedene. Als Modellarchitektur wurde ConNeXt verwendet, die aktuell als die leistungsfähigste für allgemeine 2D Bilderkennungsaufgaben gilt. Im Rahmen der Studie erbrachte das anschließende Trainieren einer logistischen Regression Zusatzschicht für die Gruppentrennung eine balancierte Genauigkeit von 59% für die vier-Klassen Unterscheidung von Alzheimer-Demenz vs. leichte kognitive Störung vs. behaviorale Variante der frontotemporalen Demenz vs. Kontrollen, und eine balancierte Genauigkeit von 80% für die zwei-Klassen Unterscheidung von Alzheimer-Demenz vs. Kontrollen. Die detailliertere Analyse der Relevanzkarten anhand des Verfahrens *Integrated Gradients* zeigte jedoch, dass das Modell als Folge der vorher festgelegten Zielfunktion für das selbst-überwachte Lernen primär bildspezifische Intensitätsvariationen als Merkmale berücksichtigt, d.h. ein Art „Fingerabdruck“ zur Wiedererkennung von gleichen koronare Schichten ermittelt und damit für die eigentlich intendierte Krankheitserkennung wenig geeignet ist.

2.2 Evaluation und Charakterisierung neuartiger Bildgebungsmarker

Vor den oben genannten Arbeiten mit künstlichen neuronalen Netzen haben wir in zahlreichen Studien experimentelle Bildgebungstechniken in großen retrospektiven sowie prospektiven multizentrischen Studien bezüglich ihrer diagnostischen Eignung evaluiert. Hierzu wurde größtenteils die Elastic-Net regularisierte Regression verwendet.

In der **Originalarbeit OA3** wurde die Diffusions-Tensor-Bildgebung untersucht, die Auskünfte über die Eigenschaften der Mikrostruktur des Nervengewebes oder der Nervenfasern gibt. Besonderes Interesse galt hier der Charakterisierung von Gruppenunterschieden bzw. Diffusionsveränderungen in der grauen Substanz unter Berücksichtigung von sogenannten Partialvolumeneffekten. Hierbei existiert das Problem, dass ein Teil der beobachteten Diffusionsveränderungen nicht aus den

eigentlichen Gewebeeigenschaften resultiert, sondern von den im Messpunkt (Voxel) ebenfalls enthaltenen Anteilen des Liquor cerebrospinalis im Randbereich der grauen Substanz. In der Arbeit wurde ein Verfahren zur Bereinigung von Diffusionswerten nach [40, 41] angewandt. Unsere Ergebnisse zeigten, dass die bereinigte mittlere Diffusion eine weitaus geringere Effektstärke im Kontinuum der Alzheimer-Krankheit aufwies und ein Großteil der Gruppenunterschiede vor allem auf die Atrophie der grauen Substanz zurückzuführen ist.

In den **Originalarbeiten OA4, OA5, OA6 und OA7** wurden die multizentrische Diffusionsbildgebung und die funktionelle MRT im Ruhezustand bezüglich der diagnostischen Eignung im Kontinuum der Alzheimer-Krankheit untersucht und insbesondere in Bezug auf die Höhe der Effektstärken im Kontext des Signal-Rausch-Verhältnisses multizentrischer Datenerhebung. Für **OA4 und OA5** wurden retrospektive funktionelle MRT Datensätze im Rahmen der *German Resting-State Initiative: PsyMRI – Dementia Arm* zusammengestellt. Hierbei wurden systematisch Indikatoren und Marker zur Erhebung der Datenqualität evaluiert. Von den zahlreichen getesteten Verfahren schienen jedoch nur wenige geeignet, Aufschluss über offensichtliche Qualitätsmängel zu geben. Darunter waren die Kennzahlen *temporal Signal to Noise Ratio*, die die Varianz der MRT-Aufnahmen im zeitlichen Verlauf bemisst, sowie *DVARS* oder *Framewise Displacement*, die Auskunft über den Umfang der Kopfbewegungen geben. Insgesamt zeigten sich die statistischen Gruppeneffekte zwischen den Zentren als sehr heterogen. Bei der automatisierten Gruppentrennung mit dem Elastic-Net regularisierten Regressionsverfahren erreichte die Trennung von Alzheimer-Demenz und Kontrollen eine AUC von 0.80 anhand der funktionellen Konnektivität.

In **OA6 und OA7** wurden Gruppentrennung und Effektstärken multimodaler MRT-Bildgebung im Rahmen der prospektiven multizentrischen *Longitudinalen Studie zu Kognitiven Beeinträchtigungen und Demenz (DELCODE)* des DZNE erhoben. Bei umfangreichen Auswertungen der Baseline Bildgebungsdaten in **OA6** konnte gezeigt werden, dass die Verwendung harmonisierter Prozeduren und Akquisitionsprotokolle die Homogenität der funktionellen MRT Aufnahmen substanziell verbessert. Bei Alzheimer-Patienten vs. Kontrollen mit positivem bzw. negativem Liquorbefund für Amyloid-Beta konnte für die funktionelle Konnektivität eine AUC von 0.88 erreicht werden, bei Patienten mit leichter kognitiver Störung vs. Kontrollen (ebenso mit pos. bzw. neg. Liquorbefund) eine AUC von 0.69. Demgegenüber stehen jedoch AUC-Werte in der Höhe von 0.94 und 0.78 für die gleichen Gruppenvergleiche und das Hippocampusvolumen[3].

[3] Differenzen in der AUC zwischen funktioneller Konnektivität und Hippocampusvolumen sind statistisch nicht signifikant aufgrund relativ hoher Variabilität der Ergebnisse.

In **OA7** konnte gezeigt werden, dass bereits früh im Kontinuum der Alzheimer-Krankheit Diffusionsveränderungen in den Fasertrakten auftreten. Während die Gruppentrennung in den Stadien der leichten kognitiven Störung und Demenz annähernd vergleichbar sind mit der Differenzierung anhand des Hippocampusvolumens, zeigen erste Ergebnisse eine bessere Unterscheidbarkeit von Probanden mit subjektiver kognitiver Beeinträchtigung und kognitiv gesunden Probanden mit einer AUC von 0.69 für Diffusionsmarker vs. 0.62 für Hippocampusvolumen. Jedoch konnte für die vorliegende Stichprobe nicht geklärt werden, ob diese Unterschiede eine psychologische Disposition (*Trait der Besorgtheit*) oder tatsächlich krankheitsbezogene Veränderungen anzeigen.

In der **Originalarbeit OA8** wurde der Marker BrainAge bei Patienten mit amyotropher Lateralsklerose (ALS) untersucht. Bei diesem radiologischen Marker wird anhand der Hirnstruktur, z.B. dem Volumen und der Verteilung der grauen Substanz, das Alter einer betreffenden Person geschätzt. Hierzu wurde das multivariate Gaußprozess-Regressionsmodell BrainAgeR (Version 2.1) verwendet, welches auf einer multizentrischen Normstichprobe mit N = 3377 gesunden Probanden trainiert worden war [42]. In der Literatur wurde für eine Reihe von neurodegenerativen Erkrankungen beschrieben, dass das BrainAge deutlich höher im Vergleich zum biologischen Alter geschätzt wird, z.B. das „Hirnalter“ bei Patienten mit einer Alzheimer-Erkrankung 8 bis 10 Jahre älter erscheint [43]. In OA8 wurde untersucht, welches BrainAge Patienten mit ALS mit und ohne Beeinträchtigung der Kognition aufweisen. Entsprechend der Hypothese zeigten ALS-Patienten mit kognitiver Beeinträchtigung und Patienten mit zusätzlicher frontotemporaler Demenz ein höheres BrainAge mit einer mittleren Altersdifferenz von +2,3 (SD 6,4) bzw. +8,6 (SD 9,2) Jahren. Im Gegensatz zur Hypothese zeigten ALS-Patienten ohne kognitive Beeinträchtigungen eine mittlere Altersdifferenz von -4,3 (SD 5,8) Jahren, d.h. hier erschien das „Hirnalter“ jünger als bei gesunden Kontrollprobanden. Zusätzliche Analysen zeigten, dass die Altersdifferenz mit der Krankheitsprogression sowie der Überlebenszeit korrelierte.

2.3 Multivariate Modellierung statistischer Zusammenhänge von Schädigungsmustern

Beim Gehirn wirken viele Prozesse (z.B. Entwicklung, Alterung, Krankheit) auf die gesamte Struktur, so dass einzelne Regionen häufig Kollinearität aufweisen, d.h. die Messwerte ähneln sich und teilen zu einem gewissen Grad ihre Varianz. Die geteilte Varianz führt jedoch zu störenden Korrelationen, da die Erhebung gewünschter statistischer Zusammenhänge (Kovarianz) überlagert und somit erschwert wird. Der Ursprung der geteilten Varianz kann mit klassischen statistischen Verfahren nicht eindeutig zugeordnet werden. In Abgrenzung zu latenten Variablenmodellen, bei denen man solche „latenten“ Variablen explizit modelliert und bestimmen möchte, versuchen die Gauß‘schen Graphenmodelle die Kovarianzmatrix einer Reihe von beobachteten Variablen zu invertieren, um die sogenannte „Precision

Matrix“ zu bestimmen, die die *partiellen* Korrelationen enthält, also den bereinigten „wahren“ statistischen Zusammenhang zwischen Variablen beschreibt.[4] Die allgemeine Matrixinvertierung ist mathematisch schwierig und kann häufig nicht exakt gelöst werden[5] [44], weshalb die Schätzungen der partiellen Korrelationen numerisch instabil ist. Zur Abschwächung dieses Problems existieren in der Literatur verschiedene Näherungsverfahren. In den **Originalarbeiten OA9** und **OA10** wurde der Zusammenhang von Amyloid-Ablagerungen, Glukose-Hypometabolismus und Atrophie der grauen Substanz modelliert und evaluiert. Dazu wurde das R-Paket BDgraph [45] verwendet und erweitert, die ein Markov-Chain-Monte-Carlo Verfahren nutzt, um die statistische Verteilung der beobachteten Variablen bzw. der zugrundeliegenden Precision Matrix abzuschätzen.

Für **OA9** wurden sechs bei der Alzheimer-Erkrankung besonders stark betroffene Hirnregionen ausgewählt und in diesen die Amyloid-Ablagerung, der Glukose-Metabolismus und das Volumen der grauen Substanz gemessen. Mittels BDgraph wurde datengetrieben die partielle Korrelation und Graphenstruktur zwischen den Regionen und Modalitäten ermittelt sowie anschließend mit drei a-priori definierten Ausbreitungshypothesen der Alzheimer-Krankheit verglichen. Die ermittelten Graphenstrukturen entsprachen hierbei am ehesten der sogenannten „Wear and Tear“ Hypothese [46], bei der einzelne zentrale Knotenregionen (Hubs) im Zentrum stehen, diese am stärksten betroffen sind und mit zahlreichen peripheren Regionen assoziiert.

In **OA10** wurde die Anzahl der Regionen auf das gesamte Gehirn erweitert, so dass je Modalität 54 Variablen bestimmt wurden. Für die vier diagnostischen Gruppen: Gesunde, frühe leichte kognitive Beeinträchtigung (EMCI), späte leichte kognitive Beeinträchtigung (LMCI) und Alzheimer-Demenz wurden jeweils die Precision Matrizen ermittelt und Graphenmetriken miteinander verglichen. Hierbei zeigte sich bei den beiden MCI Subgruppen eine Erhöhung der Matrixdichte, Clustering und Small-World Koeffizienten bzw. Verringerung der charakteristischen Pfadlänge (im Vergleich zu Gesunden und Demenzpatienten), welche auf höhere und häufigere partielle Korrelationen in diesen Gruppen zurückzuführen ist.

[4] Siehe dazu auch das vereinfachte Beispiel in *Figure 1* von OA10.

[5] Häufig ist die Kovarianzmatrix einer Stichprobe schlecht konditioniert und daher schwer zu invertieren, d.h. bereits sehr kleine Veränderungen der Eingabedaten führen zu großen Änderungen in der Precision Matrix [44].

3 Diskussion

3.1 Erklärbarkeit von Entscheidungen in tiefen neuronalen Netzen

Mit den eigenen Arbeiten zur Nachvollziehbarkeit und Erklärbarkeit von Entscheidungen in tiefen neuronalen Netzen haben wir einen wesentlichen Beitrag zur klinischen Anwendbarkeit von künstlichen neuronalen Netzen geleistet. Einige Vorarbeiten erprobten einzelne Visualisierungsverfahren empirisch, z.B. Layer-wise Relevance Propagation [47], Guided Backpropagation [47] oder Salienzkarten [48]. Bislang fehlte jedoch eine systematische Übersicht der verschiedenen Verfahren und eine Bewertung in Bezug auf deren Eignung für klinische Endnutzer. Insbesondere die Umsetzung einer Web-Applikation zur Auswertung von MRT-Datensätzen mit neuronalen Netzmodellen sowie Berechnung und Visualisierung von Relevanzkarten demonstriert das Potential dieser Verfahren für zukünftige Studien und radiologische Assistenzsysteme. Dagegen ist zu bemerken, dass die Anwendung von Relevanzattributionsmethoden bislang noch nicht in der Breite etabliert ist und sich zahlreiche aktuell erscheinende Publikationen lediglich auf die Bewertung der Leistung anhand der Metriken Erkennungsrate und F1-Score beschränken. Als mahnendes Beispiel sei hier die renommierte Studie [49] genannt, deren Modelle zur Alzheimer-Demenzerkennung neben klinisch relevanten Regionen (Hippocampus und posteriorer cingulärer Cortex) auch einen Teil des Schädelknochens sowie zwei Randbereiche der MRT-Aufnahmen (außerhalb des Kopfes) bei der Berechnung der Modellvorhersage berücksichtigen[6], ein Indiz für Bias und Überanpassung der Modelle an die Trainingsdaten.

Perspektivisch umfasst die zukünftige Arbeit in diesem Forschungsfeld die Etablierung von Ansätzen zur Quantifizierung und zum Vergleich der *Qualität* von Relevanzkarten. Hierzu wurden bisher Proxy-Metriken verwendet, in unserem Fall (OA1) die Pearson-Korrelation von Relevanzwerten mit bekannten Markern wie dem Hippocampusvolumen oder in [50] die Übereinstimmung von Relevanzkarten mit meta-analytisch bestimmten statistischen Gruppenunterschieden (Dice-Koeffizient). Weitere Ansätze werden im laufenden DFG-Projekt „ExplAInation“ untersucht.

3.2 Evaluation von experimentellen Bildgebungsmarkern

Die im Rahmen dieses Habilitationsverfahrens untersuchten multizentrischen Datensätze und experimentellen Bildgebungs- und Analyseverfahren geben uns Aufschluss über die klinische Anwendbarkeit und den Nutzen solcher Verfahren im Kontext der Demenzdiagnostik. Als wesentliche Erkenntnis konnten aus den multizentrischen Studien geschlussfolgert werden, dass die Unterschiede zwischen den Akquisitionsprotokollen und -prozeduren zwischen einzelnen Studienzenten einen

[6] Eigene Analysen der 3D Modelle aus [49], verfügbar unter https://doi.org/10.5281/zenodo.3491002, mit einem Patienten mit Demenz und einem gesunden Kontrollprobanden. Ergebnisse nicht veröffentlicht.

substantiellen Einfluss auf die erhobenen Daten haben und die Vergleichbarkeit von Ergebnissen sowie Effektstärken reduzieren. Bei der prospektiven multizentrischen DZNE-Studie DELCODE wurde durch eine stringente Abstimmung der Akquisitionsprotokolle der Zentrumseffekt im Vergleich zu retrospektiv erhobenen Datensätzen deutlich reduziert. Bei der Evaluation der diagnostischen Eignung konnte ein vergleichbares Niveau der Gruppentrennung von Diffusionsveränderungen und Hippocampusvolumen ermittelt werden. Jedoch ist dieses Bildgebungsverfahren aufgrund des Aufwands und der höheren Störanfälligkeit und Komplexität des Messverfahrens für die Demenzdiagnostik derzeit nicht zu empfehlen. Gleiches gilt für die funktionelle MRT im Ruhezustand, die im Vergleich zum Hippocampusvolumen eine schlechtere Erkennungsrate ermöglicht sowie eine schlechtere Effektstärke aufweist.

3.3 Modellierung von komplexen statistischen Zusammenhängen

Die Gauß'schen Graphenmodelle stellten einen innovativen Ansatz zur empirischen Erhebung von Kovarianzstrukturen und zur Abschätzung der Precision Matrix mit den partiellen Korrelationen dar. In Studien mit funktionellen MRT-Daten fanden diese Verfahren bereits einige Male Anwendung zur Erhebung der funktionellen Konnektivität [51-54]. In Übereinstimmung mit unseren Ergebnissen fanden die Autoren ein hohes Potential des Verfahrens, das Kollinearitätsproblem zu lösen, da die partielle Kovarianzmatrix signifikant dünner belegt ist (engl. *Sparse Matrix*). Im Gruppenvergleich stellten sich signifikante Gruppenunterschiede in den Graphenmetriken Clustering Koeffizient und charakteristische Pfadlänge dar, die für die Subgruppen mit leichter kognitiver Störung signifikant erhöht bzw. reduziert waren. Jedoch konnte nicht geklärt werden, ob diese Beobachtung auf eine höhere Dynamik von Krankheitsprozessen zurückzuführen ist oder auf die Heterogenität der betreffenden Stichprobe.[7] Letztendlich erwies sich der verwendete Algorithmus BDgraph zur Abschätzung der partiellen Kovarianz als Basis für die Graphenmodellierung für den angedachten Einsatzzweck als nur mäßig geeignet, da die Ergebnisse teilweise numerisch instabil waren und eine hohe Variabilität aufwiesen.

3.4 Konklusion und Ausblick

Im Rahmen dieses Habilitationsvorhabens wurden zahlreiche Studien zum übergreifenden Thema automatisierte Erkennung von Demenzerkrankungen durchgeführt und drei Schwerpunkte bearbeitet. Im Bereich der Generierung von visuellen Erklärungen für tiefe neuronale Faltungsnetze konnten Relevanz-Mapping-Verfahren etabliert und deren Potential für die Nutzung bei klinischen Fragestellungen demonstriert werden. Für die multimodale multizentrische Bildgebung wurde die

[7] Es ist davon auszugehen, dass in der verwendeten Stichprobe nicht alle Patienten mit leichter kognitiver Störung eine Alzheimer-Erkrankung aufweisen.

Gruppentrennung und Effektstärken für die funktionelle MRT und Diffusions-Tensor-Bildgebung charakterisiert. Derzeit untersuchen wir aufbauend auf den Ergebnissen der Ausgangsuntersuchung (Baseline) die Verlaufscharakteristik und den prädiktiven Wert multimodaler Bildgebung im Rahmen der longitudinalen Erhebung der DZNE Studien DELCODE und DESCRIBE. Für die Modellierung von komplexen statistischen Zusammenhängen wurden Ansätze basierend auf Gauß'schen Graphenmodellen zur Abschätzung der partiellen Kovarianz in multimodalen Bildgebungsdaten evaluiert. In der Zusammenschau leisteten wir mit diesen Arbeiten einen substanziellen Beitrag zum gegenwärtigen Stand der Wissenschaft, wie die hohen Zitationszahlen der jeweiligen Studien belegen.

Ausgehend von diesen Vorarbeiten erforschen wir derzeit Verfahren zur Repräsentation, Modellierung und Integration von medizinischem Vorwissen in computative KI-Systemarchitekturen zur Generierung von visuellen und textuellen Erklärungen für tiefe neuronale Netze im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projekts „ExplAInation" (Förderkennzeichen DY151/2-1, Projektnr. 454834942, Laufzeit 2021–2024). Weiter untersuchen wir auf Prozessebene Ansätze zur verteilten Bereitstellung von diagnostischen KI-Dienstleistungen im Netzwerk Universitätsmedizin (NUM) im Verbundprojekt „Open Medical Inference" (OMI) der Medizininformatik-Initiative des Bundesministeriums für Bildung und Forschung (Förderkennzeichen 01 ZZ 2315L, Laufzeit 2023–2027) sowie die Erarbeitung von Empfehlungen und Best-Practice-Ansätzen für die Einführung von KI-Assistenzsystemen in Kliniken im Ostseeraum, Verbundprojekt „Clinical AI-based Diagnostics" (CAIDX), gefördert im Rahmen des EU InterReg Baltic Sea Region Programms (Projektnummer #C005, Laufzeit 2023–2025). Perspektivisch stellt sich ebenso die Frage, welchen Einfluss diagnostische KI-Assistenzsysteme zukünftig auf die Rolle von Ärzten und das Verhältnis zwischen Patienten und Ärzten haben werden. Diese Aspekte werden im Verbundvorhaben „Theoretische, ethische und soziale Implikationen von KI für neuropsychiatrische Forschung und Praxis" (TESIComP) untersucht, gefördert vom Bundesministerium für Bildung und Forschung (Förderkennzeichen 01 GP 2216B, Laufzeit 2023–2026).

4 Publikationen

Die folgenden Publikationen sind Bestandteil dieser Habilitation und im Anhang abgedruckt.

***OA**: Originalarbeiten in internationalen Fachzeitschriften. **KB**: Methodikbezogene Konferenzbeiträge.*

Erklärbarkeit von Entscheidungen von künstlichen neuronalen Netzen

KB1 **Dyrba M**, Pallath AH, Marzban EN. 2020. Comparison of CNN Visualization Methods to Aid Model Interpretability for Detecting Alzheimer's Disease. In: Tolxdorff T, Deserno TM, Handels H, editors. *Bildverarbeitung für die Medizin 2020. Proceedings des Workshops vom 15. bis 17. März 2020 in Berlin*. Springer Fachmedien Wiesbaden. p. 307-312. DOI: 10.1007/978-3-658-29267-6_68.

KB2 **Dyrba M**, Hanzig M. 2021. Interactive Visualization of 3D CNN Relevance Maps to Aid Model Comprehensibility. In: Palm C, Deserno TM, Handels H, Maier A, Maier-Hein K, Tolxdorff T, editors. *Bildverarbeitung für die Medizin 2021. Proceedings des Workshops vom 7. bis 9. März 2021 in Regensburg.* Springer Fachmedien Wiesbaden. p. 317-322. DOI: 10.1007/978-3-658-33198-6_77.

OA1 **Dyrba M**, Hanzig M, Altenstein S, Bader S, Ballarini T, Brosseron F, Buerger K, Cantré D, Dechent P, Dobisch L, Düzel E, Ewers M, Fliessbach K, Glanz W, Haynes JD, Heneka MT, Janowitz D, Keles DB, Kilimann I, Laske C, Maier F, Metzger CD, Munk MH, Perneczky R, Peters O, Preis L, Priller J, Rauchmann B, Roy N, Scheffler K, Schneider A, Schott BH, Spottke A, Spruth EJ, Weber MA, Ertl-Wagner B, Wagner M, Wiltfang J, Jessen F, Teipel SJ. 2021. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. *Alzheimer's Research & Therapy*. 13(1):191. DOI: 10.1186/s13195-021-00924-2. PMID: 34814936. Impact Factor: 6.982.

KB3 Singh D, **Dyrba M**. 2023. Comparison of CNN Architectures for Detecting Alzheimer's Disease using Relevance Maps. In: *Bildverarbeitung für die Medizin 2023. Proceedings des Workshops vom 2. bis 4. Juli 2023 in Braunschweig.* Springer Fachmedien Wiesbaden. p. 238-243. DOI: 10.1007/978-3-658-41657-7_51.

OA2 Gryshchuk V, Teipel SJ, **Dyrba M**. Contrastive Self-supervised Learning for Neurodegenerative Disorder Classification. *In Begutachtung*.

Evaluation und Charakterisierung neuartiger Bildgebungsmarker

OA3 Henf J, Grothe MJ, Brueggen K, Teipel S, **Dyrba M**. 2017. Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction. *Neuroimage Clinical*. 17:579-586. DOI: 10.1016/j.nicl.2017.10.005. PMID: 29201644. Impact Factor: 3.869.

OA4 Teipel SJ, Wohlert A, Metzger C, Grimmer T, Sorg C, Ewers M, Meisenzahl E, Klöppel S, Borchardt V, Grothe MJ, Walter M, **Dyrba M**. 2017. Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI. *Neuroimage Clinical*. 14:183-194. DOI: 10.1016/j.nicl.2017.01.018. PMID: 28180077. Impact Factor: 4.348.

OA5 Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Klöppel S, Borchardt V, Walter M, **Dyrba M**. 2017. Robust Detection of Impaired Resting State Functional Connectivity Networks in Alzheimer's Disease Using Elastic Net Regularized Regression. *Frontiers in Aging Neuroscience*. 8:318. DOI: 10.3389/fnagi.2016.00318. PMID: 28101051. Impact Factor: 4.504.

OA6 Teipel SJ, Metzger CD, Brosseron F, Buerger K, Brueggen K, Catak C, Diesing D, Dobisch L, Fliebach K, Franke C, Heneka MT, Kilimann I, Kofler B, Menne F, Peters O, Polcher A, Priller J, Schneider A, Spottke A, Spruth EJ, Thelen M, Thyrian RJ, Wagner M, Düzel E, Jessen F, **Dyrba M**. 2018. Multicenter Resting State Functional Connectivity in Prodromal and Dementia Stages of Alzheimer's Disease. *Journal of Alzheimer's Disease*. 64(3):801-813. DOI: 10.3233/JAD-180106. PMID: 29914027. Impact Factor: 3.476.

OA7 Teipel SJ, Kuper-Smith JO, Bartels C, Brosseron F, Buchmann M, Buerger K, Catak C, Janowitz D, Dechent P, Dobisch L, Ertl-Wagner B, Fließbach K, Haynes JD, Heneka MT, Kilimann I, Laske C, Li S, Menne F, Metzger CD, Priller J, Pross V, Ramirez A, Scheffler K, Schneider A, Spottke A, Spruth EJ, Wagner M, Wiltfang J, Wolfsgruber S, Düzel E, Jessen F, **Dyrba M**. 2019. Multicenter Tract-Based Analysis of Microstructural Lesions within the Alzheimer's Disease Spectrum: Association with Amyloid Pathology and Diagnostic Usefulness. *Journal of Alzheimer's Disease*. 72(2):455-465. DOI: 10.3233/JAD-190446. PMID: 31594223. Impact Factor: 3.517.

OA8 Hermann A, Tarakdjian GN, Temp AGM, Kasper E, Machts J, Kaufmann J, Vielhaber S, Prudlo J, Cole JH, Teipel S, **Dyrba M**. 2022. Cognitive and behavioural but not motor impairment increases brain age in amyotrophic lateral sclerosis. *Brain Communications*. 4(5):fcac239. DOI: 10.1093/braincomms/fcac239. PMID: 36246047. Impact Factor 4.400.

Multivariate Modellierung statistischer Zusammenhänge von Schädigungsmustern

OA9 **Dyrba M**, Grothe MJ, Mohammadi A, Binder H, Kirste T, Teipel SJ; Alzheimer's Disease Neuroimaging Initiative. 2018. Comparison of Different Hypotheses Regarding the Spread of Alzheimer's Disease Using Markov Random Fields and Multimodal Imaging. *Journal of Alzheimer's Disease*. 65(3):731-746. DOI: 10.3233/JAD-161197. PMID: 28697557. Impact Factor: 3.731.

OA10 **Dyrba M**, Mohammadi R, Grothe MJ, Kirste T, Teipel SJ. 2020. Gaussian Graphical Models Reveal Inter-Modal and Inter-Regional Conditional Dependencies of Brain Alterations in Alzheimer's Disease. *Frontiers in Aging Neuroscience*. 12:99. DOI: 10.3389/fnagi.2020.00099. PMID: 32372944. Impact Factor: 4.362.

5 Literaturverzeichnis

1. Schmidtke, K. and M. Otto, *Alzheimer-Demenz*, in *Demenzen*, C.-W. Wallesch and H. Förstl, Editors. 2012, Thieme. p. 203–227.
2. Jack, C.R., et al., *Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.* Alzheimer's & Dementia, 2011. 7(3): p. 257-262.
3. Bickel, H., et al., *Präsenile Demenzen in Gedächtnisambulanzen — Konsultationsinzidenz und Krankheitscharakteristika.* Der Nervenarzt, 2006. 77(9): p. 1079-1085.
4. Leroy, M., et al., *Characteristics and progression of patients with frontotemporal dementia in a regional memory clinic network.* Alzheimer's Research & Therapy, 2021. 13(1).
5. Hogan, D.B., et al., *The Prevalence and Incidence of Frontotemporal Dementia: a Systematic Review.* Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques, 2016. 43(S1): p. S96-S109.
6. Rabinovici, G.D. and B.L. Miller, *Frontotemporal Lobar Degeneration.* CNS Drugs, 2010. 24(5): p. 375-398.
7. Eichler, T., et al., *Rates of formal diagnosis of dementia in primary care: The effect of screening.* Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2015. 1(1): p. 87-93.
8. Bickel, H., et al., *The Prevalence of Dementia and Cognitive Impairment in Hospitals.* Deutsches Ärzteblatt international, 2018.
9. Wippold, F.J., et al., *Neuropathology for the Neuroradiologist: Plaques and Tangles.* American Journal of Neuroradiology, 2008. 29(1): p. 18–22.
10. Mattson, M.P., *Pathways towards and away from Alzheimer's disease.* Nature, 2004. 430(7000): p. 631–639.
11. Prudlo, J., et al., *TDP-43 pathology and cognition in ALS.* Neurology, 2016. 87(10): p. 1019-1023.
12. Huk, W.J. and G. Gademann, *Magnetic resonance imaging (MRI): method and early clinical experiences in diseases of the central nervous system.* Neurosurgical review, 1984. 7(4): p. 259–280.
13. Le Bihan, D., et al., *Diffusion MR imaging: clinical applications.* American Journal of Roentgenology, 1992. 159(3): p. 591–599.
14. Herholz, K., et al., *Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET.* Neuroimage, 2002. 17(1): p. 302-316.
15. Li, Y., et al., *Regional analysis of FDG and PIB-PET images in normal aging, mild cognitive impairment, and Alzheimer's disease.* European Journal of Nuclear Medicine and Molecular Imaging, 2008. 35(12): p. 2169–2181.
16. Klunk, W.E., et al., *Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B.* Annals of Neurology, 2004. 55(3): p. 306–319.
17. Ogawa, S., et al., *Brain magnetic resonance imaging with contrast dependent on blood oxygenation.* Proceedings of the National Academy of Sciences, 1990. 87(24): p. 9868–9872.
18. Kwong, K.K., et al., *Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation.* Proceedings of the National Academy of Sciences, 1992. 89(12): p. 5675–5679.
19. Biswal, B., et al., *Functional connectivity in the motor cortex of resting human brain using echo-planar MRI.* Magn Reson Med, 1995. 34(4): p. 537–541.
20. Fox, M.D., *The human brain is intrinsically organized into dynamic, anticorrelated functional networks.* Proceedings of the National Academy of Sciences, 2005. 102(27): p. 9673–9678.
21. Biswal, B.B., J.V. Kylen, and J.S. Hyde, *Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps.* NMR in Biomedicine, 1997. 10(4/5): p. 165–170.

22. Cordes, D., et al., *Frequencies Contributing to Functional Connectivity in the Cerebral Cortex in "Resting-state" Data.* American Journal of Neuroradiology, 2001. 22(7): p. 1326–1333.
23. Sorg, C., et al., *Selective changes of resting-state networks in individuals at risk for Alzheimer's disease.* Proceedings of the National Academy of Sciences, 2007. 104(47): p. 18760–18765.
24. Supekar, K., et al., *Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease.* PLoS Computational Biology, 2008. 4(6): p. e1000100.
25. Sanz-Arigita, E.J., et al., *Loss of 'Small-World' Networks in Alzheimer's Disease: Graph Analysis of fMRI Resting-State Functional Connectivity.* PLoS ONE, 2010. 5(11): p. e13788.
26. Chen, G., et al., *Classification of Alzheimer Disease, Mild Cognitive Impairment, and Normal Cognitive Status with Large-Scale Network Analysis Based on Resting-State Functional MR Imaging.* Radiology, 2011. 259(1): p. 213–221.
27. Petrella, J.R., et al., *Default mode network connectivity in stable vs progressive mild cognitive impairment.* Neurology, 2011. 76(6): p. 511–517.
28. Koch, W., et al., *Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease.* Neurobiology of Aging, 2012. 33(3): p. 466–478.
29. Brier, M.R., et al., *Loss of Intranetwork and Internetwork Resting State Functional Connections with Alzheimer's Disease Progression.* Journal of Neuroscience, 2012. 32(26): p. 8890–8899.
30. Soldner, J., et al., *Strukturelle und funktionelle neuronale Konnektivität bei der Alzheimer-Krankheit.* Der Nervenarzt, 2012. 83(7): p. 878–887.
31. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. 67(2): p. 301-320.
32. Zentrale Ethikkommission der Bundesärztekammer, *Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz.* Deutsches Ärzteblatt, 2021. 118(33-34).
33. Murdoch, W.J., et al., *Definitions, methods, and applications in interpretable machine learning.* Proceedings of the National Academy of Sciences, 2019. 116(44): p. 22071-22080.
34. Belle, V. and I. Papantonis, *Principles and Practice of Explainable Machine Learning.* Frontiers in Big Data, 2021. 4.
35. Bach, S., et al., *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.* Plos One, 2015. 10(7).
36. Montavon, G., W. Samek, and K.-R. Müller, *Methods for interpreting and understanding deep neural networks.* Digital Signal Processing, 2018. 73: p. 1-15.
37. Lecun, Y., et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. 86(11): p. 2278-2324.
38. Deng, J., et al., *ImageNet: A large-scale hierarchical image database*, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009. p. 248-255.
39. Lin, T.-Y., et al., *Microsoft COCO: Common Objects in Context*, in *Computer Vision – ECCV 2014*. 2014. p. 740-755.
40. Koo, B., et al., *A framework to analyze partial volume effect on gray matter mean diffusivity measurements.* NeuroImage, 2009. 44(1): p. 136-144.
41. Jeon, T., et al., *Regional changes of cortical mean diffusivities with aging after correction of partial volume effects.* NeuroImage, 2012. 62(3): p. 1705-1716.
42. Cole, J.H., et al., *Brain age predicts mortality.* Molecular Psychiatry, 2017. 23(5): p. 1385-1392.
43. Franke, K. and C. Gaser, *Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?* Frontiers in Neurology, 2019. 10.
44. Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 2007. 9(3): p. 432-441.
45. Mohammadi, R. and E.C. Wit, *BDgraph: An R Package for Bayesian Structure Learning in Graphical Models.* Journal of Statistical Software, 2019. 89(3).

46. Zhou, J., et al., *Predicting Regional Neurodegeneration from the Healthy Brain Functional Connectome.* Neuron, 2012. 73(6): p. 1216-1227.
47. Böhle, M., et al., *Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification.* Frontiers in aging neuroscience, 2019. 11: p. 194.
48. Ding, Y., et al., *A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain.* Radiology, 2019. 290(2): p. 456-464.
49. Wen, J., et al., *Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation.* Medical Image Analysis, 2020. 63.
50. Wang, D., et al., *Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies.* Neuroimage, 2023. 269: p. 119929.
51. Chen, T., et al., *Estimation of resting-state functional connectivity using random subspace based partial correlation: A novel method for reducing global artifacts.* NeuroImage, 2013. 82: p. 87–100.
52. Fiecas, M., et al., *Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fMRI.* NeuroImage, 2013. 65: p. 231–241.
53. Brier, M.R., et al., *Partial covariance based functional connectivity computation using Ledoit–Wolf covariance regularization.* NeuroImage, 2015. 121: p. 29-38.
54. Wang, Y., et al., *An Efficient and Reliable Statistical Method for Estimating Functional Connectivity in Large Scale Brain Networks Using Partial Correlation.* Frontiers in Neuroscience, 2016. 10.

Selbstständigkeitserklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Verwendung anderer als den angegebenen Hilfsmitteln angefertigt habe. Es wurden keine KI-Modelle zur Synthese von Texten verwendet. Die aus anderen Quellen direkt oder indirekt übernommenen Informationen und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Regeln zur Sicherung guter wissenschaftlicher Praxis wurden beachtet.

Ich versichere, dass ich für die inhaltliche Erstellung der vorliegenden Arbeit nicht die entgeltliche Hilfe von Beratungs- und Autorendiensten in Anspruch genommen habe. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Habilitation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Dr. Martin Dyrba

Rostock, den 17. Juni 2023

Lebenslauf und wissenschaftlicher Werdegang

Dr. rer. hum. Martin Dyrba

Wallstr. 10 • 18258 Schwaan

Geboren am 5. August 1985 in Zwenkau

Verheiratet, vier Kinder

Kontakt: 0381 494 9482
0162 69 54 038
martin.dyrba@dzne.de

Aktuelle Position

Nachwuchsgruppenleiter Deep Learning Erklärbarkeit
Arbeitsgruppe Klinische Demenzforschung
Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)
Standort Rostock/Greifswald

Ausbildung

Jahr	Abschluss	Fach und Thema	Institution
2016–2024	Habilitation	Medizinische Informatik: *Verfahren der künstlichen Intelligenz zur automatisierten Erkennung von Demenz-erkrankungen*	Klinik für Psychosomatik, Universitätsmedizin Rostock
2011–2016	Dr. rer. hum.	Medizinische Informatik: *Automatisierte Erkennung manifester und prodromaler Alzheimer-Krankheit mittels multizentrisch akquirierter, multimodaler Bilddaten*	Klinik für Psychosomatik, Universitätsmedizin Rostock
2005–2011	Diplom in Informatik	Ubiquitous computing: *Design and Implementation of a SmartLab Controller Controller Based on the Subsumption Architecture*	Institute of Visual and Analytic Computing, Universität Rostock
1996–2004	Abitur		Elbe-Gymnasium Boizenburg

Wissenschaftlicher Werdegang und Praxiserfahrung

Seit Mai 2023	Lehrverantwortlicher für den Studiengang Medizinische Informatik, Europäische Fernhochschule Hamburg GmbH (freiberuflich, ca. 20% Teilzeit)
Seit Juli 2021	Betreuung von Vadym Gryshchuk (2021) und Devesh Singh (ab 2022), Doktoranden und wissenschaftliche Mitarbeiter im DFG-Projekt „*Eine Neurale Netzwerke System-architektur für diagnostische Erklärungen*“
2019–2020	Gasteditor bei Frontiers in Aging Neuroscience, Research Topic: *Deep Learning in Aging Neuroscience*
Seit 2019	Freiberuflicher IT-Berater (Teilzeit), Webseitenerstellung, technische Kunden-betreuung und Software-Gerätewartung (www.it-beratung-dyrba.de)
2017–2018	Betreuung von Eman N. Marzban, DAAD-Stipendiatin und Gastdoktorandin der Universität Kairo, Ägypten, zu Deep Learning Modellen und Visualisierungen zur Erkennung der Alzheimer-Krankheit
Seit 2013	Gutachter für internationale Fachzeitschriften, z.B. *Alzheimer's & Dementia, The Lancet Digital Health, Frontiers in Aging Neuroscience* (>125 Reviews)
Seit 2011	Wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Klinische Demenzforschung unter der Leitung von Prof. Dr. Stefan Teipel, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Rostock
2010–2011	Diplomarbeit zur Integration dynamischer Geräteensembles in smarten Besprechungsräumen, am Lehrstuhl für Mobile Multimediale Informationssysteme (MMIS) von Prof. Dr. Thomas Kirste, Institute of Visual & Analytic Computing, Universität Rostock
2009	Praktikumssemester bei der Arivis AG, Rostock, Java Programmierung von skalierbaren Nutzeroberflächen im Eclipse Framework; richtlinienbasierte Software-Testung der Vision 4D Visualisierungssoftware für hochdimensionale Bilddaten
Juli 2008 & Oktober 2009	Volontär für die Eckart Schwerin Stiftung, Linux Server/Client Systemeinrichtung und Lehrerschulung an 5 weiterführenden Schulen in der Pare Diözese, Same, Tansania
2005–2010	Studentische Hilfskraft für die Entwicklung einer Steuerungsinfrastruktur für verteilte und heterogene Geräteensembles in smarten Besprechungsräumen im Rahmen des DFG Graduiertenkollegs MuSAMA; Server/Client und GUI Programmierung in Java
2005–2009	Fachschaftsvertreter des Studiengangs Informatik an der Universität Rostock, 2008–2009 Fachschaftssprecher, studentisches Mitglied im Institutsbeirat und der Neubaukommission
2004–2005	Zivildienst als Hausmeister, Kinder- und Jugendzentrum LUNA, Boizenburg (Elbe)

Auszeichnungen

2015	Steinberg-Krupp Alzheimer-Forschungspreis für Nachwuchswissenschaftler

Lehre und Betreuung von Abschlussarbeiten

Seit 2020	Ringvorlesung „Klinische Anwendungen – Methoden der Funktionserhebung“ im Studiengang B.Sc. Medizinische Informationstechnologie (jährlich)
Seit 2017	Methodenseminar „Analyse neuronaler Bildgebungsdaten“ im Studiengang B.Sc./M.Sc. Medizinische Biotechnologie (halbjährlich)
Seit 2012	Betreuung von 12 Bachelor- und Masterarbeiten in den Studiengängen Informatik und Elektrotechnik mit dem Schwerpunkt maschinelles Lernen für Gesundheitsdaten; Zweitbetreuung von 8 medizinischen Doktorarbeiten, Schwerpunkt Verarbeitung neuronaler Bildgebungsdaten und statistische Analysemethoden

Fachvorträge und Öffentlichkeitsarbeit

2023	Vortrag bei der Helmholtz AI Konferenz, Hamburg: „Healthcare 3.0: How to Transform Machine Learning Prototypes into Functional Healthcare Applications for Diagnostic Assistance?“
2022	Vortrag beim Forschungsworkshop der Unimedizin Rostock: „Sich erklärende KI für diagnostische Expertensysteme“
2022	Schüler Chat-Sessions im Rahmen der Initiative „I’m a Scientist“ – Thema KI
2021 & 2022	Vorträge beim Schülerförderprogramm „Talente im Land – Bayern“: „Maschinelles Lernen in der Demenzforschung“
2021	Symposium bei der Jahrestagung der Deutschen Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde (DGPPN), Berlin: „Doktor-KI: sich selbst erklärende KI zur Detektion präklinischer Krankheitszeichen neuropsychiatrischer Erkrankungen“
2021	Interview (Livestream) mit KI-MV im Digitalen Innovationszentrum Rostock: „KI in der bildgebenden Diagnostik“
Seit 2021	Projektwebsite https://explaination.net mit interaktiver Demo-App und Blog
2021	Vortrag beim BR50 Workshop on AI, Berlin: „Self-Explaining AI for Diagnostic Expert Systems“
2021	Vortrag beim KI-Workshop MV, Rostock: „Jenseits der Blackbox – sich erklärende KI für diagnostische Expertensysteme“
2021	Vortrag beim European Congress of Radiology (ECR), Wien: „Towards a comprehensible convolutional neural network for the detection of Alzheimer’s disease in magnetic resonance images“ (zusammen mit Dr. Daniel Cantré)
2020	Featured Report Session (Symposium) bei der Alzheimer’s Association International Conference (AAIC), online: „Doctor AI: Making computers explain their decisions“
2016	Symposium bei der Jahrestagung der Deutschen Gesellschaft für Biologische Psychiatrie (DGBP), Würzburg: „Funktionelle MRT als Biomarker bei psychiatrischen und neurodegenerativen Erkrankungen – Ergebnisse, Relevanz und Perspektiven multizentrischer Studien“
2011–2019	Lange Nacht der Wissenschaften, Rostock: „MRT-Bildgebung als Biomarker in der Demenzforschung“

Vollständige Liste der Publikationen

1. Singh D, **Dyrba M** (2023). *Comparison of CNN Architectures for Detecting Alzheimer's Disease using Relevance Maps*. In: Bildverarbeitung für die Medizin 2023. Springer Fachmedien Wiesbaden: 238–243. DOI: 10.1007/978-3-658-41657-7_51 (ohne IF)
2. Gaubert M, Dell'Orco A, Lange C, Garnier-Crussard A, Zimmermann I, **Dyrba M**, et al. (2023). *Performance evaluation of automated white matter hyperintensity segmentation algorithms in a multicenter cohort on cognitive impairment and dementia*. Frontiers in Psychiatry 13: 1010273. DOI: 10.3389/fpsyt.2022.1010273 (IF 5.435)
3. Zhao K, Lin J, **Dyrba M** et al. (2023). *Coupling of the spatial distributions between sMRI and PET reveals the progression of Alzheimer's disease*. Network Neuroscience 7:86–101. DOI: 10.1162/netn_a_00271 (IF 4.98)
4. Nemy M, **Dyrba M**, Brosseron F, et al. (2022). *Cholinergic white matter pathways along the Alzheimer's disease continuum*. Brain 146: 2075–2088. DOI: 10.1093/brain/awac385 (IF 15.255)
5. Hermann A, Tarakdjian GN, Temp AG, Kasper E, et al., **Dyrba M** (2022). *Cognitive and behavioural but not motor impairment increases brain age in amyotrophic lateral sclerosis*. Brain Communications 4: fcac239. DOI: 10.1093/braincomms/fcac239 (IF 4.4)
6. Teipel SJ, **Dyrba M**, Ballarini T, et al. (2022). *Association of Cholinergic Basal Forebrain Volume and Functional Connectivity with Markers of Inflammatory Response in the Alzheimer's Disease Spectrum*. Journal of Alzheimer's Disease 85: 1267–82. DOI: 10.3233/JAD-215196 (IF 4.16)
7. Sakr F, **Dyrba M**, Bräuer A, Teipel S (2022). *Association of Lipidomics Signatures in Blood with Clinical Progression in Preclinical and Prodromal Alzheimer's Disease*. Journal of Alzheimer's Disease 85: 1115–1127. DOI: 10.3233/JAD-201504 (IF 4.16)
8. **Dyrba M,** Hanzig M, Altenstein S, et al. (2021). *Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: Evaluation in Alzheimer's disease*. Alzheimer's Research and Therapy 13. DOI: 10.1186/s13195-021-00924-2 (IF 6.982)
9. Zhao K, Zheng Q, **Dyrba M**, et al. (2022). *Regional Radiomics Similarity Networks Reveal Distinct Subtypes and Abnormality Patterns in Mild Cognitive Impairment*. Advanced Science (Weinh): 2104538. DOI: 10.1002/advs.202104538 (IF 17.521)
10. Teipel SJ, **Dyrba M**, Vergallo A, et al. (2021). *Partial Volume Correction Increases the Sensitivity of 18F-Florbetapir-Positron Emission Tomography for the Detection of Early Stage Amyloidosis*. Frontiers in Aging Neuroscience 13. doi: 10.3389/fnagi.2021.748198 (IF 5.75)
11. **Dyrba M**, Hanzig M (2021). *Interactive Visualization of 3D CNN Relevance Maps to Aid Model Comprehensibility – Application to the Detection of Alzheimer's Disease in MRI Images*. In: Bildverarbeitung für die Medizin 2021. Springer Fachmedien Wiesbaden: 317–322. DOI: 10.1007/978-3-658-33198-6_77 (ohne IF)
12. Zhao K, Zheng Q, **Dyrba M**, et al. (2021). *Regional radiomics similarity networks (R2SNs) in the human brain: Reproducibility, small-world properties and a biological basis*. Network Neuroscience 5: 783–797. DOI: 10.1162/netn_a_00200 (IF 4.625)
13. Hedderich DM, Menegaux A, **Dyrba M**, et al. (2021). *Aberrant Claustrum Microstructure in Humans after Premature Birth*. Cerebral Cortex 31: 5549–5559. DOI: 10.1093/cercor/bhab178 (IF 5.357)
14. Temp AG, **Dyrba M**, Büttner C, et al. (2021). *Cognitive Profiles of Amyotrophic Lateral Sclerosis Differ in Resting-State Functional Connectivity: An fMRI Study*. Frontiers in Neuroscience 15: 682100. DOI: 10.3389/fnins.2021.682100 (IF 4.677)
15. Temp AG, **Dyrba M**, Kasper E, Teipel S, Prudlo J (2021). *Case Report: Cognitive Conversion in a Non-brazilian VAPB Mutation Carrier (ALS8)*. Frontiers in Neurology 12: 668772. DOI: 10.3389/fneur.2021.668772 (IF 4.003)

16. Faraza S, Waldenmaier J, **Dyrba M**, et al. (2021). *Dorsolateral Prefrontal Functional Connectivity Predicts Working Memory Training Gains*. Frontiers in Aging Neuroscience 13. DOI: 10.3389/fnagi.2021.592261 (IF 5.75)
17. Levin F, Ferreira D, **Dyrba M**, et al. (2021). *Data-driven FDG-PET subtypes of Alzheimer's disease-related neurodegeneration*. Alzheimers Research & Therapy 13: 49. DOI: 10.1186/s13195-021-00785-9 (IF 6.982)
18. Teipel SJ, Temp AG, Levin F, **Dyrba M**, Grothe MJ (2021). *Association of TDP-43 Pathology with Global and Regional 18F-Florbetapir PET Signal in the Alzheimer's Disease Spectrum*. Journal of Alzheimer's Disease 79: 663–670. DOI: 10.3233/JAD-201032 (IF 4.472)
19. Amaefule CO, **Dyrba M**, Wolfsgruber S, et al. (2021). *Association between composite scores of domain-specific cognitive functions and regional patterns of atrophy and functional connectivity in the Alzheimer's disease spectrum*. Neuroimage: Clinical 29: 102533. DOI: 10.1016/j.nicl.2020.102533 (IF 4.881)
20. van Popering L, Tahmassebi A, **Dyrba M**, et al. (2021). *Identifying the diffusion source of dementia spreading in structural brain networks*. In: Gimi BS, Krol A, editors. Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging. SPIE: 7. DOI: 10.1117/12.2582200 (ohne IF)
21. **Dyrba M**, Pallath AH, Marzban EN (2020). *Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer's disease*. In: Bildverarbeitung für die Medizin 2020. Springer Fachmedien Wiesbaden: 307–312. DOI: 10.1007/978-3-658-29267-6_68 (ohne IF)
22. Teipel SJ, Temp AG, Levin F, **Dyrba M**, Grothe MJ (2020). *Association of PET-based stages of amyloid deposition with neuropathological markers of Aβ pathology*. Annals of Clinical and Translational Neurology 8: 29–42. DOI: 10.1002/acn3.51238 (IF 3.66)
23. Ramírez J, Górriz JM, Ortiz A, Cole JH, **Dyrba M** (2020). *Editorial: Deep Learning in Aging Neuroscience*. Frontiers in Neuroinformatics 14: 1–3. DOI: 10.3389/fninf.2020.573974 (IF 2.649)
24. Teipel SJ, **Dyrba M**, Chiesa PA, et al. (2020). *In vivo staging of regional amyloid deposition predicts functional conversion in the preclinical and prodromal phases of Alzheimer's disease*. Neurobiology of Aging 93: 98–108. DOI: 10.1016/j.neurobiolaging.2020.03.011 (IF 4.347)
25. Görß D, Kilimann I, **Dyrba M**, et al. (2020). *LATE: Nicht jede Demenz ist Alzheimer – Diskussion einer neuen Krankheitsentität am Fallbeispiel Zum aktuellen Stand der „limbic-predominant age-related TDP-43 encephalopathy" (LATE)*. Nervenarzt 92: 18–26. DOI: 10.1007/s00115-020-00922-z (IF 0.824)
26. **Dyrba M**, Mohammadi R, Grothe MJ, Kirste T, Teipel SJ (2020). *Gaussian graphical models reveal inter-modal and inter-regional conditional dependencies of brain alterations in Alzheimer's disease*. Frontiers in Aging Neuroscience. DOI: 10.3389/fnagi.2020.00099 (IF 4.362)
27. Herdick M, **Dyrba M**, Fritz H-CJ, et al. (2020). *Multimodal MRI analysis of basal forebrain structure and function across the Alzheimer's disease spectrum*. Neuroimage: Clinical 28: 102495. DOI: 10.1016/j.nicl.2020.102495 (IF 4.35)
28. Brueggen K, **Dyrba M**, Cardenas-Blanco A, et al. (2019). *Structural integrity in subjective cognitive decline, mild cognitive impairment and Alzheimer's disease based on multicenter diffusion tensor imaging*. Journal of Neurology 266: 2465–2474. DOI: 10.1007/s00415-019-09429-3 (IF 4.204)
29. Fritz H-CJ, Ray N, **Dyrba M**, et al. (2019). *The corticotopic organization of the human basal forebrain as revealed by regionally selective functional connectivity profiles*. Human Brain Mapping 40: 868–878. DOI: 10.1002/hbm.24417 (IF 4.554)
30. Sakr FA, Grothe MJ, **Dyrba M**, et al. (2019). *Applicability of in vivo staging of regional amyloid burden in a cognitively normal cohort with subjective memory complaints: the INSIGHT-preAD study*. Alzheimer's Research and Therapy 11: 15. DOI: 10.1186/s13195-019-0466-3 (IF 6.142)

31. Teipel SJ, Kuper-Smith JO, et al., **Dyrba M** (2019). *Multicenter Tract-Based Analysis of Microstructural Lesions within the Alzheimer's Disease Spectrum: Association with Amyloid Pathology and Diagnostic Usefulness*. Journal of Alzheimer's Disease 72: 455–465. DOI: 10.3233/JAD-190446 (IF 3.517)
32. Teipel S, Bakardjian H, **Dyrba M**, et al. (2018). *No association of cortical amyloid load and EEG connectivity in older people with subjective memory complaints*. Neuroimage: Clinical 17: 435–443. DOI: 10.1016/j.nicl.2017.10.031 (IF 3.869)
33. Henf J, Grothe MJ, Brueggen K, Teipel S, **Dyrba M** (2018). *Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction*. Neuroimage: Clinical 17: 579–586. DOI: 10.1016/j.nicl.2017.10.005 (IF 3.869)
34. Teipel SJ, Metzger CD, et al., **Dyrba M** (2018). *Multicenter Resting State Functional Connectivity in Prodromal and Dementia Stages of Alzheimer's Disease*. Journal of Alzheimer's Disease 64: 801–813. DOI: 10.3233/JAD-180106 (IF 3.476)
35. **Dyrba M**, Grothe M, Mohammadi R, Binder H, Kirste T, Teipel SJ (2018): *Comparison of different hypotheses regarding the spread of Alzheimer's disease using Markov random fields and multimodal imaging*. Journal of Alzheimer's Disease, 65 (3): 731–746. DOI: 10.3233/JAD-161197 (IF 3.731)
36. Mohammadi R, **Dyrba M** (2018): *Statistical modelling of brain connectivity in prodromal Alzheimer's disease*. Proceedings of the International Workshop on Statistical Modelling 2018, Bristol, UK: 119–122. (ohne IF)
37. Brueggen K, **Dyrba M**, Kilimann I, et al. (2018). *Hippocampal Mean Diffusivity for the Diagnosis of Dementia and Mild Cognitive Impairment in Primary Care*. Current Alzheimer Research 15: 1005–1012. DOI: 10.2174/1567205015666180613114829 (IF 3.289)
38. Franzmeier N, **Dyrba M** (2017). *Functional brain network architecture may route progression of Alzheimer's disease pathology* (Editorial Comment). Brain 140: 3077–3080. DOI: 10.1093/brain/awx304 (IF 10.292)
39. Grothe MJ, Barthel H, Sepulcre J, **Dyrba M**, Sabri O, Teipel SJ (2017). *In vivo staging of regional amyloid deposition*. Neurology 89: 2031–2038. DOI: 10.1212/WNL.0000000000004643 (IF 7.592)
40. Grothe MJ, Villeneuve S, **Dyrba M**, Bartres-Faz D, Wirth M (2017). *Multimodal characterization of older APOE2 carriers reveals selective reduction of amyloid load*. Neurology 88: 569–576. DOI: 10.1212/WNL.0000000000003585 (IF 7.592)
41. Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Klöppel S, Borchardt V, Walter M, **Dyrba M** (2017): *Robust Detection of Impaired Resting State Functional Connectivity Networks in Alzheimer's Disease Using Elastic Net Regularized Regression*. Frontiers in Aging Neuroscience, 8: 1564. DOI: 10.3389/fnagi.2016.00318 (IF 4.504)
42. Brueggen K, Grothe MJ, **Dyrba M**, et al. (2017): *The European DTI Study on Dementia – A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment*. NeuroImage 144: 305–308. DOI: 10.1016/j.neuroimage.2016.03.067 (IF 5.835)
43. Ochmann S, **Dyrba M**, Grothe MJ, et al. (2017). *Does Functional Connectivity Provide a Marker for Cognitive Rehabilitation Effects in Alzheimer's Disease? An Interventional Study*. Journal of Alzheimer's Disease 57: 1303–1313. DOI: 10.3233/JAD-160773 (IF 3.731)
44. Teipel SJ, Wohlert A, et al., **Dyrba M** (2017). *Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI*. Neuroimage: Clinical 14: 183–194. DOI: 10.1016/j.nicl.2017.01.018 (IF 4.348)
45. Brueggen K, Kasper E, **Dyrba M**, Bruno D, Pomara N, Ewers M, et al. (2016): *The Primacy Effect in Amnestic Mild Cognitive Impairment: Associations with Hippocampal Functional Connectivity: Associations with Hippocampal Functional Connectivity*. Frontiers in Aging Neuroscience, 8: 244. DOI: 10.3389/fnagi.2016.00244 (IF 4.348)

46. Kljajevic V, **Dyrba M**, Kasper E, Teipel S (2016): *Is the left uncinate fasciculus associated with verbal fluency decline in mild Alzheimer's disease?* Translational Neuroscience, 7: 89–91. DOI: 10.1515/tnsci-2016-0014 (IF 1.012)
47. Teipel S, Grothe MJ, Zhou J, Sepulcre J, **Dyrba M**, Sorg C, Babiloni C (2016): *Measuring Cortical Connectivity in Alzheimer's Disease as a Brain Neural Network Pathology: Toward Clinical Applications*. Journal of the International Neuropsychological Society, 22 (2): 138–163. DOI: 10.1017/S1355617715000995 (IF 2.633)
48. **Dyrba M**, Grothe M, Kirste T, Teipel SJ (2015): *Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM*. Human Brain Mapping, 36 (6): 2118–2131. DOI: 10.1002/hbm.22759 (IF 5.969)
49. **Dyrba M**, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirste T, Teipel SJ (2015): *Predicting Prodromal Alzheimer's Disease in Subjects with Mild Cognitive Impairment Using Machine Learning Classification of Multimodal Multicenter Diffusion-Tensor and Magnetic Resonance Imaging Data*. Journal of Neuroimaging, 25 (5): 738–747. DOI: 10.1111/jon.12214 (IF 1.734)
50. Brueggen K, **Dyrba M**, Barkhof F, et al. (2015): *Basal Forebrain and Hippocampus as Predictors of Conversion to Alzheimer's Disease in Patients with Mild Cognitive Impairment – A Multicenter DTI and Volumetry Study*. Journal of Alzheimer's Disease, 48: 197–204. DOI: 10.3233/JAD-150063 (IF 4.151)
51. Kljajevic V, Meyer P, **Dyrba M**, et al. (2014): *The ε4 genotype of apolipoprotein E and white matter integrity in Alzheimer's disease*. Alzheimer's & Dementia, 10: 401–404. DOI: 10.1016/j.jalz.2013.02.008 (IF 14.483)
52. Schuster C, Kasper E, **Dyrba M**, et al. (2014): *Cortical thinning and its relation to cognition in amyotrophic lateral sclerosis*. Neurobiology of Aging, 35: 240–246. DOI: 10.1016/j.neurobiolaging.2013.07.020 (IF 6.166)
53. Teipel SJ, Grothe MJ, **Dyrba M**, et al. (2014): *Fractional Anisotropy Changes in Alzheimer's Disease Depend on the Underlying Fiber Tract Architecture: A Multiparametric DTI Study using Joint Independent Component Analysis*. Journal of Alzheimer's Disease, 41: 69–83. DOI: 10.3233/JAD-131829 (IF 4.174)
54. **Dyrba M**, Ewers M, Wegrzyn M, et al. (2013): *Robust Automated Detection of Microstructural White Matter Degeneration in Alzheimer's Disease Using Machine Learning Classification of Multicenter DTI Data*. PLoS ONE 8: e64925. DOI: 10.1371/journal.pone.0064925 (IF 3.73)
55. **Dyrba M**, Ewers M, Wegrzyn M, et al. (2012): *Combining DTI and MRI for the Automated Detection of Alzheimer's Disease Using a Large European Multicenter Dataset*. In: Multimodal Brain Image Analysis 2012: 18–28. DOI: 10.1007/978-3-642-33530-3_2 (ohne IF)
56. **Dyrba M**, Nicolay R, Bader S, Kirste T (2011). *Evaluation of two Control Systems for Smart Environments*. In: Proceedings of the 8th International ICST Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services – 3rd Workshop on Context Systems Design, Evaluation and Optimisation. Copenhagen, Denmark, p. 1–12. (ohne IF)
57. **Dyrba M**, Bader S (2011): *Goalaviour-Based Control of Heterogeneous and Distributed Smart Environments*. In: Intelligent Environments 2011. Nottingham, UK. p. 142–148. DOI: 10.1109/IE.2011.33 (ohne IF)

Anhang

Comparison of CNN Visualization Methods to Aid Model Interpretability for Detecting Alzheimer's Disease

Martin Dyrba[1], Arjun H. Pallath[2], Eman N. Marzban[1,3,4]

[1] German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
[2] Institute of Visual & Analytic Computing, University of Rostock, Germany
[3] Clinic for Psychosomatic and Psychotherapeutic Medicine (KPM), University Medical Center Rostock, Germany
[4] Biomedical Engineering and Systems Dept., Faculty of Engineering, Cairo University, Giza, Egypt
martin.dyrba@dzne.de

Abstract. Advances in medical imaging and convolutional neural networks (CNNs) have made it possible to achieve excellent diagnostic accuracy from CNNs comparable to human raters. However, CNNs are still not implemented in medical trials as they appear as a black box system and their inner workings cannot be properly explained. Therefore, it is essential to assess CNN relevance maps, which highlight regions that primarily contribute to the prediction. This study focuses on the comparison of algorithms for generating heatmaps to visually explain the learned patterns of Alzheimer's disease (AD) classification. T1-weighted volumetric MRI data were entered into a 3D CNN. Heatmaps were then generated for different visualization methods using the iNNvestigate and keras-vis libraries. The model reached an area under the curve of 0.93 and 0.75 for separating AD dementia patients from controls and patients with amnestic mild cognitive impairment from controls, respectively. Visualizations for the methods deep Taylor decomposition and layer-wise relevance propagation (LRP) showed most reasonable results for individual patients matching expected brain regions. Other methods, such as Grad-CAM and guided backpropagation showed more scattered activations or random areas. For clinically research, deep Taylor decomposition and LRP showed most valuable network activation patterns.

1 Introduction

Deep convolution neural networks (CNNs) have become the state-of-the-art technique for various image classification tasks. The performance of these systems has been reported to be on par with humans. Several papers have proposed various new architectures for general-purpose image detection, which have a steady trend of improvement of model accuracy. These networks are actively being researched and developed in areas related to computer vision in many

T. Tolxdorff et al. (Hrsg.), *Bildverarbeitung für die Medizin 2020*,
Informatik aktuell, https://doi.org/10.1007/978-3-658-29267-6_68

fields such as self-driving cars, face recognition, object detection, and medical imaging. These systems when applied in medical imaging, could aid physicians in the early diagnosis of diseases and highlight the concerning areas in medical scans. However, there is a lack of transparency in the accuracy of results derived from these networks, as there is no direct way to identify on what basis the network performs the classification. Recently, several methods have been proposed to calculate CNN relevance maps ([1] for an extensive overview). These maps highlight regions of the input images that the network focuses on when classifying the disease. The regions identified by the network and the regions known to be affected by the disease can then be compared to check whether they match. Such plausibility checks could lead to more robust, reliable, and trustworthy CNN models; and, therefore, would also improve the clinical utility of such models.

In the literature, we only found three papers [2, 3, 4] providing CNN visualizations for 3D MRI data and disease prediction. However, a direct comparison with the most recent approaches such as layer-wise relevance propagation (LRP) and deep Taylor decomposition is still lacking.

Alzheimer's disease (AD) is the major cause of dementia in elderly people above 65 years of age. AD is characterized by the death of nerve cells (neurons) causing irreversible changes and atrophy (volume reduction) in the brain, leading to memory loss, behavioral changes, speech impairment, and difficulties in activities of daily living. AD is difficult to diagnose in its early stages due to the slow progress of the disease and due to the difficulty of discriminating accelerated atrophy in AD from normal age-related atrophy. People suffering from mild cognitive impairment (MCI), especially when involving memory, are being seen at high risk for progressing to AD dementia.

This study aims to i) detect AD or MCI using a 3D CNN for T1-weighted volumetric MRI data, and ii) compare different visualization methods with respect to the clinical utility of the derived heatmaps for indicating areas that most contribute to the classification of the scans.

2 Materials and methods

Study sample T1-weighted volumetric MRI data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1]. In total, the sample included 662 cases consisting of 198 patients with AD dementia, 219 patients with amnestic MCI, and 254 cognitively normal controls. The subjects' demographics are shown in Tab. 1. Using common SPM8 and VBM8 software, MRI scans were segmented into grey and white matter, spatially normalized to an in-house ageing/AD-specific brain template [5] using the DARTEL algorithm, and finally modulated. Additionally, all scans were cleaned for the effects of the covariates age, gender, total intracranial volume and scanner magnetic field strength using linear regression [6]. For each voxel, models were fitted for the

[1] More information about the ADNI can be found on http://adni.loni.usc.edu

Table 1. Sample characteristics.

	Controls	MCI	AD	p-value
Sample size (female)	254 (130)	219 (93)	189 (80)	0.149
Age (SD)	75.4 ± 6.6	74.1 ± 8.1	75.0 ± 8.0	<0.001
Education (SD)	16.4 ± 2.7	16.2 ± 2.8	15.9 ± 2.7	0.227
MMSE (SD)	29.1 ± 1.2	27.6 ± 1.9	22.6 ± 3.2	<0.001
Delayed recall (SD)	7.6 ± 4.1	3.2 ± 3.7	0.8 ± 1.9	<0.001

healthy control subjects. Subsequently, these models were applied to all scans, i.e. the residualized images were taken as input for the CNN. Due to memory limitations, we defined a field-of-view including the whole brain in axial and saggital directions, but only a range of 32 coronal slices covering the temporal lobe and the hippocampus area known to be most affected by AD. The field-of-view is illustrated on the left of Fig. 1.

Validation strategy We used a ten-fold cross-validation approach, such that the sample was divided into ten test sets (10%, n=67) for determining the accuracy of the model, and nested splits into training set (80%, n=535) and validation set (10%, n=60) for model training. The test sets included approximately 17 AD dementia patients, 24 MCI patients and 26 controls. Prior to training, the training sets were augmented by adding copies of the training scans shifted by ±2 voxels in x/y/z-direction resulting in training samples of n=3745 images.

CNN model layout and parameterization The CNN model was implemented in Keras/Tensorflow 1.15. The general 3D CNN model layout is shown in Fig. 1. Prior to training, class labels were merged for AD dementia and MCI to have a binary classification task. We specified the categorical cross-entropy as the loss function and the accuracy as the performance metric. The models were optimized by Adam running for 100 epochs with a batch size of 64 and default learning rate of 0.001. Training took approximately 55 minutes per cross-validation iteration.

Visualization methods We used the iNNvestigate library [1] implementing various visualization methods. In detail, we tested deconvnet, guided backpropagation, deep Taylor decomposition, input*gradient, and layer-wise relevance propagation (LRP) with the Z, epsilon, and alpha=1,beta=0 rules. Additionally, we used keras-vis [7] for the Grad-CAM approach. As the intensity range of the relevance maps differed greatly between approaches, it was scaled linearly to a fixed range allowing a visual comparison. In addition to the raw relevance maps overlayed on the original input data (with 50% transparency), we provide a smoothed and thresholded version containing the most prominent clusters only, i.e. the top 30 percentile of intensity values.

3 Results

For the test data, we obtained a mean accuracy of 75.2% with an area under the curve (AUC) of 0.83 for the combined dataset. When looking at the origi-

nal diagnosis subsets, the CNN achieved an AUC of 0.93 for AD dementia vs. controls and 0.75 for separating MCI patients from controls.

Exemplary relevance maps for different individuals are presented in Fig. 2 for an AD dementia case, a patient with MCI, and for a healthy control.

4 Discussion

The CNN model achieved excellent diagnostic accuracy for separating AD dementia from controls, comparable to other approaches from the literature [3, 6]. For separating MCI patients from controls, accuracy was reasonable and in line with previous studies. Notably, as computational complexity is considerably higher for 3D CNN models compared to 2D CNN models used for general purpose image detection tasks, there is a high potential of model overfitting, in contrast to a very limited number of MRI scans available for training. We addressed this problem by in three ways. Firstly, we applied a sophisticated image preprocessing pipeline including segmentation, spatial normalization, and covariate cleaning as common for voxel-based statistical analyzes. Secondly, we reduced the number of layers compared to other approaches [2, 3] resulting in a more shallow network. Our CNN model included three convolutional layers with in total approximately 6,400 parameters on the cost of being less rotation/translation-invariant compared to deeper CNNs. Thirdly, we used data augmentation to multiply the data available for training and to improve the stability and robustness of the model.

The CNN relevance maps obtained from the various approaches showed diverging quality with respect to focus, smoothness and scatter (Fig. 2). This result is in line with two previous papers testing a subset of methods [2, 3]. Approximately the same image regions were highlighted across the visualization methods. As expected, the hippocampus area showed the highest relevance for the AD and MCI patients. However, the directionality of weighting (positive vs. negative) differed between the algorithms. Notably, for Grad-CAM the relevance maps substantially differed with respect to the smoothness. This is due to the approach of calculating low-resolution activations at the fully-connected layer followed by upscaling (interpolation) to the original input image resolution. The two methods deep Taylor decomposition and LRP with alpha=1,beta=0 rule showed the most promising relevance maps with strongest focus. Also, these approaches mainly showed positive relevance scores for the AD class and sup-

Fig. 1. Convolutional neural network model layout.

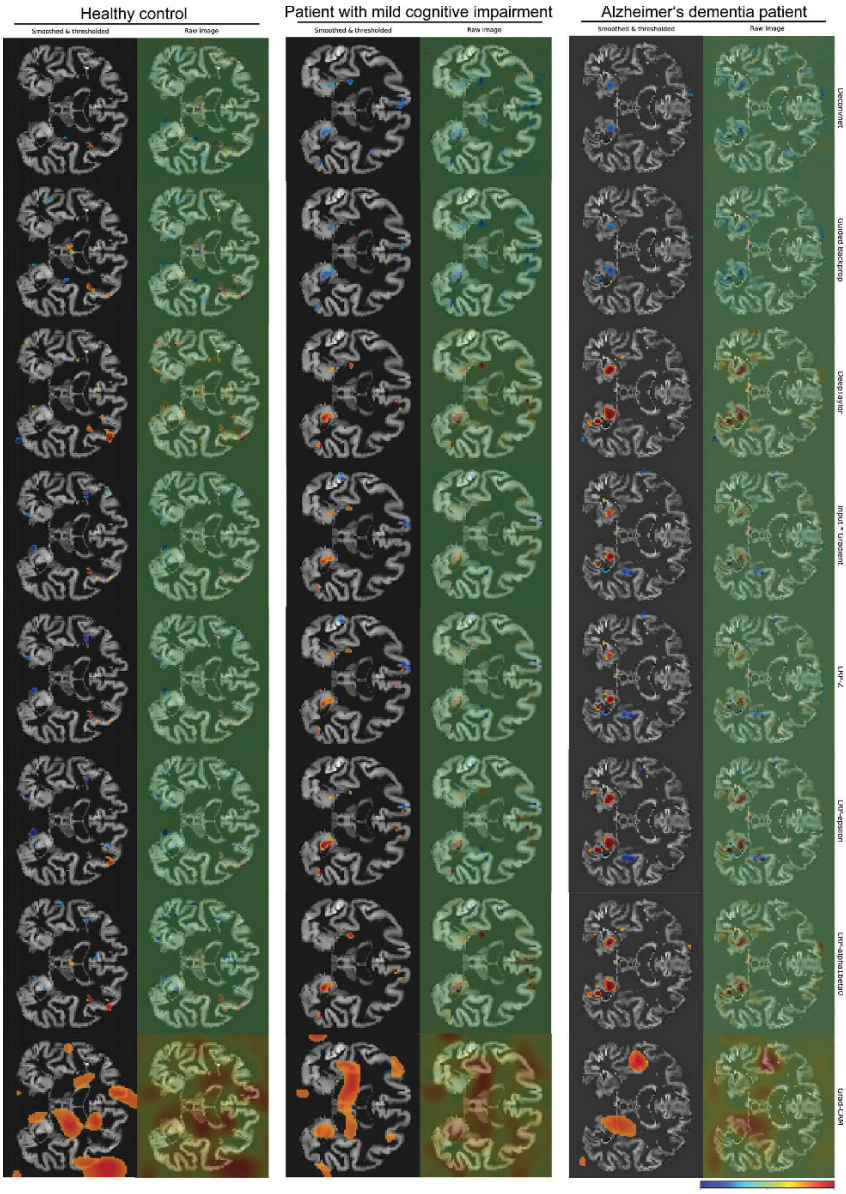


Fig. 2. Relevance maps for an individual with Alzheimer's dementia, amnestic mild cognitive impairment, and normal cognition. Raw figures in top row, smoothed and thresholded figures in bottom row. Methods by column: deconvnet, guided backpropagation, deep Taylor decomposition, input*gradient, LRP-Z, LRP-epsilon, LRP-alpha1beta0, grad-CAM. Red and blue color indicate high positive or negative activation.

pressed the negative relevance against AD. This might be valuable for a multi-class model, where negative relevance cannot be interpreted as clearly as in a binary classification task.

Two limitations have to be mentioned for the present work. Firstly, the CNN structure and parameters need to be systematically evaluated and performance has to be validated on an independent dataset, which we will do in the near future. Secondly, intensity normalization of the relevance maps is open research question. As the distribution of values differed between visualization methods and patients, we rescaled the range linearly and applied a percentile-based threshold. For clinical use, it should be considered to allow users to interactively adjust the color scale and threshold.

In conclusion, we presented a CNN structure providing both excellent diagnostic accuracy as well as relevance maps highlighting expected regions such as the hippocampus. For clinically oriented research, deep Taylor decomposition and LRP with alpha1,beta0 rule showed most valuable network activation patterns with high focus and less scatter.

Acknowledgement. We would like to thank the ADNI and contributors [2] for sharing their data.

References

1. Alber M, Lapuschkin S, Seegerer P, et al. iNNvestigate neural networks! J Mach Learn Res. 2019;20:1–8.
2. Rieke J, Eitel F, Weygandt M, et al. Visualizing vonvolutional networks for MRI-based diagnosis of Alzheimer's disease. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer; 2018. p. 24–31.
3. Böhle M, Eitel F, Weygandt M, et al. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
4. Zintgraf LM, Cohen TS, Adel T, et al. Visualizing deep neural network decisions: prediction difference analysis. In: International Conference on Learning Representations (ICLR); 2017. .
5. Grothe M, Heinsen H, Teipel S. Longitudinal measures of cholinergic forebrain atrophy in the transition from healthy aging to Alzheimer's disease. Neurobiol Aging. 2013;34(4):1210–1220.
6. Dyrba M, Barkhof F, Fellgiebel A, et al. Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter DTI and MRI data. J Neuroimaging. 2015;25(5):738–747.
7. Kotikalapudi R, contributors. keras-vis. GitHub; 2019.

[2] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Interactive Visualization of 3D CNN Relevance Maps to Aid Model Comprehensibility

Application to the Detection of Alzheimer's Disease in MRI Images

Martin Dyrba[1], Moritz Hanzig[1,2]

[1]German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
[2]Institute of Visual & Analytic Computing, University of Rostock, Germany
martin.dyrba@dzne.de

Abstract. Relevance maps derived from convolutional neural networks (CNN) indicate the influence of a particular image region on the decision of the CNN model. Individual maps are obtained for each single input 3D MRI image and various visualization options need to be adjusted to improve information content. In the use case of model prototyping and comparison, the common approach to save the 3D relevance maps to disk is impractical given the large number of combinations. Therefore, we developed a web application to aid interactive inspection of CNN relevance maps. For the requirements analysis, we interviewed several people from different stakeholder groups (model/visualization developers, radiology/neurology staff) following a participatory design approach. The visualization software was conceptually designed in a Model–View–Controller paradigm and implemented using the Python visualization library Bokeh. This framework allowed a Python server back-end directly executing the CNN model and related code, and a HTML/Javascript front-end running in any web browser. Slice-based 2D views were realized for each axis, accompanied by several visual guides to improve usability and quick navigation to image areas with high relevance. The interactive visualization tool greatly improved model inspection and comparison for developers. Owing to the well-structured implementation, it can be easily adapted to other CNN models and types of input data.

1 Introduction

Convolutional neural networks (CNN) achieved a high accuracy for the automated detection of disease patterns in MRI scans. Several relevance mapping algorithms have been proposed to generate heatmaps that indicate the influence of a particular image region on the decision of the CNN model [1, 2]. Two previous studies compared CNN relevance mapping algorithms with respect to brain regions driving the detection of Alzheimer's disease in structural T1-weighted MRI [3, 4]. These relevance maps were found to greatly improve CNN comprehensibility and identification of reasons why a model failed [1, 3]. Notably,

C. Palm et al. (Hrsg.), *Bildverarbeitung für die Medizin 2021*,
Informatik aktuell, https://doi.org/10.1007/978-3-658-33198-6_77

these approaches generate 3D relevance maps for each single input image (=MRI scan). In addition, several post-processing steps are required in order to improve their visual appeal and information content. These steps include smoothing, color scale transformation, relevance score and cluster size thresholding, which are not implemented in the feature portfolio of common MRI viewers. Further, preparation of static images from a specific parameter set yields large amounts of output files, which is prone to loosing track of the particular parameter settings used to generating these files in explorative research with various CNN model and post-processing parameter combinations.

In this paper, we present an interactive visualization toolkit for the online generation, parameterization, and inspection of CNN relevance maps for individual MRI scans. In the following sections, we describe the conceptual considerations and implementation, followed by a demonstration of the realized user interface and use case.

2 Materials and methods

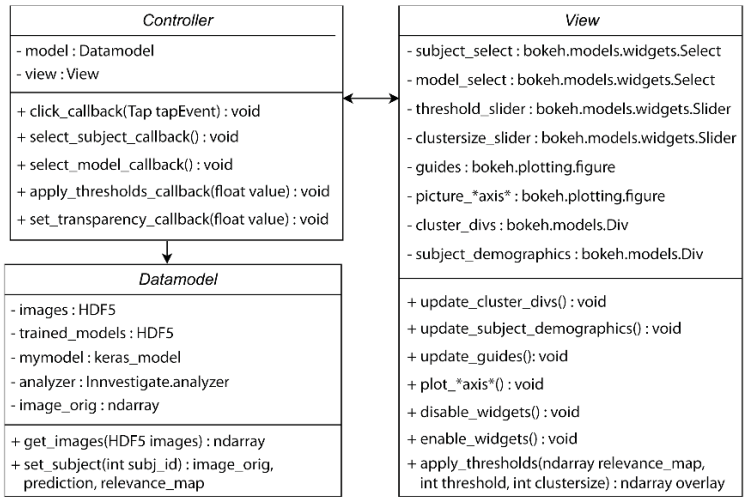
We used pretrained CNN models obtained from [4, 5], which were implemented in Keras 2.2.4 and Tensorflow 1.15. The CNN visualization library iNNvestigate 1.0.8 [2] was used to derive the relevance maps. After drafting a first prototype user interface for the visualization, we collected a list of key requirements from a range of stakeholders following a participatory design approach. Therefore, we interviewed two physicians trained in radiology/neurology, two experienced visualization developers, and two machine learning model developers. From their comments, we defined the list of requirements:

- Directly run in CNN modeling environment (Python)
- Optional: remote display for the case where data handling and model execution need to be run remotely
- Visualization as slice-based 2D plots, which clinical users are familiar with
- For regular users: interactive selection of MRI scans, adjustable relevance and cluster size thresholds
- For expert users: selection of alternative CNN models and relevance mapping algorithms

The Python visualization library Bokeh [6] met the requirements with respect to Python runtime environment and remote viewing instance in a web browser. It provides a Python server instance back-end and Javascript browser libraries front-end to remotely trigger Python function calls and return execution results to the web browser for displaying.

We divided the implementation into three components following the well-established Model–View–Controller design pattern. Fig. 1 provides an overview of implemented methods and Fig. 2 shows a sequence diagram of function calls being executed when selecting a new MRI scan. In addition to the key requirements, we implemented various visual guides in order to facilitate parameterization and quick navigation to brain regions with high relevance scores (Fig. 3).

Fig. 1. Class diagram illustrating core components and functions.



Among them are (a) a histogram providing the distribution of cluster sizes next to the cluster size threshold slider, (b) plots visualizing the amount of positive and negative relevance per slice next to the slice selection sliders, and (c) statistical information on the currently selected cluster. Further, assuming spatially normalized MRI data in MNI reference space, we added (d) atlas-based anatomical region lookup for the current cursor/cross-hair position and (e) the option to display the outline of the anatomical region to simplify visual comparison with the cluster location.

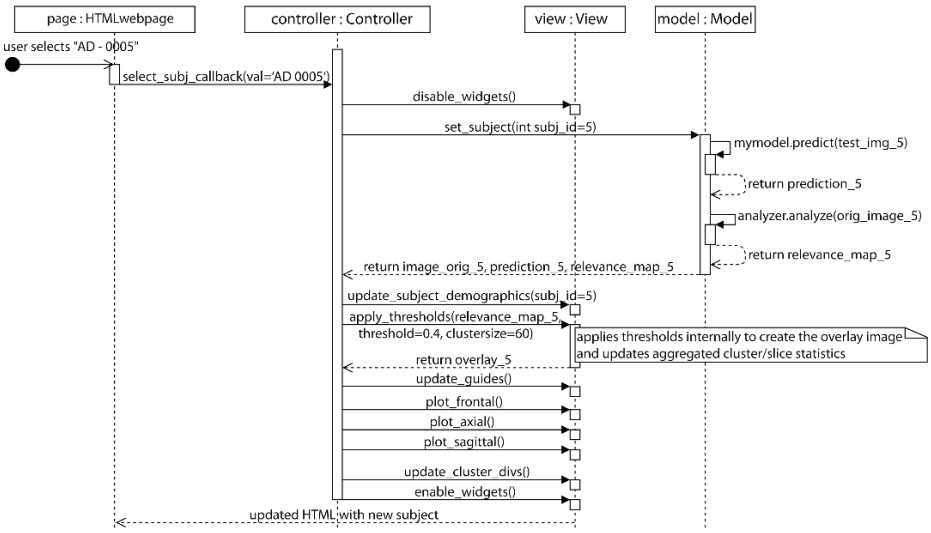


Fig. 2. Sequence diagram of functions being executed when loading a new MRI scan.

3 Results

The realized user interface is shown in Fig. 3. The source code is freely available on GitHub: https://github.com/martindyrba/DeepLearningInteractiveVis. The employed relevance mapping algorithm was initially fixed to layer-wise relevance propagation (LRP) as this method was already applied previously [3, 4]. The distribution and location of clusters with highest relevance scores varied between people with most consistent contributions from hippocampus, putamen and thalamus (Fig. 4). Notably, the highlighted regions mostly indicated actual gray matter atrophy as visible from the background images. This was confirmed by a quantitative comparison in which automatically derived hippocampus volume measures highly correlated with the aggregated relevance scores in the hippocampus region (Pearson's $r \approx -0.81$, see [5] for further details).

The used high-level programming interfaces of Keras, iNNvestigate, and Bokeh enabled a clean and structured programming of the web application. The complete application was realized in approximately 700 lines of code including view specification and program logic (model, controller). The app does not require a GPU to be available on the host, i.e. works smoothly on CPU. Loading of a new person currently takes ≈5 sec and loading a new model takes ≈15 sec. Adjusting the other sliders or directly clicking on the brain image/clusters updates the visualizations with a short latency of ≈200 ms. With the web page front-end, the visualization can also be run remotely on mobile devices such as tablets or smartphones without modification.

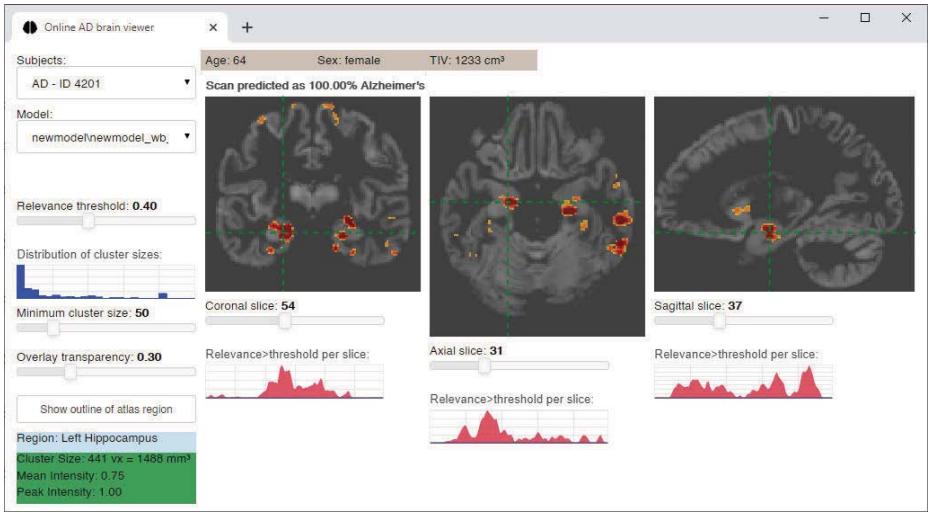


Fig. 3. Interactive web application user interface for 3D CNN relevance map visualization for a patient with dementia due to Alzheimer's disease.

4 Discussion

The presented visualization framework allows the inspection of CNN relevance maps for individuals and the assessment of drivers of the CNN decision. There-

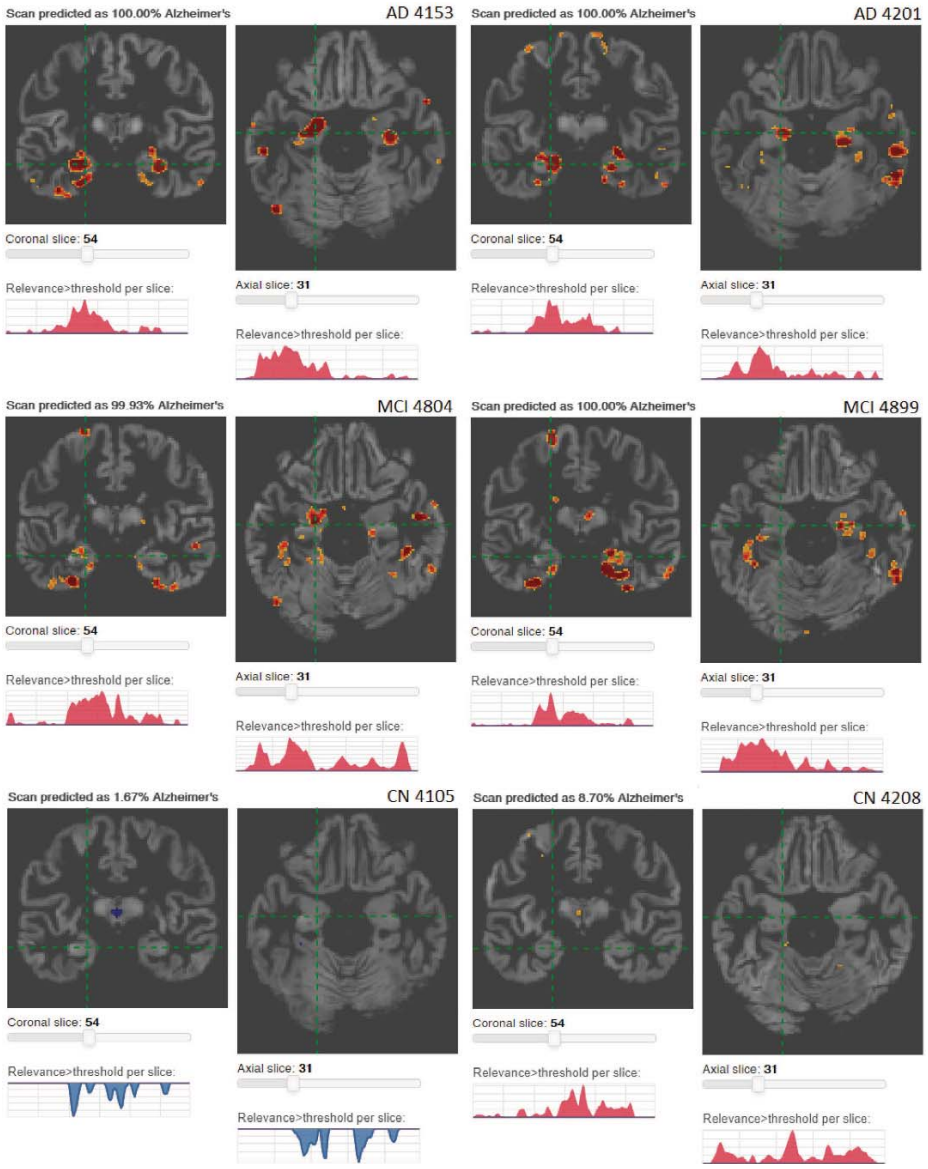


Fig. 4. Comparison of 3D CNN relevance maps for six randomly selected MRI scans. Top row: two patients with dementia due to Alzheimer's disease, middle row: two patients with mild cognitive impairment, bottom row: two controls with normal cognition.

fore, it contributed to model comprehensibility in the sense that it revealed the regional sensitivity of the CNN models, which is currently not assessed in the majority of papers with focus on model accuracy. As previously reported for other application domains [1], we found an association of model performance and relevant regions, which means that less accurate models mainly considered brain regions of low clinical relevance for AD. Thus, relevance maps provide a useful tool for CNN model 'debugging'.

Consulting experts from various disciplines for the requirements analysis greatly improved usability of the initial prototype application with many useful comments and recommendations such as adding the visual guides. The presented application was implemented as model- and data-agnostic tool such that it can be easily adjusted for other types of 3D input data.

The interactive web application met all the requirements defined initially (see section 2 above). A clear disadvantage is the small latency in reactivity causing a short delay of navigation actions. This is due to the data handling on the server and transfer of the relevance map slices to the client as byte stream, which could only be circumvented if both data model and viewer components run on the same system as native Python application.

For future work, there are ideas on additional features, for instance the import of new MRI scans, export for relevance maps, and 3D rendering view components. Further evaluation of the actual improvement of CNN comprehensibility for end users such as clinical staff is advised. Most importantly, more research is required to define evaluation metrics for relevance map quality and plausibility.

In summary, we presented a concept and implementation of an interactive online visualization application to inspect 3D CNN relevance maps and adjust display parameters as appropriate. Highlighting the individual's image regions with highest contribution on the particular decision of the CNN model in a simple and intuitive way makes this tool greatly enhancing model inspection and comparison for developers.

References

1. Samek W, Binder A, Montavon G, et al. Evaluating the visualization of what a deep neural network has learned. IEEE T Neur Net Lear. 2017;28(11):2660–2673.
2. Alber M, Lapuschkin S, Seegerer P, et al. iNNvestigate neural networks! J Mach Learn Res. 2019;20:1–8.
3. Böhle M, Eitel F, Weygandt M, et al. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
4. Dyrba M, Pallath AH, Marzban EN. Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer's disease. Proc BVM. 2020; p. 307–312.
5. Dyrba M, Hanzig M, Altenstein S, et al.. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease; 2020. https://arxiv.org/abs/2012.10294.
6. Bokeh contributors. The bokeh visualization library; 2020. https://bokeh.org.

Dyrba *et al. Alzheimer's Research & Therapy* (2021) 13:191
https://doi.org/10.1186/s13195-021-00924-2

Alzheimer's Research & Therapy

RESEARCH Open Access

Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease

Martin Dyrba[1*], Moritz Hanzig[1,2], Slawek Altenstein[3,4], Sebastian Bader[2], Tommaso Ballarini[5], Frederic Brosseron[5,6], Katharina Buerger[7,8], Daniel Cantré[9], Peter Dechent[10], Laura Dobisch[11], Emrah Düzel[11,12], Michael Ewers[7,8], Klaus Fliessbach[5,6], Wenzel Glanz[11], John-Dylan Haynes[13], Michael T. Heneka[5,6], Daniel Janowitz[8], Deniz B. Keles[14], Ingo Kilimann[1,15], Christoph Laske[16,17,18], Franziska Maier[19], Coraline D. Metzger[11,12,20], Matthias H. Munk[16,18,21], Robert Perneczky[7,22,23,24], Oliver Peters[3,14], Lukas Preis[3,14], Josef Priller[3,4,25], Boris Rauchmann[22], Nina Roy[5], Klaus Scheffler[26], Anja Schneider[5,6], Björn H. Schott[27,28,29], Annika Spottke[5,30], Eike J. Spruth[3,4], Marc-André Weber[9], Birgit Ertl-Wagner[31,32], Michael Wagner[5,6], Jens Wiltfang[27,28,33], Frank Jessen[5,19,34] and Stefan J. Teipel[1,15] for the ADNI, AIBL, DELCODE study groups

Abstract

Background: Although convolutional neural networks (CNNs) achieve high diagnostic accuracy for detecting Alzheimer's disease (AD) dementia based on magnetic resonance imaging (MRI) scans, they are not yet applied in clinical routine. One important reason for this is a lack of model comprehensibility. Recently developed visualization methods for deriving CNN relevance maps may help to fill this gap as they allow the visualization of key input image features that drive the decision of the model. We investigated whether models with higher accuracy also rely more on discriminative brain regions predefined by prior knowledge.

Methods: We trained a CNN for the detection of AD in $N = 663$ T1-weighted MRI scans of patients with dementia and amnestic mild cognitive impairment (MCI) and verified the accuracy of the models via cross-validation and in three independent samples including in total $N = 1655$ cases. We evaluated the association of relevance scores and hippocampus volume to validate the clinical utility of this approach. To improve model comprehensibility, we implemented an interactive visualization of 3D CNN relevance maps, thereby allowing intuitive model inspection.

Results: Across the three independent datasets, group separation showed high accuracy for AD dementia versus controls ($AUC \geq 0.91$) and moderate accuracy for amnestic MCI versus controls ($AUC \approx 0.74$). Relevance maps indicated that hippocampal atrophy was considered the most informative factor for AD detection, with additional contributions from atrophy in other cortical and subcortical regions. Relevance scores within the hippocampus were highly correlated with hippocampal volumes (Pearson's $r \approx -0.86$, $p < 0.001$).

*Correspondence: martin.dyrba@dzne.de
[1] German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany
Full list of author information is available at the end of the article

Conclusion: The relevance maps highlighted atrophy in regions that we had hypothesized a priori. This strengthens the comprehensibility of the CNN models, which were trained in a purely data-driven manner based on the scans and diagnosis labels. The high hippocampus relevance scores as well as the high performance achieved in independent samples support the validity of the CNN models in the detection of AD-related MRI abnormalities. The presented data-driven and hypothesis-free CNN modeling approach might provide a useful tool to automatically derive discriminative features for complex diagnostic tasks where clear clinical criteria are still missing, for instance for the differential diagnosis between various types of dementia.

Keywords: Alzheimer's disease, Deep learning, Convolutional neural network, MRI, Layer-wise relevance propagation

Introduction

Alzheimer's disease (AD) is characterized by widespread neuronal degeneration, which manifests macroscopically as cortical atrophy that can be detected in vivo using structural magnetic resonance imaging (MRI) scans. Particularly at earlier stages of AD, atrophy patterns are relatively regionally specific, with volume loss in the medial temporal lobe and particularly the hippocampus. Therefore, hippocampus volume is currently the best-established MRI marker for diagnosing Alzheimer's disease at the dementia stage as well as at its prodromal stage amnestic mild cognitive impairment (MCI) [1, 2]. Automated detection of subtle brain changes in early stages of Alzheimer's disease could improve diagnostic confidence and early access to intervention [1, 3].

Convolutional neural networks (CNNs) provide a powerful method for image recognition. Various studies have evaluated the performance of CNNs for the detection of Alzheimer's disease in MR images with promising results regarding both separation of diagnostic groups and the prediction of conversion from MCI to manifest dementia. Despite the high accuracy levels achieved by CNN models, a major drawback is their algorithmic complexity, which renders them black-box systems. The poor intuitive comprehensibility of CNNs is one of the major obstacles which hinder the clinical application.

Novel methods for deriving relevance maps from CNN models [4, 5] may help to overcome the black-box problem. In general, relevance or saliency maps indicate the amount of information or contribution of a single input feature on the probability of a particular output class. Previous methodological approaches like gradient-weighted class activation mapping (Grad-CAM) [6], occlusion sensitivity analyses [7, 8], and local interpretable model-agnostic explanations (LIME) [9] had the limitation that deriving the relevance or saliency maps provided only group-average estimates, required long runtime [10], or provided only low spatial resolution [11, 12]. In contrast, more recent methods such as guided backpropagation [13] or layer-wise relevance propagation (LRP) [4, 5] use back-tracing of neural activation through the network paths to obtain high-resolution relevance maps.

Recently, three studies compared LRP with other CNN visualization methods for the detection of Alzheimer's disease in T1-weighted MRI scans [11, 12, 14]. The derived relevance maps showed the strongest contribution of medial and lateral temporal lobe atrophy, which matched the a priori expected brain regions of high diagnostic relevance [15, 16]. These preliminary findings provided the first evidence that CNN models and LRP visualization could yield reasonable relevance maps for individual people. We investigated whether this approach could be used as the basis for neuroradiological assistance systems to support the examination and diagnostic evaluation of MRI scans. Furthermore, we wanted to develop a data-driven and hypothesis-free CNN modeling approach that is capable of automatically deriving discriminative features and, therefore, might support complex diagnostic tasks where clear clinical criteria are still missing such as the differential diagnosis of various types of dementia.

In the current study, our aims were threefold: First, we trained robust CNN models that achieved a high diagnostic accuracy in three independent validation samples. Second, we developed a visualization software to interactively derive and inspect diagnostic relevance maps from CNN models for individual patients. Here, we expected high relevance to be shown in brain regions with strong disease-related atrophy, primarily in the medial temporal lobe. Third, we evaluated the validity of relevance maps in terms of correlation of hippocampus relevance scores and hippocampus volume, which is the best-established MRI marker for Alzheimer's disease [15, 16]. We expected a high consistency of both measures, which would strengthen the overall comprehensibility of the CNN models.

State of the art

Neural network models to detect Alzheimer's disease

An overview of neuroimaging studies which applied neural networks in the context of AD is provided in

Table 1. We focused on the aspects whether the studies used independent validation samples to assess the generalizability of their models and whether they evaluated which image features contributed to the models' decision. Studies reported very high classification performances to differentiate AD dementia patients and cognitively healthy participants, typically with accuracies around 90% (Table 1). For the separation of MCI and controls, accuracies were substantially lower ranging between 75 and 85%. However, there is a high variation of the accuracy levels depending on various factors such as (i) differences in diagnostic criteria across samples, (ii) included data types, (iii) differences in image preprocessing procedures, and (iv) differences between machine learning methods [27].

CNN performance estimation and model robustness are still open challenges. Wen and colleagues [27] actually showed only a minor effect of the particular CNN model parameterization or network layer configuration on the final accuracy, which means that the fully trained CNN models achieved almost identical performance. Different CNN approaches exist for MRI data [27] based on (i) 2D convolutions for single slices, often reusing pre-trained models for general image detection, such as AlexNet [29] and VGG [30]; (ii) so-called 2.5D approaches running 2D convolutions on each of the three slice orientations, which are then combined at higher layers of the network; and (iii) 3D convolutions, which are at least theoretically superior in detecting texture and shape features in any direction of the 3D volume. Although final accuracy is almost comparable between all three approaches for detecting MCI and AD [27], the 3D models require substantially more parameters to be estimated during training. For instance, a single 2D convolutional kernel has $3 \times 3 = 9$ parameters whereas the 3D version requires $3 \times 3 \times 3 = 27$ parameters. Here, relevance maps and related methods enable the assessment of learnt CNN models with respect to overfitting to clinically irrelevant brain regions and the detection of potential biases present in the training samples, which cannot be directly identified just from the model accuracy.

Approaches to assess model comprehensibility

In the literature, the most often applied methods to assess model comprehensibility and sensitivity were (i) the visualization of model weights, (ii) occlusion sensitivity analysis, and (iii) more advanced CNN methods such as guided backpropagation or LRP (Table 1). Notably, studies using approaches i and ii showed visualizations characterizing the whole sample or group averages. In contrast, studies applying iii also presented relevance maps for single participants [11, 14].

Böhle and colleagues [14] pioneered the application of LRP in neuroimaging and reported a high sensitivity of this method to actual regional atrophy. Eitel and colleagues [12] assessed the stability and reproducibility of CNN performance results and LRP relevance maps. After training ten individual models based on the same training dataset, they reported the highest consistency and lowest deviation of relevance maps for LRP and guided backpropagation among five different methods [12]. Recently, we compared various methods for relevance and saliency attribution [11]. Visually, all tested methods provided similar relevance maps except for Grad-CAM, which provided much lower spatial resolution, and, hence, lost a high amount of regional specificity. For the other methods, the main difference was the amount of "negative" relevance which indicates evidence against a particular diagnostic class. Notably, [12, 14] did not include patients in the prodromal stage of MCI and [11] focused on a limited range of coronal slices covering the temporal lobe. All three studies did not validate their results in independent samples.

Materials and methods

Study samples

Data for *training* the CNN models were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations, with the primary goal of testing whether neuroimaging, neuropsychological, and other biological measurements can be used as reliable in vivo markers of Alzheimer's disease pathogenesis. A complete description of ADNI, up-to-date information, and a summary of diagnostic criteria are available at https://www.adni-info.org. We selected a sample of $N = 663$ participants from the ADNI-GO and ADNI-2 phases, based on the availability of concurrent T1-weighted MRI and amyloid AV45-PET scans. Notably, we used only one (i.e., the first) available scan from each ADNI participant in our analyses. The sample characteristics are shown in Table 2. We included 254 cognitively normal controls, 220 patients with (late) amnestic mild cognitive impairment (MCI), and 189 patients with Alzheimer's dementia (AD). Amyloid-beta status of the participants was determined by the UC Berkeley [31] based on the AV45-PET standardized uptake value ratio (SUVR) cutoff 1.11.

For *validation* of the diagnostic accuracy of the CNN models, we obtained MRI scans from three independent cohorts. The sample characteristics and demographic information are summarized in Table 2. The

Table 1 Overview of previous studies applying neural networks for the detection of AD and MCI

Study (chronologic order)	Data type	Sample			Algorithm	Performance				Addressed model comprehensibility
		AD	MCI c/nc	CN		Groups	Accuracy	Balanced accuracy	*AUC*	
Suk et al. [17]	MRI GM and FDG-PET	93	76/128	101	RBM on class discriminative patches selected by statistical significance tests	AD/CN MCI/CN MCIc/MCInc	95.4% 85.7% 74.6%	94.9% 80.6% 71.6%	0.988 0.881 0.747	Visualization of selected features (image patches) and RBM model weights projected on MRI scan
Li et al. [18]	MRI and FDG-PET	51	43/56	52	RBM for feature learning, SVM for classification	AD/CN MCI/CN MCIc/MCInc	91.4% 77.4% 57.4%			No
Ortiz et al. [19]	MRI GM and FDG-PET	70	39/64	68	RBM for feature learning, SVM for classification	AD/CN MCIc/CN MCIc/MCInc	90% 83% 78%		0.95 0.95 0.82	Visualization of SVM model weights projected on MRI scan
Aderghal et al. [20]	MRI and DTI	188	339	228	CNN for hippocampus region of interest only	AD/CN MCI/CN	92.5% 80.0%	92.5% 82.9%		No
Liu et al. [21]	FDG-PET	93	146	100	CNN and RNN	AD/CN MCI/CN	91.2% 78.9%		0.953 0.839	Visualization of most contributing brain areas obtained from occlusion sensitivity analysis
Liu et al. [22]	MRI	199	–	229	CNN on landmarks selected by statistical significance tests	AD/CN MCIc/CN	90.6%		0.957	Visualization of top 50 anatomical landmarks used as input for the CNN
Lin et al. [23]	MRI	188	169/193	229	CNN	AD/CN MCIc/MCInc	88.8% 79.9%		0.861	No
Böhle et al. [14]	MRI	211	–	169	CNN	AD/CN	88.0%			Visualization of LRP relevance and guided backpropagation maps, comparison of LRP relevance scores by group and brain region
Li et al. [24]	MRI	Training 192 Test 225	383 479	228 639	CNN for hippocampus only	AD/CN MCIc/MCInc	92.9%		0.958 0.891	Visualization of most contributing hippocampus areas obtained from CNN class activation mapping
Dyrba et al. [11]	MRI	189	219	254	CNN for coronal slices covering hippocampus	AD/CN MCI/CN			0.93 0.75	Visualization of LRP and other methods' relevance maps and comparison by diagnostic group
Lian et al. [25]	MRI	Training 199 Test 159	167/226 38/239	229 200	CNN	AD/CN MCIc/MCInc	90.3% 80.9%		0.951 0.781	Visualization of most contributing image areas obtained from CNN class activation mapping
Qiu et al. [26]	MRI	Training 188 $Test_1$ 62 $Test_2$ 29 $Test_3$ 209	– – – –	229 320 73 356	FCN	AD/CN_1 AD/CN_2 AD/CN_3	87.0% 76.6% 81.8%		0.870 0.892 0.881	Visualization of most contributing brain areas obtained from occlusion sensitivity analysis

Table 1 (continued)

Study (chronologic order)	Data type	Sample			Algorithm	Performance				Addressed model comprehensibility
		AD	MCI c/nc	CN		Groups	Accuracy	Balanced accuracy	AUC	
Wen et al. [27]	MRI	Training 336 $Test_1$ 76 $Test_2$ 78	295/298 20/13 –	330 429 76	CNN	AD/CN_1 $MCIc/MCInc_1$ AD/CN_2		86% 50% 70%		No
Thibeau-Sutre et al. [8]	MRI	Training 336 Test 76	– –	330 429	CNN	AD/CN		90%		Visualization of most contributing brain areas obtained from occlusion sensitivity analysis
Jo et al. [28]	Tau-PET	66	–	66	CNN	AD/CN	90.8%			Visualization of LRP relevance maps, visualization of most contributing brain areas obtained from occlusion sensitivity analysis

Empty cells in the performance columns indicate that the respective values were not reported

AD Alzheimer's dementia, *MCI* mild cognitive impairment, *MCIc* MCI converted to dementia, *MCInc* non-converter/stable MCI, *CN* cognitively normal controls, *DTI* diffusion tensor imaging, *FCN* fully connected network, *RBM* restricted Boltzmann machine, *RNN* recurrent neural network, *CNN* convolutional neural network, *MRI* T1-weighted magnetic resonance imaging, *GM* gray matter volume, *FDG-PET* glucose metabolism derived from fluorodeoxyglucose positron emission tomography

Table 2 Summary of sample characteristics

Sample	CN	MCI	AD
ADNI-GO/2 (training) *N* = 663			
Sample size (female)	254 (130)	220 (93)	189 (80)
Age (*SD*)	75.4 (6.6)	74.1 (8.1)	75.0 (8.0)
Education (*SD*)	16.4 (2.7)	16.2 (2.8)	15.9 (2.7)
MMSE (*SD*)	29.1 (1.2)	27.6 (1.9)	22.6 (3.2)
RAVLT Delayed recall (*SD*)	7.6 (4.1)	3.2 (3.7)	0.8 (1.9)
WMS-LM Delayed recall (*SD*)	13.9 (3.7)	5.1 (3.8)	1.5 (2.1)
Hippocampus volume (*SD*) mm^3	6235 (756)	5619 (963)	4834 (930)
Amyloid status (neg/pos)	177/77	79/141	28/161
MRI field strength (1.5T/3T)	71/183	49/171	35/154
ADNI-3 (validation) *N* = 575			
Sample size (female)	326 (211)	187 (85)	62 (27)
Age (*SD*)	70.0 (7.5)	72.2 (7.5)	74.8 (7.7)
Education (*SD*)	16.6 (2.2)	16.6 (2.5)	16.5 (2.4)
MMSE (*SD*)	29.1 (1.1)	27.8 (2.0)	23.1 (3.3)
RAVLT Delayed recall (*SD*)	8.3 (4.4)	4.7 (4.7)	0.3 (0.9)
WMS-LM Delayed recall (*SD*)	13.0 (3.5)	7.2 (3.9)	2.0 (2.8)
Hippocampus volume (*SD*) mm^3	6583 (649)	6112 (902)	4839 (978)
Amyloid status (neg/pos)	75/39	19/27	3/17
MRI field strength (1.5T/3T)	0/326	0/187	0/62
AIBL (validation) *N* = 606			
Sample size (female)	448 (260)	96 (46)	62 (36)
Age (*SD*)	72.4 (6.2)	74.3 (6.9)	73.2 (7.3)
MMSE (*SD*)	28.7 (1.2)	27.0 (2.2)	21.2 (5.3)
WMS-LM Delayed recall (*SD*)	11.2 (4.3)	4.9 (4.0)	1.0 (1.9)
Hippocampus volume (*SD*) mm^3	6362 (704)	5712 (1028)	4940 (1055)
Amyloid status (neg/pos)	316/101	34/54	6/53
MRI field strength (1.5T/3T)	55/393	7/89	2/60
DELCODE (validation) *N* = 474			
Sample size (female)	215 (124)	155 (72)	104 (61)
Age (*SD*)	69.5 (5.5)	73.0 (5.7)	75.2 (6.2)
Education (*SD*)	14.7 (2.7)	14.0 (3.1)	12.9 (3.1)
MMSE (*SD*)	29.5 (0.8)	27.8 (2.0)	23.1 (3.2)
WMS-LM Delayed recall (*SD*)	14.3 (3.6)	7.4 (5.2)	1.8 (2.8)
Hippocampus volume (*SD*) mm^3	6543 (679)	5665 (950)	4610 (944)
Amyloid status (neg/pos)	58/28	30/57	5/49
MRI field strength (1.5T/3T)	0/215	0/155	0/104

Numbers indicate mean and standard deviation (*SD*) if not indicated otherwise. Years of education were not available for the AIBL dataset. RAVLT Delayed recall scores were not available for the AIBL and DELCODE samples

CN cognitively normal controls, *MCI* amnestic mild cognitive impairment, *AD* Alzheimer's dementia, *SD* standard deviation, *MMSE* Mini Mental State Examination, *RAVLT* Rey Auditory Verbal Learning Test, *WMS-LM* Wechsler Memory Scale Logical Memory Test, *MRI* magnetic resonance imaging

first dataset was compiled from $N = 575$ participants of the recent ADNI-3 phase. The second dataset included MR images from $N = 606$ participants of the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) (https://aibl.csiro.au), provided via the ADNI system. A summary of the diagnostic criteria and additional information is available at https://aibl.csiro.au/about. For AIBL, we additionally obtained amyloid PET scans which were available for 564 participants (93%). The PET scans were processed using the Centiloid SPM pipeline and converted to Centiloid values as recommended for the different amyloid PET traces [32–34]. Amyloid-beta status of the participants was determined using the cutoff 24.1 CL [33]. As a third sample, we included data from $N = 474$ participants of the German Center for Neurodegenerative Diseases (DZNE) multicenter observational study on Longitudinal Cognitive Impairment and Dementia (DELCODE) [35]. Comprehensive information on the diagnostic criteria and study design are provided in [35]. For the DELCODE sample, cerebrospinal fluid (CSF) biomarkers were available for a subsample of 227 participants (48%). Amyloid-beta status was determined using the Aβ42/Aβ40 ratio with a cutoff 0.09 [35].

Image preparation and processing

All MRI scans were preprocessed using the Computational Anatomy Toolbox (CAT12, v9.6/r7487) [36] for Statistical Parametric Mapping 12 (SPM12, v12.6/r1450, Wellcome Centre for Human Neuroimaging, London, UK). Images were segmented into gray and white matter, spatially normalized to the default CAT12 brain template in Montreal Neurological Institute (MNI) reference space using the DARTEL algorithm, resliced to an isotropic voxel size of 1.5 mm, and modulated to adjust for expansion and shrinkage of the tissue. Initially and after all processing steps, all scans were visually inspected to check for image quality. In all scans, effects of the covariates age, sex, total intracranial volume (TIV), and scanner magnetic field strength (FS) were reduced using linear regression. This step was performed, as these factors are known to affect the voxel intensities or regional brain volume [37, 38]. For each voxel vx_{ij}, linear models were fitted on the healthy controls:

$$vx_{ij} = \beta_{i0} + \beta_{i1} age_j + \beta_{i2} sex_j + \beta_{i3} TIV_j + \beta_{i4} FS_j + \varepsilon_{ij} \quad (1)$$

with i being the voxel index, j being the healthy participant index, β_i being the respective model coefficients (for each voxel), and ε_i being the error term or residual. Subsequently, the predicted voxel intensities were subtracted from all participants' gray matter maps to obtain the residual images:

$$res_{ij} = vx_{ij} - \left(\beta_{i0} + \beta_{i1} age_j + \beta_{i2} sex_j + \beta_{i3} TIV_j + \beta_{i4} FS_j\right) \quad (2)$$

Notably, we performed the estimation process (1) only for the healthy ADNI-GO/2 participants. Then, (2) was applied to all other participants and the validation samples. This method was applied as brain volume, specifically in the temporal lobe and hippocampus, is substantially decreasing/shrinking in old age independently of the disease process [37, 38], and we expected this approach to increase accuracy. As sensitivity analysis, we also repeated CNN training on the raw gray matter volume maps for comparison. Patients with MCI and AD were combined into one disease-positive group. On the one hand, this was done as we observed a low sensitivity of machine learning models for MCI when trained only on AD cases, due to the much larger and more heterogeneous patterns of atrophy in AD than in MCI, where atrophy is specifically present in medial temporal and parietal regions [39]. On the other hand, combining both groups substantially increased the training sample, which was required to reduce the overfitting of the CNN models.

CNN model structure and training

The CNN layer structure was adapted from [14, 27], which was inspired by the prominent 2D image detection networks AlexNet [29] and VGG [30]. The model was implemented in Python 3.7 with Keras 2.2.4 and Tensorflow 1.15. The layout is shown in Fig. 1. The residualized/raw 3D images with a resolution of 100 × 100 × 120 voxels were fed as input into the neural network and processed by three consecutive convolution blocks including 3D convolutions (5 filters of 3 × 3 × 3 kernel size) with rectified linear activation function (ReLU), maximum pooling (2 × 2 × 2 voxel patches), and batch normalization layers (Fig. 1). Then, three dropout (10%) and fully connected layers with ReLU activation followed, each consisting of 64, 32, and 2 neurons, respectively. The weights of last two layers were regularized with the L2 norm penalty. The last layer had the softmax activation function that rescaled the class activation values to likelihood scores. The network required approximately 700,000 parameters to be estimated.

The whole CNN pipeline was evaluated by stratified tenfold cross-validation, partitioning the ADNI-GO/2 sample into approximately 600 training and 60 test images with almost equal distribution of CN, MCI, and AD cases. Additionally, data augmentation was used. All images included in the respective training subsamples were flipped along the coronal (L/R) axis and also translated by ±10 voxels in each direction ($x/y/z$), yielding fourteen times increased number of samples per epoch of approximately 8350 images. The CNN model was then trained with the ADAM optimizer, applying the categorical cross-entropy loss function, the learning rate of 0.0001, and a batch size of 20. As the training group sizes were imbalanced, we set class weights of 1.31 for controls and 0.81 for MCI/AD in order to circumvent biased predictions. The weights were determined using the formula $0.5n/n_i$ as recommended in the TensorFlow tutorial [40]. To select the optimal models during training, we set the number of epochs to ten and saved the model state (epoch) which performed best on the test partition. On a Windows 10 computer with Intel Core i5-9600 hexa-core CPU, 64 GB working memory, and NVIDIA GeForce GTX 1650 CUDA GPU, training took approximately 35 min per fold and 12 h in total. All ten models were saved to disk for further inspection and validation. As control

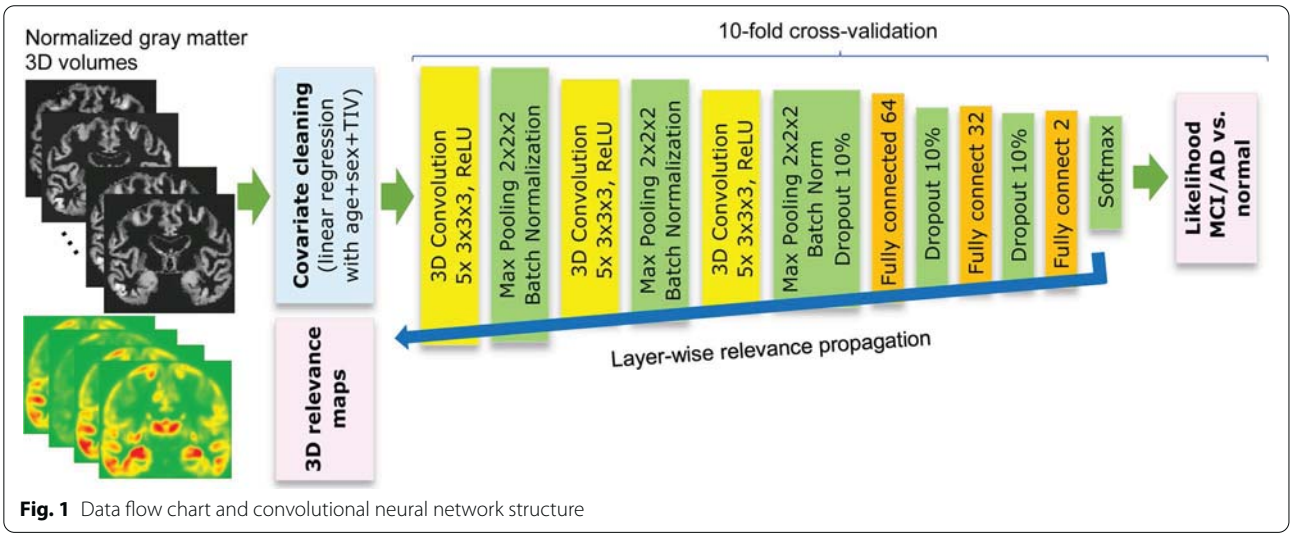


Fig. 1 Data flow chart and convolutional neural network structure

analysis, we also repeated the whole procedure based on the raw image data (normalized gray matter volumes) instead of using the residuals as CNN input. Here, we set the number of epochs to 20 due to slower convergence of the models.

We also trained CNN models on the whole ADNI-GO/2 sample for further evaluation. Here, we fixed the number of epochs to 4 for the residualized data and 8 for the raw data. These values provided the highest average accuracy and lowest loss in the previous cross-validation.

Model evaluation

The balanced accuracy and area under the receiver operating characteristic curve (*AUC*) were calculated for the independent validation samples. We report first the numbers for the model trained on the whole ADNI-GO/2 dataset and second the average values for the models obtained via cross-validation.

As an internal validity benchmark, we compared CNN model performance and group separation using hippocampus volume, the best-established MRI marker for Alzheimer's disease. Automated extraction of hippocampus volume is already implemented in commercial radiology software to aid physicians in diagnosing dementia. We extracted total hippocampus volume from the modulated and normalized MRI scans using the Automated Anatomical Labeling (AAL) atlas [41]. The extracted volumes were corrected for the effects of age, sex, total intracranial volume, and magnetic field strength of the MRI scanner in the same way as described above for the CNN input (see the section "Image preparation and processing"). Here, a linear model was estimated based on the normal controls of the ADNI-GO/2 training sample, and then, the parameters were applied to the measures of all other participants and validation samples to obtain the residuals. Subsequently, the residuals of the training sample were entered into a receiver operating characteristic analysis to obtain the *AUC*. The optimal threshold providing the highest accuracy was selected based on the Youden index. We obtained two thresholds. One for the separation of MCI and controls, which was the residual volume of −0.63 ml. That means participants with the deviation of individual hippocampus volume from the expected value (for that age, sex, total brain volume, and magnetic field strength) below −0.63 ml were classified as MCI. The other threshold for AD dementia and controls was −0.95 ml. Additionally, we repeated the same cross-validation training/test splits as used for CNN training to compare the variability of the derived thresholds and performance measures.

CNN relevance map visualization

Relevance maps were derived from the CNN models using the LRP algorithm [4] implemented in the Python package iNNvestigate 1.0.9 [42]. LRP has previously been demonstrated to yield relevance maps with high spatial resolution and clinical plausibility [11, 14]. In this approach, the final network activation scores for a given input image are propagated back through the network layers. LRP applies a relevance conservation principle that means that the total amount of relevance per layer is kept constant during the back-tracing procedure to reduce numerical challenges that occur in other methods [4]. Several rules exist, which apply different weighting to positive (excitatory) and negative (inhibitory) connections such that network activation for and against a specific class can be considered differentially. Here, we applied the so-called $\alpha = 1, \beta = 0$ rule that only considers positive relevance as proposed by [11, 14]. In this case, the relevance of a network neuron R_j was calculated from all connected neurons k in the subsequent network layer using the formula:

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j \left(a_j w_{jk}^+\right)} R_k \tag{3}$$

with a_j being the activation of neuron j, w_{jk}^+ being the positive weight of the connection between neurons j and k, and R_k being the relevance attributed to neuron k [5]. As recent studies reported further improvements in LRP relevance attribution [43, 44], we applied the LRP $\alpha = 1, \beta = 0$ composition rule that applies (3) to the convolutional layers, and the slightly extended ϵ rule [5] to the fully connected layers. In the ϵ rule, (3) is being extended by a small constant term added to the denominator, i.e., $\epsilon = 10^{-10}$ in our case, which is expected to reduce relevance when the activation of neuron k is weak or contradictory [5].

To facilitate model assessment and quick inspection of relevance maps, we implemented an interactive Python visualization application that is capable of immediate switching between CNN models and participants. More specifically, we used the Bokeh Visualization Library 2.2.3 (https://bokeh.org). Bokeh provides a webserver backend and web browser frontend to directly run Python code that dynamically generates interactive websites containing various graphical user interface components and plots. The Bokeh web browser JavaScript libraries handle the communication between the browser and server instance and translate website user interaction into Python function calls. In this way, we implemented various visualization components to adjust plotting parameters and provide easy navigation for the 2D slice views obtained from the 3D MRI volume.

The application is structured following a model–view–controller paradigm. An overview of implemented functions is provided in Supplementary Fig. 1. A sequence diagram illustrating function calls when selecting a new person is provided in Supplementary Fig. 2. The source code and files required to run the interactive visualization are publicly available via https://github.com/martindyrba/DeepLearningInteractiveVis.

As core functionality, we implemented the visualization in a classical 2D multi-slice window with axial, coronal, and sagittal views, cross-hair, and sliders to adjust the relevance threshold as well as minimum cluster size threshold (see Fig. 2). Here, a cluster refers to groups of adjacent voxels with high relevance above the selected relevance threshold. The cluster size is the number of voxels in this group and can be controlled in order to reduce the visual noise caused by single voxels with high relevance. Additionally, we added visual guides to improve usability, including (a) a histogram providing the distribution of cluster sizes next to the cluster size threshold slider, (b) plots visualizing the amount of positive and negative relevance per slice next to the slice selection sliders, and (c) statistical information on the currently selected cluster. Furthermore, assuming spatially normalized MRI data in MNI reference space, we added (d) atlas-based anatomical region lookup for the current cursor/cross-hair position and (e) the option to display the outline of the anatomical region to simplify visual comparison with the cluster location.

CNN model comprehensibility and validation

As quantitative metrics for assessing relevance map quality are still missing, we compared CNN relevance scores in the hippocampus with hippocampus volume. Here, we used the same AAL atlas hippocampus masks as for deriving hippocampus volume and applied it on the relevance maps obtained from all ADNI-GO/2 participants for each model. The sum of relevance score of each voxel inside the mask was considered as hippocampus relevance. Hippocampus relevance and volume were compared using Pearson's correlation coefficient.

Additionally, we visually examined a large number of scans from each group to derive common relevance patterns and match them with the original MRI scans. Furthermore, we calculated mean relevance maps for each group. We also extracted the relevance for all lobes of the brain and subcortical structures to test the specificity of relevance distribution across the whole brain. These masks were defined based on the other regions included in the AAL atlas [41].

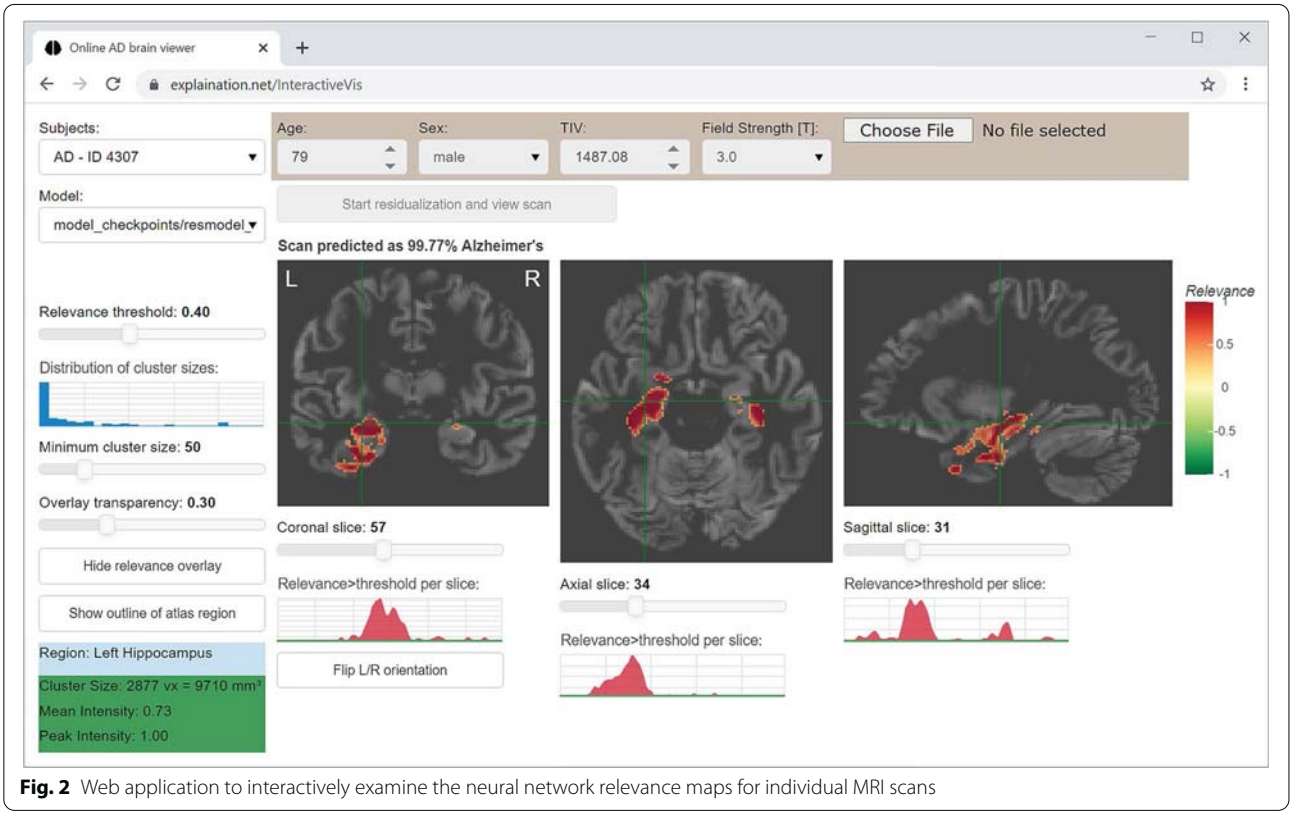


Fig. 2 Web application to interactively examine the neural network relevance maps for individual MRI scans

In an occlusion sensitivity analysis, we evaluated the influence of local atrophy on the prediction of the model and the derived relevance scores. Here, we slid a cube of 20 voxels = 30 mm edge size across the brain. Within the cube, we reduced the intensity of the voxel by 50%, simulating gray matter atrophy in this area. We selected a normal control participant from the DELCODE dataset without visible CNN relevance, a prediction probability for AD/MCI of 20%, and hippocampus volume residual of 0 ml, i.e., the hippocampus volume matched the reference volume expected for this person. For each position of the cube, we derived the probability of AD predicted by the model obtained from the whole ADNI-GO/2 sample. Additionally, we calculated the total amount of relevance in the scan.

Results

Group separation

The accuracy and *AUC* for diagnostic group separation are shown in Table 3. Additional performance measures are provided in Supplementary Table 1. The CNN reached a balanced accuracy between 75.5 and 88.3% across validation samples with an *AUC* between 0.828 and 0.978 for separating AD dementia and controls. For MCI vs. controls, the group separation was substantially lower with balanced accuracies between 63.1 and 75.4% and an *AUC* between 0.667 and 0.840. These values were only slightly better than the group separation performance of hippocampus volume (Table 3). The performance results for the raw gray matter volume data as input for the CNN are provided in Supplementary Table 2. In direct comparison to the CNN results for the residualized data, the balanced accuracies and *AUC* values did not show a clear difference (Table 3, Supplementary Table 2).

Model comprehensibility and relevance map visualization

The implemented web application frontend is displayed in Fig. 2. The source code is available at https://github.com/martindyrba/DeepLearningInteractiveVis and the web application can be publicly accessed at https://explaination.net/demo. In the left column, the user can select a study participant and a specific model. Below, there are controls (sliders) to adjust the thresholds for displayed relevance score, cluster size, and overlay transparency. As we used the spatially normalized MRI images as CNN input, we can directly obtain the anatomical reference location label from the automated

Table 3 Group separation performance for hippocampus volume and the convolutional neural network models

Sample	Hippocampus volume (residuals)		3D convolutional neural network	
	Balanced accuracy (mean ± *SD*)	*AUC*	Balanced accuracy (mean ± *SD*)	*AUC* (mean ± *SD*)
ADNI-GO/2				
MCI vs. CN	(70.0% ± 6.8%)	(0.773 ± 0.091)	*(74.5% ± 6.2%)*	*(0.785 ± 0.078)*
AD vs. CN	(84.4% ± 3.6%)	(0.945 ± 0.024)	*(88.9% ± 5.3%)*	*(0.949 ± 0.029)*
MCI^+ vs. CN^-	(75.6% ± 7.1%)	(0.831 ± 0.080)	*(86.7% ± 10.3%)*	*(0.925 ± 0.071)*
AD^+ vs. CN^-	(86.2% ± 4.2%)	(0.954 ± 0.025)	*(94.9% ± 3.8%)*	*(0.985 ± 0.017)*
ADNI-3				
MCI vs. CN	62.8% (63.1% ± 1.4%)	0.683	63.1% (63.6% ± 1.5%)	0.684 (0.677 ± 0.020)
AD vs. CN	83.4% (83.4% ± 0.4%)	0.917	84.4% (81.7% ± 2.9%)	0.913 (0.899 ± 0.013)
MCI^+ vs. CN^-	69.1% (69.2% ± 2.7%)	0.791	69.8% (68.3% ± 4.4%)	0.810 (0.742 ± 0.024)
AD^+ vs. CN^-	83.6% (82.0% ± 1.8%)	0.882	80.2% (75.5% ± 4.2%)	0.830 (0.828 ± 0.028)
AIBL				
MCI vs. CN	67.4% (67.6% ± 0.5%)	0.741	68.2% (67.3% ± 2.7%)	0.763 (0.749 ± 0.012)
AD vs. CN	84.1% (85.3% ± 1.5%)	0.927	85.0% (82.3% ± 3.0%)	0.950 (0.926 ± 0.007)
MCI^+ vs. CN^-	78.5% (78.8% ± 0.9%)	0.874	75.4% (73.6% ± 3.1%)	0.828 (0.814 ± 0.022)
AD^+ vs. CN^-	87.2% (89.1% ± 2.4%)	0.976	88.3% (85.3% ± 3.3%)	0.978 (0.958 ± 0.011)
DELCODE				
MCI vs. CN	69.0% (69.0% ± 9.6%)	0.774	71.0% (69.7% ± 2.6%)	0.775 (0.772 ± 0.017)
AD vs. CN	88.4% (86.4% ± 3.0%)	0.943	85.5% (80.5% ± 4.0%)	0.953 (0.938 ± 0.013)
MCI^+ vs. CN^-	77.4% (77.8% ± 0.7%)	0.867	72.2% (74.9% ± 3.5%)	0.840 (0.830 ± 0.017)
AD^+ vs. CN^-	88.2% (87.6% ± 1.8%)	0.954	83.3% (82.2% ± 4.0%)	0.968 (0.956 ± 0.012)

Reported values are for the single model trained on the whole ADNI-GO/2 dataset. In parenthesis, the mean values and standard deviation for the ten models trained in the tenfold cross-validation procedure are provided to indicate the variability of the measures. Values for the ADNI-GO/2 sample (in italics) may be biased as the respective test subsamples were used to determine the optimal model during training. We still report them for better comparison of the model performance across samples

anatomical labeling (AAL) atlas [41] given the MNI coordinates at the specific cross-hair location, which is displayed in the light blue box. The green box displays statistics on the currently selected relevance cluster such as number of voxels and respective volume. In the middle part of Fig. 2, the information used as covariates (age, sex, total intracranial volume, MRI field strength) and the CNN likelihood score for AD are depicted above the coronal, axial, and sagittal views of the 3D volume. We further added sliders and plots of cumulated relevance score per slices as visual guides to facilitate navigation to slices with high relevance. All user interactions are directly sent to the server, evaluated internally, and updated in the respective views and control components in real-time without major delay. For instance, adjusting the relevance threshold directly changes the displayed brain views, the shape of the red relevance summary plots, and the blue cluster size histogram. A sequence diagram of internal function calls when selecting a new participant is illustrated in Supplementary Fig. 2.

Individual people's relevance maps are illustrated in Fig. 3. The group mean relevance maps for the DELCODE validation sample are shown in Fig. 4 and those for the ADNI-GO/2 training sample in Supplementary Fig. 3. They are very similar to traditional statistical maps obtained from voxel-based morphometry, indicating the highest contribution of medial temporal brain regions, more specifically the hippocampus, amygdala, thalamus, middle temporal gyrus, and middle/posterior cingulate cortex. Also, they were highly consistent between samples (Supplementary Fig. 3). The occlusion sensitivity analysis also showed identical brain regions' atrophy to contribute to the model's decision (Fig. 5). Interestingly, the occlusion relevance maps showed a ring structure around the most contributing brain areas, indicating that relevance was highest when the occluded area just touched the salient regions, leading to a thinning-like shape of the gray matter.

The correlation of individual DELCODE participants' hippocampus relevance score and hippocampus volume for the model trained on the whole ADNI-GO/2 dataset is displayed in Fig. 6. For this model, the correlation was $r = -0.87$ for bilateral hippocampus volume ($p < 0.001$). Across all ten models obtained using cross-validation, the median correlation of total hippocampus relevance and volume was $r = -0.84$ with a range of -0.88 and -0.44 (all with $p < 0.001$). Cross-validation models with higher correlation between hippocampus relevance and volume showed a tendency for better *AUC* values for MCI vs. controls ($r = 0.61$, $p = 0.059$). To test whether hippocampus volume and relevance measures were specific to the hippocampus, we also compared the correlation between hippocampus volume and other regions' and whole-brain relevance. Here, the correlations were lower, with $r = -0.62$ ($p < 0.001$) between hippocampus volume and whole-brain relevance. More detailed results are provided as a correlation matrix in Supplementary Fig. 4.

Discussion

Neural network comprehensibility

We have presented a CNN framework and interactive visualization application for obtaining class-specific relevance maps for disease detection in MRI scans, yielding human-interpretable and clinically plausible visualizations of key features for image discrimination. To date, most CNN studies focus on model development and optimization, which are undoubtedly important tasks and there are still several challenges to tackle. However, as black-box models, it is typically not feasible to judge, why a CNN fails or which image features drive a particular decision of the network. This gap might be closed with the use of novel visualization algorithms such as LRP [4] and deep Taylor decomposition [5]. In our application, LRP relevance maps provided a useful tool for model inspection to reveal the brain regions which contributed most to the decision process encoded by the neural network models.

Currently, there is no ground truth information for relevance maps, and there are no appropriate methods available to quantify relevance map quality. Samek and colleagues [45] proposed the information-theoretic measures relevance map entropy and complexity, which mainly characterize the scatter or smoothness of images. Furthermore, adapted from classical neural network sensitivity analysis, they assessed the robustness of relevance maps using perturbation testing where small image patches were replaced by random noise, which was also applied in [46]. Already for 2D data, this method is computationally very expensive and only practical for a limited number of input images. Instead of adding random noise, we simulated gray matter atrophy by lowering the image intensities by 50% in a cube-shaped area. As visible from Fig. 5, the brain areas contributing to the model's AD probability nicely matched the areas shown in the mean relevance maps (Fig. 4). Notably, the ring-shaped increase in relevance around the salient regions (Fig. 5, bottom) indicates that the model is sensitive to intensity jumps occurring when the occlusion cube touches the borderline of those regions. Most probably, this means that the model was more sensitive to thinning patterns of gray matter than to equally distributed volume reduction. However, our findings have to be seen as preliminary, as we only assessed this analysis in one normal control

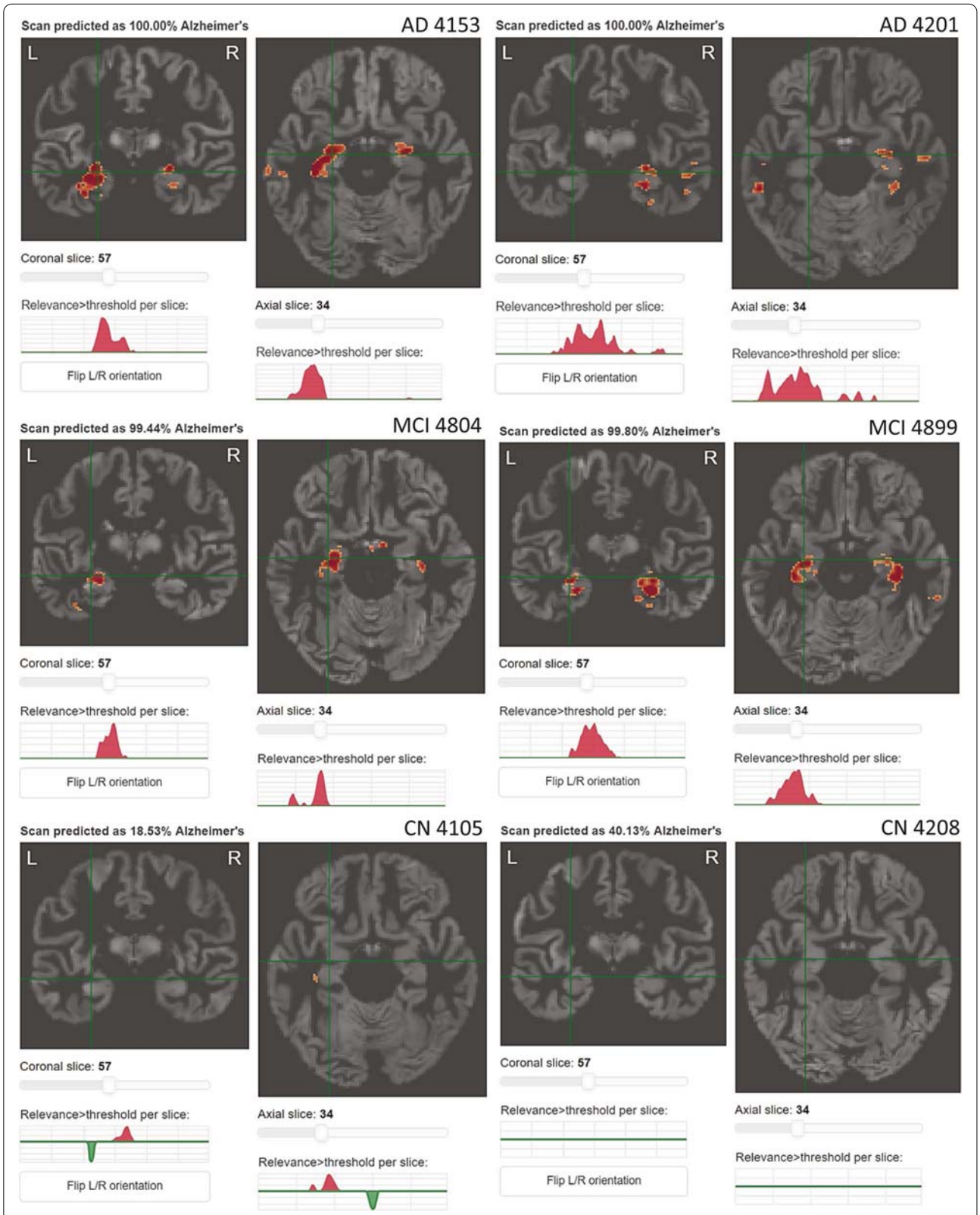


Fig. 3 Example relevance maps obtained for different people. Top row: Alzheimer's dementia patients, middle row: patients with mild cognitive impairment, bottom row: cognitively normal controls

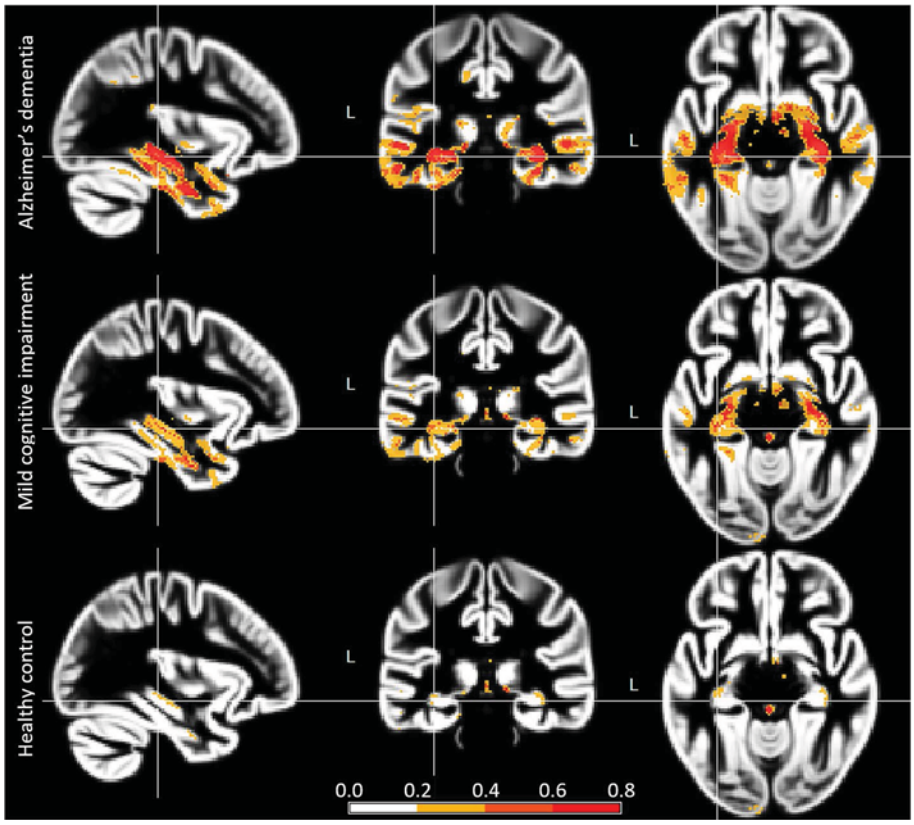


Fig. 4 Mean relevance maps for Alzheimer's dementia patients (top row), patients with mild cognitive impairment (middle row), and healthy controls (bottom row) for the DELCODE validation sample. Relevance maps thresholded at 0.2 for better comparison

participant due to the computational effort, and therefore, it requires more extensive research in future studies.

Based on the extensive knowledge about the effect of Alzheimer's disease on brain volume as presented in T1-weighted MRI scans [15, 16], we selected a direct quantitative comparison of relevance maps with hippocampus volume as a validation method. Here, we obtained very high correlations between hippocampus relevance scores and volume (median correlation $r = -0.84$), underlining the clinical plausibility of learnt patterns to differentiate AD and MCI patients from controls. In addition, visual inspection of relevance maps also revealed several other clusters with gray matter atrophy in the individual participants' images that contributed to the decision of the CNN (Figs. 2 and 3). Böhle and colleagues [14] proposed an atlas-based aggregation of CNN relevance maps to be used as "disease fingerprints" and to enable a quick comparison between patients and controls, a concept that has also been proposed previously for differential diagnosis of dementia based on heterogeneous clinical data and other machine learning models [47, 48].

Notably, the CNN models presented here were solely based on the combinations of input images with their corresponding diagnostic labels to determine which brain features were diagnostically relevant. Traditionally, extensive clinical experience is required to define relevant features (e.g., hippocampus volume) that discriminate between a clinical population (here: AD, MCI) and a healthy control group. Also, typically, only few predetermined parameters are used (e.g., hippocampus volume or medial temporal lobe atrophy score [15, 16]). Our results demonstrate that the combination of CNN and relevance map approaches constitutes a promising tool for improving the utility of CNN in the classification of MRIs of patients with suspected AD in a clinical context. By referring back to the relevance maps, trained clinicians will be enabled to compare classification results to comprehensible features visible in the relevance images and thereby more readily interpret the classification results in clinically ambiguous situations. Perspectively, the relevance map approach might also provide a helpful tool to reveal features for more complex diagnostic challenges such as differential diagnosis between various types of dementia,

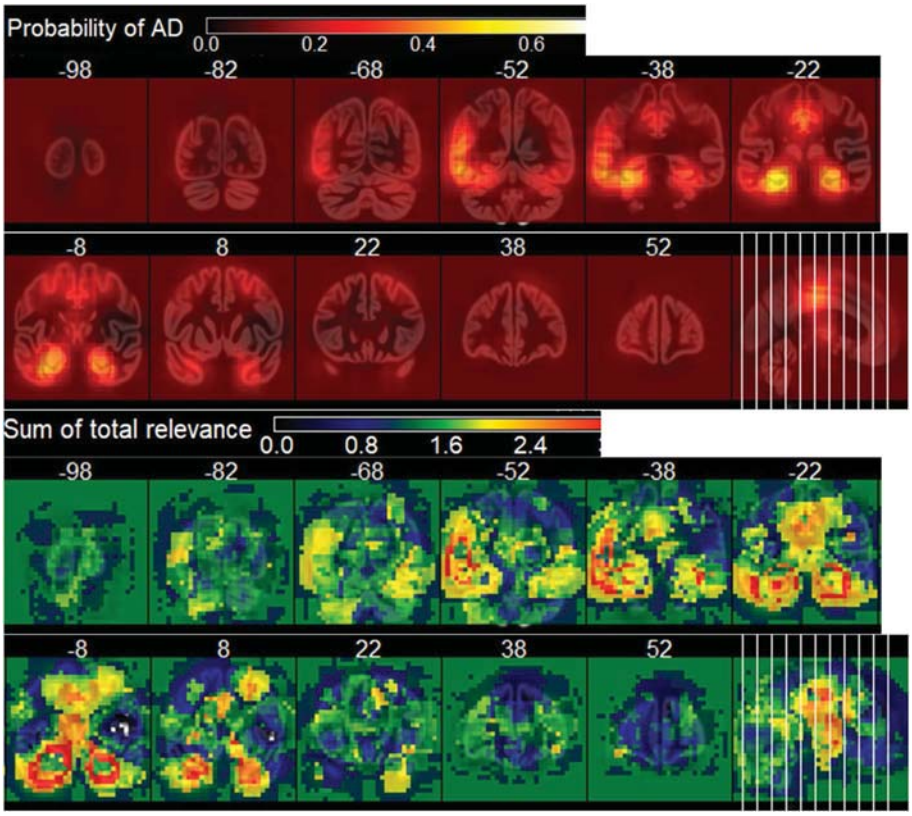


Fig. 5 Results from the occlusion sensitivity analysis. A gray matter volume loss of 50% was simulated in a cube of 30-mm edge length. Each voxel encodes the derived values when centering the cube at that position. Top: probability of AD for the areas with simulated atrophy. Bottom: total sum of image relevance depending on simulated atrophy. Numbers indicate the *y*-axis slice coordinates in MNI reference space

for instance the differentiation between AD, frontotemporal dementia, and dementia with Lewy bodies.

CNN performance

As expected, CNN-based classification reached an excellent $AUC \geq 0.91$ for the group separation of AD compared to controls but a substantially lower accuracy for group separation between MCI and controls ($AUC \approx 0.74$, Table 3). When restricting the classification to amyloid-positive MCI versus amyloid-negative controls, group separation improved to $AUC = 0.84$ in DELCODE, highlighting the heterogeneity of MCI as a diagnostic entity and the importance of biomarker stratification [1, 2]. In summary, these numbers are also reflected by the recent CNN literature as shown in Table 1. Notably, [27] reported several limitations and issues in the performance evaluation of some other CNN papers, such that it is not easy to finally conclude on the group separation capabilities of the CNN models in realistic settings. To overcome such challenges, we validated the models on three large independent cohorts (Table 3), providing strong evidence for their generalizability and for the robustness of our CNN approach.

To put the CNN model performance into perspective, we compared the accuracy of the CNN models with the accuracy achieved by assessing hippocampus volume, the key clinical MRI marker for neurodegeneration in Alzheimer's disease [1, 2]. Interestingly, there were only minor differences in the achieved *AUC* values across all samples (Table 3). The MCI group of the ADNI-3 sample, which yielded the worst group separation of all samples ($AUC = 0.68$), was actually the group with the largest average hippocampus volumes and, therefore, the lowest group difference compared to the controls (Table 2). Obviously, our results here indicate a limited value of using CNN models instead of traditional volumetric markers for the detection of Alzheimer's dementia and mild cognitive impairment. Previous MRI CNN papers have not reported the baseline accuracy reached by hippocampus volume for comparison. However, as noted above, CNNs might provide a useful tool to automatically derive discriminative features for complex diagnostic tasks where

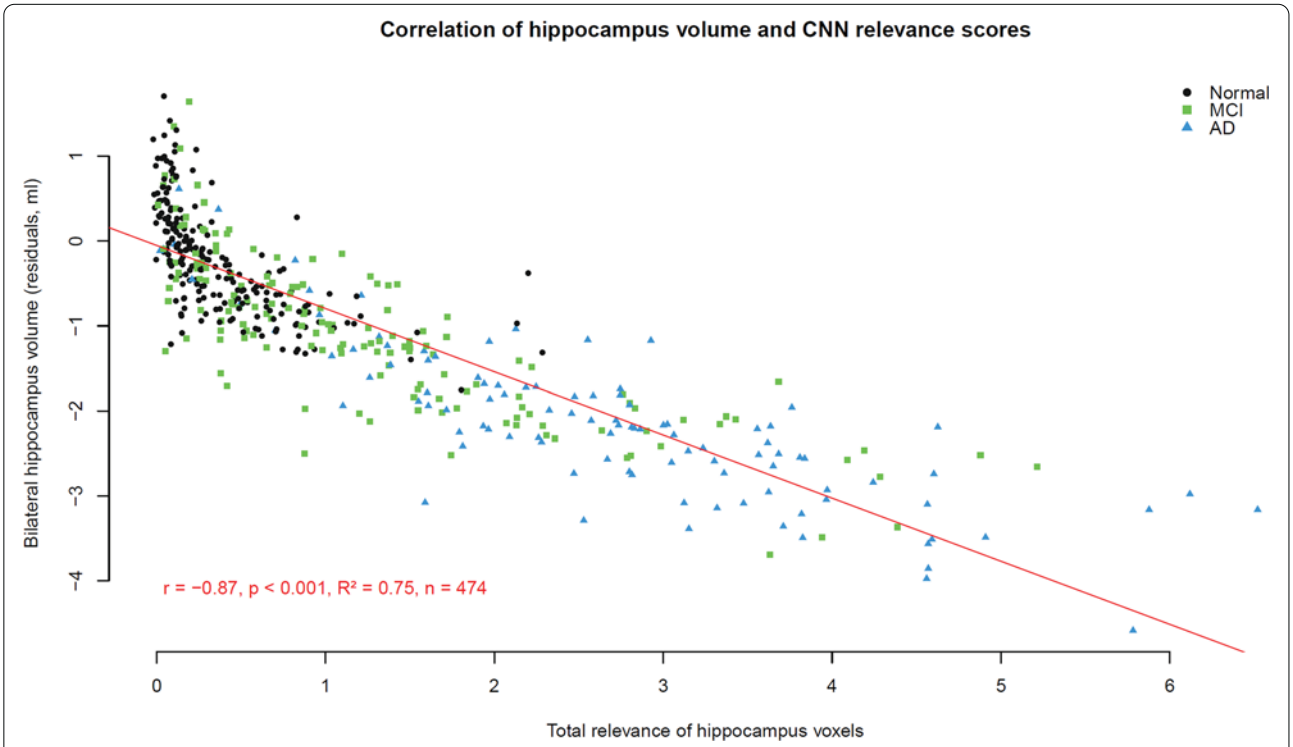


Fig. 6 Scatter plot and correlation of bilateral hippocampus volume and neural network relevance scores for the hippocampus region for the DELCODE sample ($r = -0.87$, $p < 0.001$)

clear clinical criteria are still missing, for instance for the differential diagnosis between various types of dementia.

Limitations

As already mentioned above, visual inspection of relevance maps also revealed several other regions with gray matter atrophy in the individual participants' images that contributed to the decision of the CNN. These additional regions were not further assessed, as a priori knowledge regarding their diagnostic value is still under debate in the scientific community [1, 2]. Also, we did not perform a three-way classification between AD dementia, MCI, and CN due to the limited availability of cases for training. Additionally, MCI itself is a heterogeneous diagnostic entity [1, 2]. Here, all the studies involved in our analysis tried to increase the likelihood of underlying Alzheimer's pathology by focusing on MCI patients with memory impairment. But markers of amyloid-beta pathology were only available for a subset of participants such that we could not stratify by amyloid status for the training of the CNN models. However, we optionally applied this stratification for the validation of the CNN performances to improve the diagnostic confidence.

Future prospects

Several studies focused on CNN models for the integration of multimodal imaging data, e.g., MRI and fluorodeoxyglucose (FDG)-PET [17–19], or heterogeneous clinical data [49]. Here, it will be beneficial, to directly include the variables we used as covariates (such as age and sex) as input to the CNN model rather than performing the variance reduction directly on the input data before applying the model. In this context, relevance mapping visualization approaches need to be developed that allow for a direct comparison of the relevance magnitude for images and clinical variables simultaneously. Another aspect is the automated generation of textual descriptions and diagnostic explanations from images [50–52]. Given the recent technical progress, we suggest that the approach is now ready for interdisciplinary exchange to assess how clinicians can benefit from CNN assistance in their diagnostic workup, and which requirements must be met to increase clinical utility. Beyond the technical challenges, regulatory and ethical aspects and caveats must be carefully considered when introducing CNN as part of clinical decision support systems and medical software—and the discussion of these issues has just recently begun [53, 54].

Conclusion

We presented a framework for obtaining diagnostic relevance maps from CNN models to improve model comprehensibility. These relevance maps have revealed reproducible and clinically plausible atrophy patterns in AD and MCI patients, with a high correlation with the well-established MRI marker of hippocampus volume. The implemented web application allows a quick and versatile inspection of brain regions with a high relevance score in individuals. With the increased comprehensibility of CNNs provided by the relevance maps, the data-driven and hypothesis-free CNN modeling approach might provide a useful tool to aid differential diagnosis of dementia and other neurodegenerative diseases, where fine-grained knowledge on discriminating brain alterations is still missing.

Abbreviations

AAL: Automated anatomical labeling; AD: Alzheimer's disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; AIBL: Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; *AUC*: Area under the receiver operating characteristic curve; CAM: Class activation mapping; *CI*: Confidence interval; CN: Cognitively normal participants; CNN: Convolutional neural network; CSF: Cerebrospinal fluid; DELCODE: DZNE multicenter observational study on Longitudinal Cognitive Impairment and Dementia; DZNE: Deutsches Zentrum für Neurodegenerative Erkrankungen (German Center for Neurodegenerative Diseases); FDG: Fluorodeoxyglucose; GM: Gray matter; LRP: Layer-wise relevance propagation; MCI: Mild cognitive impairment; MNI: Montreal Neurological Institute; MRI: Magnetic resonance imaging; PET: Positron emission tomography; ReLU: Rectified linear activation function; *SD*: Standard deviation; SUVR: Standardized uptake value ratio; TIV: Total intracranial volume.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13195-021-00924-2.

Additional file 1: Supplementary Table 1. Group separation performance for hippocampus volume and the convolutional neural network models for residualized data (extended).

Additional file 2: Supplementary Table 2. Group separation performance for hippocampus volume and the convolutional neural network models for raw input data.

Additional file 3: Supplementary Figure 1. UML diagram of the interactive visualization application.

Additional file 4: Supplementary Figure 2. Sequence diagram of function calls when selecting a new person.

Additional file 5: Supplementary Figure 3. Comparison of mean relevance maps between samples.

Additional file 6: Supplementary Figure 4. Correlation matrix of hippocampus volume (residualized) and several brain regions' relevance scores for DELCODE participants and the model trained on the whole ADNI-GO/2 dataset.

Acknowledgements

The data samples were provided by the DELCODE study group of the Clinical Research Unit of the German Center for Neurodegenerative Diseases (DZNE). Details and participating sites can be found at www.dzne.de/en/research/studies/clinical-studies/delcode. The DELCODE study was supported by Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin; Center for Cognitive Neuroscience Berlin (CCNB) at Freie Universität Berlin; Bernstein Center for Computational Neuroscience (BCCN), Berlin; Core Facility MR-Research in Neurosciences, University Medical Center Goettingen; Institute for Clinical Radiology, Ludwig Maximilian University, Munich; Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center; and Magnetic Resonance research center, University Hospital Tuebingen.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. AIBL researchers are listed at aibl.csiro.au.

Authors' contributions

MD: conceptualization, methodology, data curation and processing, coding and software development, visualization, and writing of the original draft. MH: coding and software development and visualization. ST: conceptualization, methodology, data collection and curation, writing, review and editing, supervision, and clinical validation. All others: data acquisition, collection and curation, substantial intellectual contribution on study design and methodology, review and editing, and clinical validation. All authors read and approved the final manuscript.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 454834942, funding code DY151/2-1. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

Data used for training/evaluation of the models is available from the respective initiatives (ADNI: http://adni.loni.usc.edu/data-samples/access-data, AIBL: https://aibl.csiro.au, DELCODE: https://www.dzne.de/en/research/studies/clinical-studies/delcode).

The source code, a demo dataset, the trained CNN models, and all additional files required to run the interactive visualization are publicly available at GitHub: https://github.com/martindyrba/DeepLearningInteractiveVis.

Declarations

Ethics approval and consent to participate

Data collecting within ADNI and AIBL was approved by participating institutions. See https://adni.loni.usc.edu and https://aibl.csiro.au for details. The DELCODE study was approved by participating institutions. See [35] for details. All study participants or their representatives provided written informed consent to participate in the respective studies and also agreed to sharing of their data. The retrospective analysis, study design, and interactive visualization of relevance maps were approved by the internal review board of the Rostock University Medical Center, reference number A 2020-0182.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany. [2]Institute of Visual and Analytic Computing, University of Rostock, Rostock, Germany. [3]German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany. [4]Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Charité Mitte, Berlin, Germany. [5]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany. [6]Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany. [7]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany. [8]Institute for Stroke and Dementia Research (ISD), University Hospital, Ludwig Maximilian University, Munich, Germany. [9]Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center, Rostock, Germany. [10]MR-Research in Neurosciences, Department of Cognitive Neurology, Georg-August-University, Goettingen, Germany. [11]German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany. [12]Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University, Magdeburg, Germany. [13]Bernstein Center for Computational Neuroscience, Berlin, Germany. [14]Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin, Berlin, Germany. [15]Department of Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany. [16]German Center for Neurodegenerative Diseases (DZNE), Tuebingen, Germany. [17]Section for Dementia Research, Hertie Institute for Clinical Brain Research, Tuebingen, Germany. [18]Department of Psychiatry and Psychotherapy, University of Tuebingen, Tuebingen, Germany. [19]Department of Psychiatry, Medical Faculty, University of Cologne, Cologne, Germany. [20]Department of Psychiatry and Psychotherapy, Otto-von-Guericke University, Magdeburg, Germany. [21]Systems Neurophysiology, Department of Biology, Darmstadt University of Technology, Darmstadt, Germany. [22]Department of Psychiatry and Psychotherapy, University Hospital, Ludwig Maximilian University, Munich, Germany. [23]Munich Cluster for Systems Neurology (SyNergy), Ludwig Maximilian University, Munich, Germany. [24]Ageing Epidemiology Research Unit (AGE), School of Public Health, Imperial College London, London, UK. [25]Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University Munich, Munich, Germany. [26]Department for Biomedical Magnetic Resonance, University of Tuebingen, Tuebingen, Germany. [27]German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany. [28]Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, Goettingen, Germany. [29]Leibniz Institute for Neurobiology, Magdeburg, Germany. [30]Department of Neurology, University Hospital Bonn, Bonn, Germany. [31]Institute for Clinical Radiology, Ludwig Maximilian University, Munich, Germany. [32]Department of Medical Imaging, University of Toronto, Toronto, Canada. [33]Neurosciences and Signaling Group, Institute of Biomedicine (iBiMED), Department of Medical Sciences, University of Aveiro, Aveiro, Portugal. [34]Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany.

Received: 2 March 2021 Accepted: 25 October 2021
Published: 23 November 2021

References

1. Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011;7(3):257–62.
2. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. 2014;13(6):614–29.
3. Vemuri P, Fields J, Peter J, Klöppel S. Cognitive interventions in Alzheimer's and Parkinson's diseases. Curr Opin Neurol. 2016;29(4):405–11.
4. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10(7).
5. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Process. 2018;73:1–15.
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 618-626.
7. Zeiler MD, Fergus R: Visualizing and understanding convolutional networks. In: Computer Vision – ECCV 2014. 2014: 818-833.
8. Thibeau-Sutre E, Colliot O, Dormont D, Burgos N, Landman BA, Išgum I. Visualization approach to assess the robustness of neural networks for medical image classification. In: Medical Imaging 2020: Image Processing. 2020.
9. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1135-1144.
10. Alber M: Software and application patterns for explanation methods. In: Explainable AI: interpreting, explaining and visualizing deep learning. 2019: 399-433.
11. Dyrba M, Pallath AH, Marzban EN: Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer's disease. In: Bildverarbeitung für die Medizin. 2020: 307-312.
12. Eitel F, Ritter K: Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. 2019: 3-11.
13. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M: Striving for simplicity: the all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings. Edited by Bengio Y, LeCun Y.
14. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
15. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry. 1992;55(10):967–72.
16. Teipel S, Drzezga A, Grothe MJ, Barthel H, Chételat G, Schuff N, et al. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. Lancet Neurol. 2015;14(10):1037–53.
17. Suk H-I, Lee S-W, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage. 2014;101:569–82.
18. Li F, Tran L, Thung K-H, Ji S, Shen D, Li J. A robust deep model for improved classification of AD/MCI patients. IEEE J Biomed Health Inform. 2015;19(5):1610–6.
19. Ortiz A, Munilla J, Gorriz JM, Ramirez J. Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. Int J Neural Syst. 2016;26(7):1650025.
20. Aderghal K, Khvostikov A, Krylov A, Benois-Pineau J, Afdel K, Catheline G: Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). 2018: 345-350.
21. Liu M, Cheng D, Yan W. Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. Front Neuroinformat. 2018;12.
22. Liu M, Zhang J, Nie D, Yap P-T, Shen D. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J Biomed Health Inform. 2018;22(5):1476–85.
23. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, et al. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci. 2018;12.
24. Li H, Habes M, Wolk DA, Fan Y. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimers Dement. 2019;15(8):1059–70.
25. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell. 2020;42(4):880–93.
26. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920–33.
27. Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal. 2020;63.

28. Jo T, Nho K, Risacher SL, Saykin AJ. Deep learning detection of informative features in tau PET for Alzheimer's disease classification. BMC Bioinformatics. 2020;21(S21).
29. Krizhevsky A, Sutskever I, Hinton GE: ImageNet Classification with deep convolutional neural networks. In: Advances in neural information processing systems 25. Edited by Pereira F, Burges CJC, Bottou L, Weinberger KQ: Curran Associates, Inc; 2012: 1097–1105.
30. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015). Edited by Bengio Y, LeCun Y; 2015.
31. Landau SM, Mintun MA, Joshi AD, Koeppe RA, Petersen RC, Aisen PS, et al. Amyloid deposition, hypometabolism, and longitudinal cognitive decline. Ann Neurol. 2012;72(4):578–86.
32. Klunk WE, Koeppe RA, Price JC, Benzinger TL, Devous MD, Jagust WJ, et al. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. Alzheimers Dement. 2015;11(1):1–15.e14.
33. Navitsky M, Joshi AD, Kennedy I, Klunk WE, Rowe CC, Wong DF, et al. Standardization of amyloid quantitation with florbetapir standardized uptake value ratios to the Centiloid scale. Alzheimers Dement. 2018;14(12):1565–71.
34. Battle MR, Pillay LC, Lowe VJ, Knopman D, Kemp B, Rowe CC, et al. Centiloid scaling for quantification of brain amyloid with [18F] flutemetamol using multiple processing methods. EJNMMI Res. 2018;8(1).
35. Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, et al. Alzheimers Res Ther. 2018;10(1).
36. Kurth F, Gaser C, Luders E. A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (SPM). Nat Protoc. 2015;10(2):293–304.
37. Dima D, Modabbernia A, Papachristou E, Doucet GE, Agartz I, Aghajani M, et al. Subcortical volumes across the lifespan: data from 18,605 healthy individuals aged 3–90 years. Hum Brain Mapp. 2021.
38. Jack CR, Wiste HJ, Weigand SD, Knopman DS, Vemuri P, Mielke MM, et al. Age, sex, andAPOEε4 effects on memory, brain structure, and β-amyloid across the adult life span. JAMA Neurol. 2015;72(5).
39. Grothe MJ, Teipel SJ. Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks. Hum Brain Mapp. 2016;37(1):35–53.
40. TensorFlow Tutorial. Classification on imbalanced data [https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#class_weights]
41. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage. 2002;15(1):273–89.
42. Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. iNNvestigate neural networks! J Mach Learn Res. 2019;20:1–8.
43. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: 2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1-7.
44. Sixt L, Granz M, Landgraf T: When explanations lie: why many modified BP attributions fail. In: Proceedings of the 37th International Conference on Machine Learning; Proceedings of Machine Learning Research: Edited by Hal D, III, Aarti S. PMLR 2020: 9046--9057.
45. Samek W, Binder A, Montavon G, Lapuschkin S, Muller K-R. Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst. 2017;28(11):2660–73.
46. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Red Hook: Curran Associates Inc; 2018. p. 9525–36.
47. Tolonen A, Rhodius-Meester HFM, Bruun M, Koikkalainen J, Barkhof F, Lemstra AW, et al. Data-driven differential diagnosis of dementia using multiclass disease state index classifier. Front Aging Neurosci. 2018;10.
48. Bruun M, Frederiksen KS, Rhodius-Meester HFM, Baroni M, Gjerum L, Koikkalainen J, et al. Impact of a clinical decision support tool on prediction of progression in early-stage dementia: a prospective validation study. Alzheimers Res Ther. 2019;11(1).
49. Candemir S, Nguyen XV, Prevedello LM, Bigelow MT, White RD, Erdal BS. Neuroimaging Initiative AsD: predicting rate of cognitive decline at baseline using a deep neural network with multidata analysis. J Med Imaging. 2020;7(04).
50. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2577-2586.
51. Zhang Z, Xie Y, Xing F, McGough M, Yang L: MDNet: a semantically and visually interpretable medical image diagnosis network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017: 6428–6436.
52. Lucieri A, Bajwa MN, Braun SA, Malik MI, Dengel A, Ahmed S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: International Joint Conference on Neural Networks International Joint Conference on Neural Networks (IJCNN-2020), July 19-24, Glasgow, United Kingdom. IEEE; 2020.
53. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) [https://www.fda.gov/media/122535/download]
54. Ethics Guidelines for Trustworthy AI [https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419]

Publisher's Note

Supplementary material

Supplementary Table 1 Group separation performance for hippocampus volume and the convolutional neural network models for residualized data (extended).

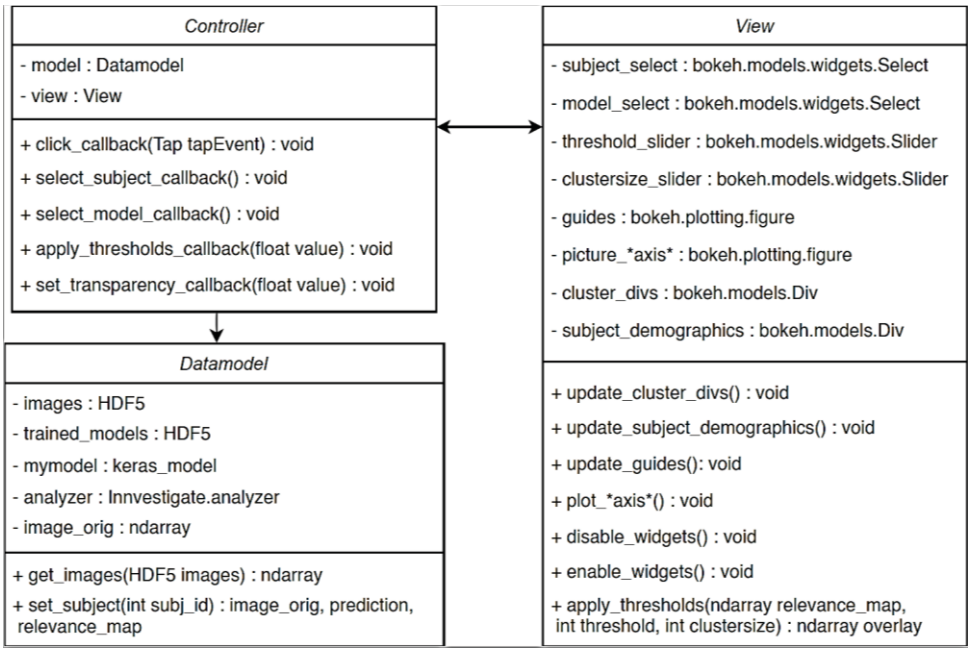
Sample	Hippocampus volume (residuals)		3D convolutional neural network						
	Balanced accuracy (mean ± SD)	AUC	Balanced accuracy (mean ± SD)	Sensitivity (mean ± SD)	Specificity (mean ± SD)	F1-score (mean ± SD)	Positive predictive value (mean ± SD)	Negative predictive value (mean ± SD)	AUC (mean ± SD)
ADNI-GO/2									
MCI vs. CN	(70.0 % ± 6.8 %)	(0.773 ± 0.091)	*(74.5 % ± 6.2 %)*	*(0.655 ± 0.108)*	*(0.836 ± 0.081)*	*(0.707 ± 0.080)*	*(0.781 ± 0.095)*	*(0.741 ± 0.063)*	*(0.785 ± 0.078)*
AD vs. CN	(84.4 % ± 3.6 %)	(0.945 ± 0.024)	*(88.9 % ± 5.3 %)*	*(0.942 ± 0.053)*	*(0.836 ± 0.081)*	*(0.872 ± 0.061)*	*(0.815 ± 0.084)*	*(0.952 ± 0.043)*	*(0.949 ± 0.029)*
MCI^+ vs. CN^-	(75.6 % ± 7.1 %)	(0.831 ± 0.080)	*(86.7 % ± 10.3 %)*	*(0.790 ± 0.173)*	*(0.943 ± 0.042)*	*(0.843 ± 0.127)*	*(0.916 ± 0.067)*	*(0.850 ± 0.116)*	*(0.925 ± 0.071)*
AD^+ vs. CN^-	(86.2 % ± 4.2 %)	(0.954 ± 0.025)	*(94.9 % ± 3.8 %)*	*(0.956 ± 0.038)*	*(0.943 ± 0.042)*	*(0.941 ± 0.043)*	*(0.927 ± 0.051)*	*(0.966 ± 0.029)*	*(0.985 ± 0.017)*
ADNI-3									
MCI vs. CN	62.8 % (63.1 % ± 1.4 %)	0.683	63.1 % (63.6 % ± 1.5 %)	0.421 (0.496 ± 0.082)	0.850 (0.775 ± 0.080)	0.492 (0.523 ± 0.031)	0.611 (0.570 ± 0.060)	0.716 (0.730 ± 0.015)	0.684 (0.677 ± 0.020)
AD vs. CN	83.4 % (83.4 % ± 0.4 %)	0.917	84.4 % (81.7 % ± 2.9 %)	0.839 (0.858 ± 0.036)	0.850 (0.775 ± 0.080)	0.638 (0.573 ± 0.068)	0.515 (0.438 ± 0.088)	0.965 (0.967 ± 0.006)	0.913 (0.899 ± 0.013)
MCI^+ vs. CN^-	69.1 % (69.2 % ± 2.7 %)	0.791	69.8 % (68.3 % ± 4.4 %)	0.556 (0.615 ± 0.124)	0.840 (0.752 ± 0.086)	0.556 (0.523 ± 0.031)	0.556 (0.479 ± 0.057)	0.840 (0.847 ± 0.031)	0.810 (0.742 ± 0.024)
AD^+ vs. CN^-	83.6 % (82.0 % ± 1.8 %)	0.882	80.2 % (75.5 % ± 4.2 %)	0.765 (0.759 ± 0.043)	0.840 (0.752 ± 0.086)	0.619 (0.573 ± 0.068)	0.520 (0.424 ± 0.080)	0.940 (0.932 ± 0.012)	0.830 (0.828 ± 0.028)
AIBL									
MCI vs. CN	67.4 % (67.6 % ± 0.5 %)	0.741	68.2 % (67.3 % ± 2.7 %)	0.552 (0.596 ± 0.111)	0.812 (0.749 ± 0.086)	0.455 (0.437 ± 0.020)	0.387 (0.351 ± 0.057)	0.894 (0.898 ± 0.016)	0.763 (0.749 ± 0.012)
AD vs. CN	84.1 % (85.3 % ± 1.5 %)	0.927	85.0 % (82.3 % ± 3.0 %)	0.887 (0.897 ± 0.051)	0.812 (0.749 ± 0.086)	0.547 (0.523 ± 0.055)	0.396 (0.350± 0.090)	0.981 (0.982 ± 0.007)	0.950 (0.926 ± 0.007)
MCI^+ vs. CN^-	78.5 % (78.8 % ± 0.9 %)	0.874	75.4 % (73.6 % ± 3.1 %)	0.685 (0.713 ± 0.095)	0.823 (0.759 ± 0.089)	0.503 (0.464 ± 0.051)	0.398 (0.356 ± 0.082)	0.939 (0.940 ± 0.013)	0.828 (0.814 ± 0.022)
AD^+ vs. CN^-	87.2 % (89.1 % ± 2.4 %)	0.976	88.3 % (85.3 % ± 3.3 %)	0.943 (0.947 ± 0.048)	0.823 (0.759 ± 0.089)	0.629 (0.573 ± 0.085)	0.472 (0.420 ± 0.104)	0.989 (0.989 ± 0.008)	0.978 (0.958 ± 0.011)
DELCODE									
MCI vs. CN	69.0 % (69.0 % ± 9.6 %)	0.774	71.0 % (69.7 % ± 2.6 %)	0.652 (0.724 ± 0.048)	0.767 (0.670 ± 0.084)	0.660 (0.664 ± 0.017)	0.669 (0.618 ± 0.051)	0.753 (0.771 ± 0.009)	0.775 (0.772 ± 0.017)
AD vs. CN	88.4 % (86.4 % ± 3.0 %)	0.943	85.5 % (80.5 % ± 4.0 %)	0.942 (0.939 ± 0.017)	0.767 (0.670 ± 0.084)	0.778 (0.719 ± 0.046)	0.662 (0.585 ± 0.062)	0.965 (0.958 ± 0.011)	0.953 (0.938 ± 0.013)
MCI^+ vs. CN^-	77.4 % (77.8 % ± 0.7 %)	0.867	72.2 % (74.9 % ± 3.5 %)	0.737 (0.809 ± 0.046)	0.707 (0.690 ± 0.085)	0.724 (0.762 ± 0.027)	0.712 (0.723 ± 0.049)	0.732 (0.787 ± 0.031)	0.840 (0.830 ± 0.017)
AD^+ vs. CN^-	88.2 % (87.6 % ± 1.8 %)	0.954	83.3 % (82.2 % ± 4.0 %)	0.959 (0.955 ± 0.021)	0.707 (0.690 ± 0.085)	0.832 (0.824 ± 0.033)	0.734 (0.726 ± 0.053)	0.953 (0.949 ± 0.023)	0.968 (0.956 ± 0.012)

Reported values are for the single model trained on the whole ADNI-GO/2 dataset. In parenthesis, the mean values and standard deviation for the ten models trained in the tenfold cross-validation procedure are provided to indicate the variability of the measures. Values for the ADNI-GO/2 sample (in italics) may be biased as the respective test subsamples were used to determine the optimal model during training. We still report them for better comparison of the model performance across samples.

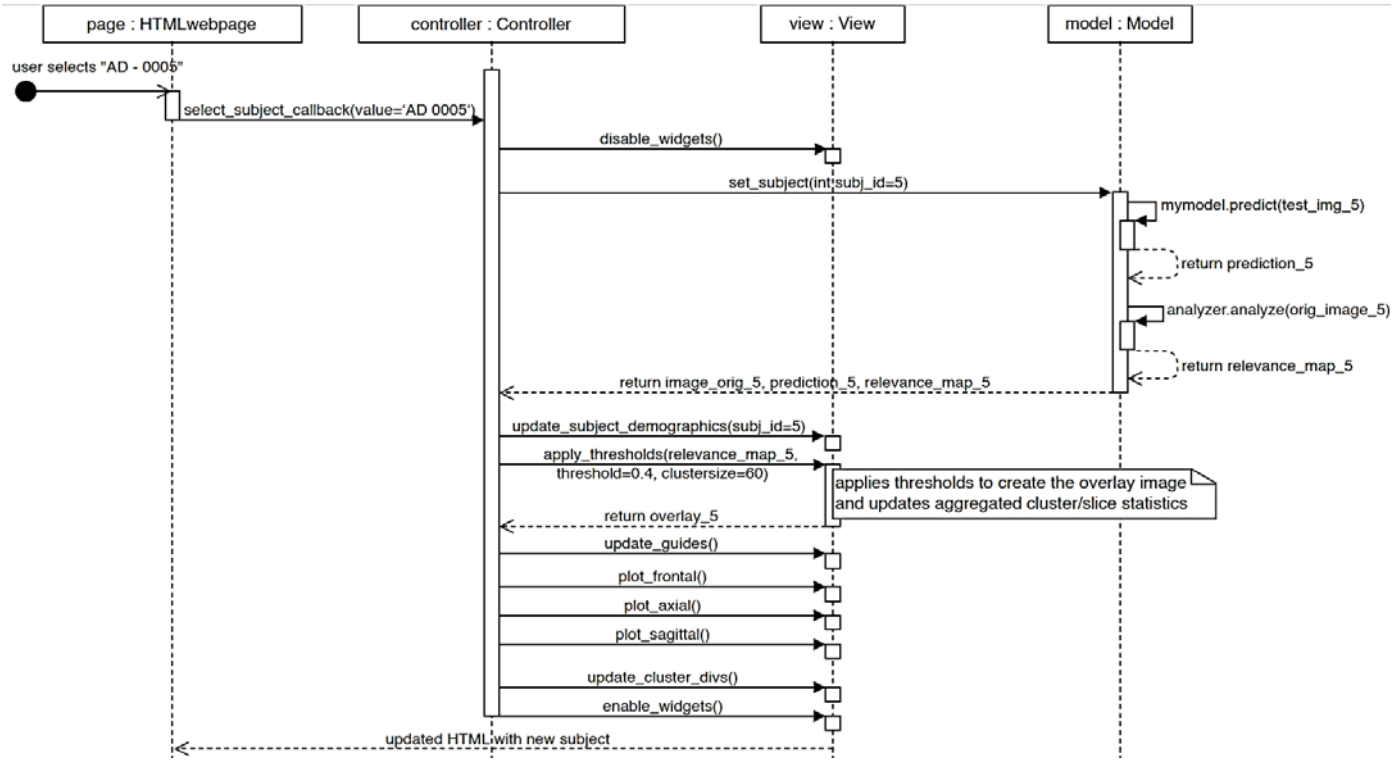
Supplementary Table 2 Group separation performance for hippocampus volume and the convolutional neural network models for raw input data.

Sample	3D convolutional neural network	
	Balanced accuracy (mean ± SD)	**AUC (mean ± SD)**
ADNI-GO/2		
MCI vs. CN	*(71.1 % ± 5.7 %)*	*(0.731 ± 0.070)*
AD vs. CN	*(84.6 % ± 6.5 %)*	*(0.921 ± 0.024)*
MCI^+ vs. CN^-	*(80.7 % ± 7.9 %)*	*(0.881 ± 0.069)*
AD^+ vs. CN^-	*(92.4 % ± 3.9 %)*	*(0.974 ± 0.015)*
ADNI-3		
MCI vs. CN	62.6 % (60.7 % ± 2.6 %)	0.629 (0.626 ± 0.017)
AD vs. CN	86.1 % (82.1 % ± 5.8 %)	0.919 (0.907 ± 0.028)
MCI^+ vs. CN^-	71.8 % (70.6 % ± 4.9 %)	0.769 (0.745 ± 0.021)
AD^+ vs. CN^-	82.2 % (78.8 % ± 5.2 %)	0.873 (0.877 ± 0.026)
AIBL		
MCI vs. CN	69.1 % (64.8 % ± 3.2 %)	0.735 (0.713 ± 0.016)
AD vs. CN	83.7 % (80.2 % ± 6.3 %)	0.922 (0.924 ± 0.006)
MCI^+ vs. CN^-	78.0 % (73.3 % ± 4.5 %)	0.837 (0.817 ± 0.025)
AD^+ vs. CN^-	86.3 % (83.7 % ± 6.8 %)	0.959 (0.959 ± 0.007)
DELCODE		
MCI vs. CN	69.8 % (69.0 % ± 2.4 %)	0.779 (0.761 ± 0.017)
AD vs. CN	89.8 % (83.5 % ± 6.0 %)	0.947 (0.937 ± 0.023)
MCI^+ vs. CN^-	72.5 % (72.5 % ± 5.9 %)	0.853 (0.814 ± 0.049)
AD^+ vs. CN^-	92.5 % (86.0 % ± 7.1 %)	0.982 (0.967 ± 0.028)

Reported values are the respective measures for the single model trained on the whole ADNI-GO/2 dataset. In parenthesis, the mean values and standard deviation for the ten models trained in the tenfold cross-validation procedure are provided to indicate the variability of the measures. Values for the ADNI-GO/2 sample (in italics) may be biased as the respective test subsamples were used to determine the optimal model during training. We still report them for better comparison of the model performance across samples.

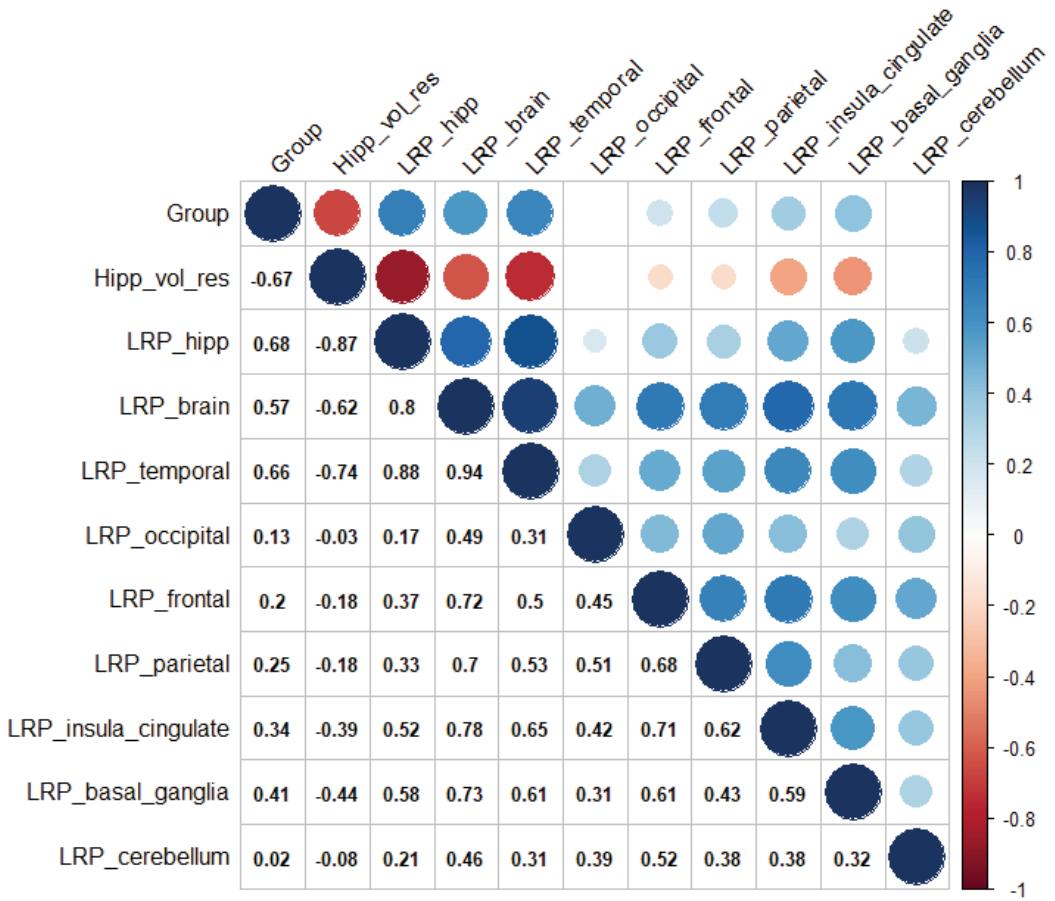


Supplementary Figure 1 UML diagram of the interactive visualization application.



Supplementary Figure 2 Sequence diagram of function calls when selecting a new person.

Supplementary Figure 3 Comparison of mean relevance maps between samples. Left: ADNI-GO/2, Right: DELCODE.



Supplementary Figure 4 Correlation matrix of hippocampus volume (residualized) and several brain regions' relevance scores for DELCODE participants and the model trained on the whole ADNI-GO/2 dataset. The correlation between hippocampus volume and hippocampus relevance was highest (–0.87). Upper right triangle entries were thresholded a $p<0.001$. For simplicity, group was numerically encoded as CN=1, MCI=2, AD=3.

Comparison of CNN Architectures for Detecting Alzheimer's Disease using Relevance Maps

Devesh Singh, Martin Dyrba

Deutsches Zentrum für Neurodegnerative Erkrankungen (DZNE), Rostock, Germany
devesh.singh@dzne.de

Abstract. Alzheimer's disease (AD) is a neurodegenerative disorder which can be detected using T1-weighted MRI scans. Recent developments in convolutional neural networks (CNN) achieved promising results in various image classification tasks. Specifically, four CNN architectures are widely used: AlexNet, VGG, ResNet, and DenseNet. Feature attribution methods such as layer-wise relevance propagation allow to trace back the information flow in CNNs to derive relevance heatmaps, which approximate the contribution of the input image regions on the model decision. We addressed the open question, which of these CNN architectures is best suited for medical image detection, i.e. AD classification based on MRI data. We adapted the CNN architectures to be used with 3D brain MRI data and trained the models on a heterogeneous dataset with N>2200 from four large studies. We applied tenfold cross-validation and additionally validated results in an independent test dataset. DenseNet and ResNet provided best results, although the overall differences in accuracy did not reach statistical significance. DenseNet provided the most focused relevance maps best matching a-priori expectations of brain regions contributing to the detection of AD, i.e. atrophy in medial temporal lobe.

1 Introduction

In 2020, it was reported that approximately 1.8 million people suffer from dementia in Germany. The most common reason for dementia is Alzheimer's disease (AD), which accounts for almost 60-70% of all the dementia cases. The pathophysiological development of AD begins almost a decade before the first symptoms of the disease appear. AD is a neuro-degenerative disease, which causes gradual and irreversible damage to the brain. Specifically, this damage includes accumulation of the protein beta-amyloid (also called plaques), deaths of neurons caused by neurofibrillary tangles, and gray matter volume reduction (atrophy) in hippocampus, medial temporal lobe, and later-on more widespread cortical areas. These reductions in brain volume are visible in T1 weighted MRI scans, already in early stages of AD [1]. AD causes behavioural changes like memory loss, deteriorating orientation and executive functioning, deregulated emotions, motor control loss and speech impairment. Due to the gradual nature of AD, it is difficult to clinically diagnose its early stages. The first symptomatic stage of AD is called mild cognitive impairment (MCI). It is reported that approximately 10%-15% of the people with MCI will convert into dementia each year[1].

[1] https://alz.org/facts

T. M. Deserno et al. (Hrsg.), *Bildverarbeitung für die Medizin 2023*,
Informatik aktuell, https://doi.org/10.1007/978-3-658-41657-7_51

State of the art research shows that deep convolutional neural networks (CNN) have been successfully applied for detecting AD using brain MRI scans [2–6]. Notably, CNNs have solved a variety of machine vision problems like image classification, image segmentation and object detection, within and outside the medical domain. Some of the most widely applied CNN architectures are: AlexNet, VGG, ResNet, Inception-net and DenseNet. Though, due to the black box nature of CNNs, it is difficult to include them in clinical decision support systems, as they lack transparency and comprehensibility. To address this issue, attribution methods have been proposed to produce relevance maps highlighting key features from input samples which contributed to a particular model decision. Few studies systematically compared attribution methods for CNN models for AD classification [4–6], proposing the layer-wise relevance propagation (LRP) or Integrated Gradients methods as most useful.

Notably, to our knowledge, the studies mentioned above compared attribution methods for a single CNN model architecture, but little is known about the influence of the different CNN architectures on the derived relevance maps. In this study we address this gap by i) the comparison of four different CNN architectures with respect to model performance, and ii) the face to face comparison of derived relevance maps.

2 Materials and methods

In our study, T1-weighted volumetric MRI scans were obtained from five study sources: The Alzheimer's disease neuroimaging initiative (ADNI)[2] study phases ADNI2 and ADNI3, the Australian imaging, biomarker & lifestyle Flagship Study of Ageing (AIBL)[3], the DZNE Longitudinal Study on cognitive ampairment and dementia (DELCODE)[4], and the European DTI study on dementia (EDSD)[5]. Data obtained from ADNI2, AIBL, DELCODE, and EDSD was used as training data and ADNI3 was used as independent test data. Sample characteristics are listed in Tab. 1. The image preparation pipeline included the N4ITK bias field correction and the SyN algorithm from ANTs to perform an affine registration of each scan to the MNI space. The brain scans were segmented into the compartments gray and white matter, and cerebrospinal fluid using ANTs Atropos. Only the normalized gray matter segments were used as model input. Finally, each gray matter map was cropped to the size of 169 × 208 × 179, with 1 mm isotropic voxel size. Notably, only one (the first) MRI scan from each participant was considered in our study, in case scans from multiple timepoints were available.

Four different convolutional neural network (CNN) architectures were tested in this study, which were chosen because of their successful application across various machine vision problems: AlexNet, VGG, ResNet and DenseNet (Fig. 1). The shallower and simpler network AlexNet was successfully applied before in [3]. Recent studies additionally applied more sophisticated model architectures such VGG [2], ResNet, and DenseNet. We hypothesized that more complex CNNs - ResNet and DenseNet, which use skip connections, should perform better than other simpler CNNs - VGG and

[2]More information about the ADNI can be found on https://adni.loni.usc.edu/

[3]https://aibl.csiro.au/ for further details

[4]https://www.dzne.de/en/research/studies/clinical-studies/delcode/

[5]https://www.gaaindata.org/partner/EDSD

Tab. 1. Sample statistics per diagnosis state. The training dataset is an aggregation of ADNI2, AIBL, DELCODE, and EDSD study datasets, while independent ADNI3 study data is used for testing. CN: a cognitively normal, MCI: mild cognitive impairment, AD: dementia due to Alzheimer's disease, MMSE: mini-mental state examination score, F: female, M: male.

	CN	MCI	AD
Training dataset N	1109	640	487
Age (SD)	71.8 ± 6.6	73.0 ± 7.3	74.1 ± 7.7
MMSE (SD)	29.0 ± 1.2	27.3 ± 2.1	22.0 ± 4.3
Education (SD)	14.9 ± 3.3	14.5 ± 3.5	13.4 ± 3.9
Sex (F/M)	610 / 499	287 / 353	253 / 234
Testing dataset N	325	185	62
Age (SD)	70.2 ± 6.4	72.2 ± 7.5	74.8 ± 7.7
MMSE (SD)	29.1 ± 1.1	27.8 ± 2.0	23.1 ± 3.3
Education (SD)	16.6 ± 2.2	16.6 ± 2.5	16.5 ± 2.4
Sex (F/M)	210 / 115	84 / 101	27 / 35

Alexnet, which use linear feature processing steps. Our implementation of these models is available via GitHub[6].

To train these models, a ten-fold cross-validation approach was used, which allowed unbiased comparison of these model architectures. The samples aggregated from ADNI2, AIBL, DELCODE, and EDSD were used as training dataset and ADNI3 study data was used as independent test dataset. We hypothesized that learning from a mixed cohort of pooled mdatasets should help the model learn more general features. Below we report two model performance estimates, first, on the left out validation sets of the cross-validation and second, on the independent test set (ADNI3).

The models were trained with a binary classification target, where Alzheimer's dementia (AD) patients and patients with amnestic mild cognitive impairment (MCI) were merged into one disease-positive class, which was compared to the cognitively normal (CN) participants as the control class. Categorical cross-entropy was chosen as the loss function. An early stopping regularization method was applied, monitoring the training set accuracy as performance metric over epochs, to reduce model over-fitting. The models were optimised using Adam optimiser with the default parameter settings.

It has been previously shown that feature attribution visualisation methods can improve clinical users' understanding of a model's prediction. Specifically, relevance propagation methods help in explaining the neural network's prediction by propagating activation backwards and eventually highlighting features from the input sample as heatmaps or relevance maps. In the context of AD, Böhle et al. [4] showed that the layer-wise relevance propagation (LRP) method highlights neuro-anatomically specific and stable input features and should be preferred over other relevance propagation methods like guided backpropagation (GB). Here, we also applied the compositional $\text{LRP}_{\alpha=1,\beta=0}$ rule as shown by Dyrba et al. [3] to produce clinically valuable relevance maps. To choose signal over noise while visualising the relevance maps, we first re-scaled the relevance intensities using their $q = 0.9999$ quantile value and clipped values between

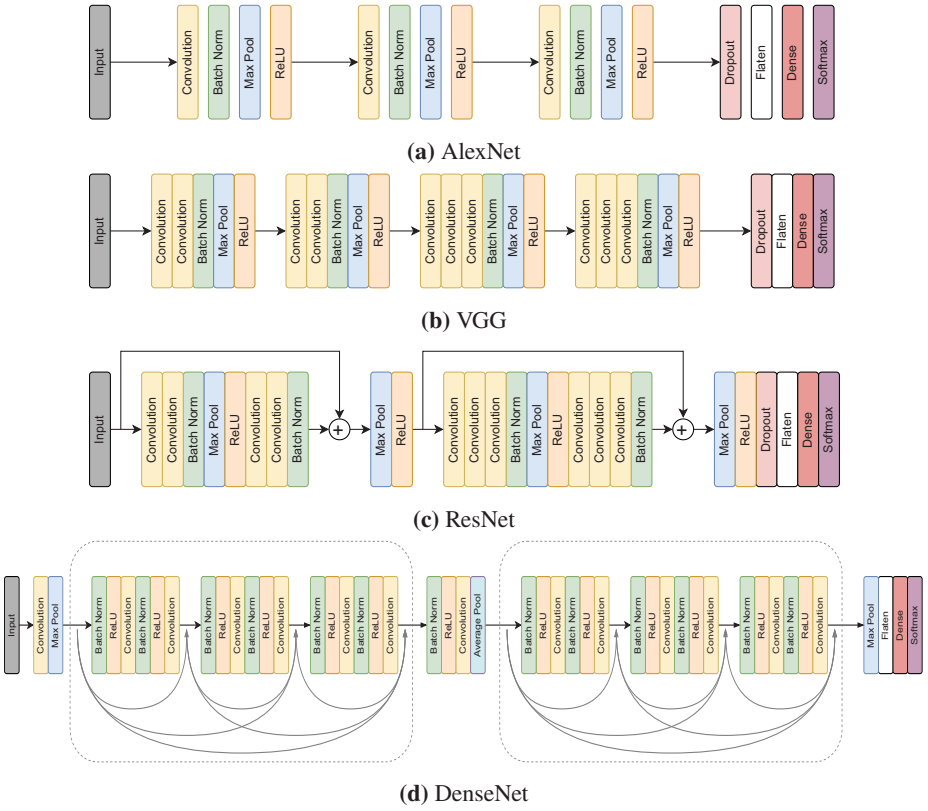
[6]Source code is available via GitHub. URL: `https://github.com/martindyrba/VisualAiD`

the range $[-1, 1]$. Then, the relevance values were smoothed using a Gaussian blur with a kernel size of 6 mm. We used Keras/Tenserflow 1.15 for our implementation of the CNN models and the iNNvestigate[7] library for deriving the relevance maps.

3 Results

Tab. 2 shows the mean accuracy obtained from the various CNN models. ResNet and DenseNet achieved the highest mean accuracy levels on the tasks of AD vs CN and MCI vs CN, respectively.

The mean relevance maps of the AD dementia patients group and MCI patients were very similar to each other, both in the training sample as well as in the independent test dataset. Comparing different CNN architectures, the mean relevance maps for the MCI group of the ADNI3 test dataset were plotted (Fig. 2). These maps were calculated only for a single CNN model of each architecture, namely the second cross-validation iteration which had the highest average accuracy among all ten iterations.



(a) AlexNet

(b) VGG

(c) ResNet

(d) DenseNet

Fig. 1. CNN model architectures used for classifying dementia.

[7] https://github.com/albermax/innvestigate

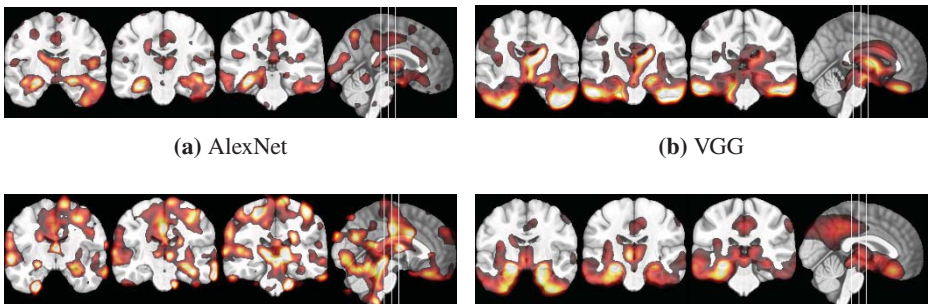
Tab. 2. Mean and standard deviation of accuracy obtained for the cross-validation dataset (top) and independent ADNI3 test data (bottom). AD: Alzheimer's dementia, CN: a cognitively normal state, MCI: mild cognitive impairment.

	AlexNet	VGG	ResNet	DenseNet
Validation-set accuracy				
AD vs CN	0.68 ± 0.12	0.70 ± 0.04	0.75 ± 0.09	0.71 ± 0.05
MCI vs CN	0.63 ± 0.08	0.64 ± 0.03	0.67 ± 0.06	0.65 ± 0.05
Independent test-set accuracy				
AD vs CN	0.72 ± 0.18	0.76 ± 0.08	0.81 ± 0.14	0.76 ± 0.07
MCI vs CN	0.64 ± 0.08	0.66 ± 0.03	0.66 ± 0.05	0.67 ± 0.03

4 Discussion

The mean accuracy levels revealed a consistently higher test accuracy than the validation accuracy (Tab. 2). This suggests that separating the classes for the pooled and more heterogeneous study datasets is more challenging for the models than performing this task on a relatively homogeneous prospective cohort such as ADNI3. This is in agreement with our hypothesis that training on pooled heterogeneous datasets from different study cohorts is beneficial and improves robustness of the models. Though these accuracy levels are lower than those found in AD literature [5, 6], which could be attributed to combining AD dementia and MCI groups into a single disease-positive group during training, which possibly needs another model fine-tuning iteration to optimize different thresholds to better demarcate CN, MCI, and AD.

The accuracy levels also suggest that the group separation task of AD vs CN was simpler than MCI vs CN (Tab. 2), this corroborates with other AD literature findings. Upon comparing the different CNN architectures with each other, we found that complex



(a) AlexNet **(b)** VGG

(c) ResNet **(d)** DenseNet

Fig. 2. Mean relevance maps for the MCI group of the ADNI3 dataset obtained using the $\text{LRP}_{\alpha=1,\beta=0}$ relevance propagation method overlaid on the MNI brain template. Coronal slices show Y=[-10,-20,-30] mm in MNI reference space are shown. Bright yellow represent the most relevant regions while the dark red regions are of lower relevance. Relevance maps were created following proportional scaling of the activations.

architectures with skip connections - ResNet and DenseNet, performed slightly better than the two other models. This is in agreement with our other hypothesis that more complex models perform better. Notably, the differences in performance did not reach statistical significance overall.

From the relevance maps derived from all the models (Fig. 2), we see that DenseNet mainly focuses on medial temporal lobe and posterior cingulate cortex, also highlighted in [1]. The relevance map for the ResNet model is more noisy and heterogeneous, suggesting that the model's relative high test accuracy for AD vs CN is achieved by considering more widespread brain atrophy or potential artifacts. In contrast to AlexNet and ResNet, the DenseNet model seems to be less sensitive to noise. Its dense connections between layers enabled a highly efficient information flow at various scales.

In future, we will test the CNN models on each individual study dataset (AIBL, EDSD etc.) to replicate previous studies, and investigate the gap between our reported accuracy levels and other studies' accuracy levels. It will also be of interest to examine relevance maps of individuals and the gain of relevance values for specific brain regions.

In this study, we showed that having a pooled training dataset from different study cohorts is beneficial. We also showed that DenseNet utilised an efficient information flow at various scales to generate relevance maps which focused on clinically relevant features. While ResNet derived its high accuracy levels from learning heterogenerous patterns or potential artifacts. This demonstrated the added value of a holistic evaluation of models where relevance maps are being used in combination with classical performance metrics like accuracy, F1-score or ROC-AUC.

References

1. Bernard C, Helmer C, Dilharreguy B, et al. Time course of brain volume changes in the preclinical phase of Alzheimer's disease. Alzheimers Dement. 2014;10(2):143–51.
2. Dyrba M, Hanzig M, Altenstein S, et al. Improving 3D CNN comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. Alzheimers Res Ther. 2021:1–18.
3. Dyrba M, Pallath AH, Marzban EN. Comparison of CNN visualization methods to aid model interpretability for detecting Alzheimer's disease. Proc BVM. 2020:307–12.
4. Böhle M, Eitel F, Weygandt M, et al. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
5. Qiu S, Joshi PS, Miller MI, et al. Development and validation of an interpretable deep learning framework for Azheimer's disease classification. Brain. 2020;143(6):1920–33.
6. Wang D, Honnorat N, Fox PT, et al. Deep neural network heatmaps capture Alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. Neuroimage. 2023:119929.

Contrastive Self-supervised Learning for Neurodegenerative Disorder Classification

Vadym Gryshchuk[a,*], Stefan Teipel[a,b], Martin Dyrba[a,**], and for the ADNI, AIBL, FTLDNI study groups

[a]*Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Rostock, Germany*
[b]*Department of Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany*

Abstract

Background and Objective: Neurodegenerative disorders such as Alzheimer's disease or frontotemporal lobar degeneration are characterized by specific brain volume loss, which can be assessed in-vivo using T1-weigted magnetic resonance imaging (MRI). Most machine learning approaches applied to the classification of neurodegenerative disorders using MRI scans rely on supervised methods, that means, they typically require for each training sample the corresponding diagnosis label which can be burdensome or difficult to obtain for large numbers of data. Self-supervised learning (SSL) provides a new perspective on training machine learning models without sample labels. We investigated the application of contrastive learning as a type of SSL and, more specifically, if this methodology can be used to reliably distinguish various neurodegenerative disorders from each other.

Methods: We propose a model that is composed of two parts, a feature extraction and a classification component. A deep convolutional neural network trained in a contrastive self-supervised way is utilized as the feature extractor. It learns a latent high-dimensional representation of the given data. The classification component is a simple neural network, here a single layer perceptron, that performs the classification of neurodegenerative disorders based on the latent representation vectors.

Results: Our experiments showed that the feature extractor trained in a self-supervised way could provide generalizable representations for the clas-

*Corresponding authors: vadym.gryshchuk@protonmail.com,
**martin.dyrba@dzne.de

Preprint submitted to Computer Methods and Programs in Biomedicine January 10, 2023

sification component. This component achieved a balanced accuracy for Alzheimer's dementia vs. cognitively normal controls of 0.81 for the holdout test subset and 0.79 for an independent validation dataset. Feature attribution heatmaps obtained by the Integrated Gradient method highlighted widespread contributions of the gray and white matter tissue compartments to the classification.
Conclusions: The balanced accuracy of our model was comparable to the accuracy of other state-of-the-art supervised deep learning approaches. Our findings suggest that the proposed contrastive SSL methodology can successfully make use of unannotated neuroimaging datasets as training data.

Keywords: contrastive learning, self-supervised learning, neurodegenerative disorders, deep learning, structural magnetic resonance imaging

1. Introduction

Neurodegenerative disorders such as Alzheimer's disease (AD) and frontotemporal dementia (FTD) are characterized by specific brain volume loss, which can be assessed in-vivo using structural magnetic resonance imaging (MRI). The usual radiological evaluation of MRI scans is mainly based on visual examination and is very time-consuming. Assistance systems for the automated detection of disease-specific patterns could be very useful for better clinical diagnosis, as they can significantly decrease the evaluation time for the radiologists and neurologists.

Convolutional neural networks (CNNs) act as a powerful method for the automated identification of neurodegenerative diseases from MRI scans and achieve state-of-the-art results in image classification tasks. Due to recent technological advances and novel software frameworks we have witnessed an emergence of different CNN architectures - VGG, ResNet, Inception, DenseNet, EfficientNet are probably the most notable ones. Such rapid developments had a great impact on the neuroimaging community that is interested in automatic feature learning. Thus, various methodologies relying on the use of CNNs were developed [1, 2, 3, 4, 5].

Recently, advances in the natural language processing domain have influenced the design of new convolutional networks. The architectures based on Transfomers like Vision Transformers [6, 7] set a new direction in the imaging community. Currently, CNN architectures have been primarily trained in a *supervised* way by using an external label which indicates whether a certain

scan belongs to a particular group or class. Generating such labels for a large number of data is often burdensome and costly. Furthermore, CNN models require much data to achieve appropriate results. Such large data is not always easily available in the medical domain due to the high cost of data collection and the associated privacy regulation.

Such constraints led us to reconsider the use of CNN models trained in a *supervised* way and adhere to *self-supervised learning (SSL)*. The SSL methods learn without any sample labels by utilizing the internal structure of the data to acquire representative features. Architectures trained in a self-supervised way are more biologically plausible, provide extensive feature space, and compete with supervised approaches [8]. We utilized a contrastive SSL method that allows learning of general features from data by contrasting similar and different data samples. Furthermore, we hypothesized that SSL methods could learn meaningful salient structural representations better than supervised approaches, thus increasing interpretability.

The main goal of the current study was to develop an approach for learning general features from structural MRI data, thus enabling better generalization and interpretability. Therefore, in this paper, we propose to train a CNN model on structural MRI data without any sample labels and use the trained CNN model as a feature extractor for another linear model trained in a supervised way from the provided features that predicts a specific type of a neurodegenerative disorder. The main research question was defined as: *Does contrastive self-supervised learning outperform supervised learning in terms of predictive power?*

Our main contributions are: i) An architecture for the prediction of neurodegenerative disorders that uses contrastive self-supervised learning. ii) A comparison between models trained on structural MRI data using self-supervised and supervised approaches in terms of classification power.

1.1. Related Work

Most self-supervised approaches applied to MRI data deal with reconstruction or segmentation [9, 10]. To the best of our knowledge, there are no approaches for direct comparison, though the methods that apply contrastive learning methods to MRI data exist and showed promising results [9]. Probably the most comparable work to ours was done by Taleb et al. [9]. In the following, we review two self-supervised methods evaluated by Taleb et al. [9] that are most similar to the methods that we used:

3D Contrastive Predictive Coding (3D-CPC): The 3D-CPC method [9] uses the contrastive learning technique that predicts latent representations of data. 3D-CPC operates on 3D data by splitting it into patches. The InfoNCE loss is used to optimize the negative log probability of classifying a positive patch's representation among a set of negative patches' representations correctly. To produce positive and negative representations, 3D-CPC utilizes encoder and context models. An encoder model maps an input patch to latent space and a context model summarizes latent representations in the form of a context vector that captures high-level information about input patches. Thus, a context vector can be used to predict latent representations of the next patches. Therefore, a positive sample is the predicted representation of a patch, and negative samples are representations of patches taken randomly from other locations in 3D data.

3D Exemplar networks (3D-Exe): The 3D-Exe method [9] is based on the Exemplar network and derives positive and negative representations via augmentation techniques. A positive sample is an augmented sample, while a negative sample is a different sample. The model uses a triplet loss, which pulls representations of negative samples apart and representations of positive samples in the embedding space together.

Though Taleb et al. [9] provided an extensive evaluation of different self-supervised approaches, they did not compare the achieved results with supervised approaches in terms of sample classification in the target domain.

Another methodology for self-supervised learning from MRI data was proposed by Hu et al. [10] that is based on parallel training of two networks with the objective of minimizing the reconstruction loss. The authors achieved competitive results with supervised approaches. However, their method was evaluated only for image reconstruction, and thus is not comparable with our approach that is based on the detection of a neurodegenerative disorder.

Other recent publications such as [11, 12, 13, 14] applied SSL to longitudinal AD MRI datasets in order to i) study methods for combining information from multiple imaging modalities, or ii) to predict the trajectories of cognitive performance and/or cognitive decline.

In our work, however, we used the methods that were successfully applied for images produced by conventional shutter cameras and were not previously applied to 2D MRI scans. Additionally, we tried to interpret a model's prediction through attribution methods which are a popular technique for the visualization of input regions that are relevant to the prediction [15, 16].

2. Methods

2.1. ConvNeXt – A modern convolutional neural network

Convolutional neural networks have been dominating computer vision for a long time. However, Transformer neural networks showing state-of-the-art results in the natural language processing domain started to influence the design of CNNs. Transformers take advantage of the hierarchical structure of input sequences and so-called multi-head self-attention mechanisms to achieve superior results. Thus, Transformers have been adopted as de facto standard for building the most powerful language models. Their success inspired researchers in the development of modern CNNs. The designers of the ConvNeXt model [17] attempted to create a CNN layout using experience gained from the building blocks of Transformers. With ConvNeXt, Lui et al. [17] achieved state-of-the-art results for image classification based on the ImageNet dataset, providing higher accuracy than competing Vision Transformer models while being computationally more efficient. In the present study, we used and adapted the ConvNeXt model as feature extractor for the contrastive SSL approach. In the following, we briefly describe the main architectural principles of the ConvNeXt network:

Architectural block: The widespread adoption of consecutive stacking of i) a convolution operation, ii) an activation function, and iii) a pooling function led to the incorporation of separate "blocks". Each block has the same sequential operations, but their scaling properties with respect to the number of input and output channels differ. The ConvNeXt model layout is illustrated in the left part of Figure 1. We used the "tiny" ConvNeXt model [17] that has the number of repeated sequential blocks set to $(3, 3, 9, 3)$ with the number of output channels equalling to $(96, 192, 384, 768)$.

Inverted bottleneck: The ResNet architecture uses residual blocks called bottlenecks. These bottlenecks compress the number of input channels. Since ResNet makes use of skip-connections to learn a residual function that references the inputs of previous layers, those channels are then expanded to match the input. ConvNeXt utilizes an inverted residual bottleneck by expanding the number of channels, thus reducing the number of network parameters.

Self-attention: Conventional convolution operation is a function of all input channels. ConvNeXt relies on depthwise convolution that operates per channel, thus enabling the separation of the channel and spatial information.

Kernel size, normalization, and activation function: ConvNeXt uses a convolution kernel size of 7×7. A layer normalisation is used instead of batch normalization which is widely popularized in traditional CNNs. Additionally, ConvNeXt uses the Gaussian Error Linear Unit (GELU) as an activation function. For more details refer to the ConvNeXt paper [17].

2.2. Contrastive self-supervised learning

To train ConvNeXt without using actual sample labels, we used contrastive SSL which has shown promising results in learning representative features [18]. Contrastive learning is based on the use of positive and negative data pairs [19]. A *positive* pair (i, j) consists of two views of the same data sample that can be created by using augmentation techniques. For instance, refer to Figure 2 for examples of augmented images providing such positive pairs. A *negative* pair encompasses two different object samples. Contrastive learning utilizes a contrastive loss that tries to learn similar embeddings for positive pairs while pulling negative pairs apart in a latent space. Most SSL approaches for contrastive learning have an "encoder" and a "projection head" as illustrated in Figure 1. An encoder is a convolutional neural network that serves as a feature extractor. A projection head can be a small neural network, for example, a multilayer perceptron (MLP), that takes as input the output from the encoder. The contrastive loss ℓ operates directly on the output of a projection head $\mathbf{z}$ which serves as a mapping of CNN features to a latent space. The contrastive loss for a positive pair is defined as follows [19]:

$$\ell(i,j) = -log\frac{exp(sim(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k\neq 1}^{2N} \mathbb{1}_{[k\neq i]} exp(sim(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{1}$$

where τ is a scaling factor called temperature, $\mathbb{1}$ is an indicator function with output values being 0 or 1, N is the number of training samples, and $\text{sim}(\cdot)$ is a cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$, formally defined as [19]:

$$sim(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}. \tag{2}$$

The final loss for the batch of data results in [19]:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)], \tag{3}$$

where $2k-1$ and $2k$ represent the indices of the same augmented object.

We used the Nearest-Neighbor Contrastive Learning (NNCLR) method [20] as the main framework for learning representations from structural MRI data without any sample labels. NNCLR extends a common contrastive SSL framework by keeping a record of recent embeddings of augmented views in a queue Q. Thus, the pairs are not directly compared, but a projection embedding that is most similar to a view is selected from Q for the comparison with another view. The contrastive loss is defined as follows:

$$\ell(i,j) = -log\frac{exp(sim(\mathcal{S}(\mathbf{z}_i, Q), \mathbf{z}_j)/\tau)}{\sum_{k\neq 1}^{2N} \mathbb{1}_{[k\neq i]} exp(sim(\mathcal{S}(\mathbf{z}_i, Q), \mathbf{z}_k)/\tau)}, \tag{4}$$

where $S(\mathbf{z}, Q)$ is the nearest neighbour function defined as follows:

$$\mathcal{S}(\mathbf{z}, Q) = \arg\min_{\mathbf{q}\in Q} \|\mathbf{z} - \mathbf{q}\|_2 . \tag{5}$$

2.3. Study data of neurodegenerative disorders

We used T1-weighted brain MRI scans from publicly available neuroimaging repositories. Each of the neuroimaging initiatives obtained internal review board approvals and met all ethical standards in the collection of data. See the online documentation linked below for more detailed information on the study protocols and procedures. In the following, we describe the datasets:

ADNI[1]: The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a research study that provides clinical and imaging data for the investigation of Alzheimer's disease. The ADNI was launched in 2003 as a public-private partnership, led by principal investigator Michael W. Weiner, MD. The primary goal of ADNI is to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early dementia due to Alzheimer's disease (AD). We included only MRI scans from ADNI2 and ADNI3 studies. Participants with mild cognitive impairment (MCI), dementia due to Alzheimer's disease (AD), and cognitive normal (CN) states were considered. The total number of MRI scans was 2365 from 844 people for ADNI3, and 1836 from 743 people for ADNI2, respectively.

[1]ADNI: `https://adni.loni.usc.edu/`

AIBL[2]: The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) is a similar study to ADNI that focus on the early detection of Alzheimer's disease and possible lifestyle interventions. Data was collected by the AIBL study group. We selected participants with the same diagnoses as we used in ADNI. The total number of MRI scans was 991 from 583 participants.

FTLDNI[3]: The Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI) was funded through the National Institute of Aging and started in 2010. The primary goals of FTLDNI were to identify neuroimaging modalities for tracking frontotemporal lobar degeneration (FTLD) and to assess the value of imaging versus other biomarkers in diagnostic roles. The Principal Investigator of FTLDNI was Dr. Howard Rosen, MD, at the University of California, San Francisco. Three variants of frontotemporal degeneration are differentiated: semantic variant (SV), behavioural variant (BV), and progressive non-fluent aphasia (PNFA). In our studies, we used all three variants along with samples of healthy participants. The total number of MRI scans was 614 from 273 participants.

Table 1 summarizes the sample statistics of the different data sources we used. We applied the "t1-linear pipeline" of the Clinica library [21, 4] to preprocess the raw MRI scans. The pipeline uses the N4ITKmethod for bias field correction and the SyN algorithm from ANTs to perform an affine registration for the alignment of each scan to the Montreal Neurological Institute (MNI) reference space. During the execution of the pipeline, some MRI samples were excluded due to quality checking or some missing information. Additionally, each scan was cropped to the size of $169 \times 208 \times 179$ voxels with 1 mm isotropic resolution.

2.4. Proposed contrastive learning pipeline

Our proposed method utilizes 2D coronal slices of the brain and consists of two modules: a feature extractor and a classification component. The feature extractor is a convolutional neural network trained without any sample labels in a self-supervised way. The classification component is a simple neural network subsequently trained in a supervised way. The proposed architecture is shown in Figure 1.

[2]AIBL: `https://aibl.csiro.au/`

[3]FTLDNI: `https://memory.ucsf.edu/research-trials/research/allftd`

Table 1: Sample statistics of study data per diagnosis state. CN: a cognitively normal state, AD: dementia due to Alzheimer's disease, MCI: mild cognitive impairment, BV: behavioural variant of frontotemporal neurodegeneration, SV: semantic variant of frontotemporal neurodegeneration, PNFA: progressive non-fluent aphasia, μ: mean, σ: standard deviation, MMSE: mini-mental state examination, F: female, M: male.

	CN	AD	MCI	
ADNI3				
Age: $\mu(\sigma)$	74 (7)	77 (8.3)	74.6 (8)	
MMSE: $\mu(\sigma)$	29.4 (0.7)	20.8 (4.5)	27.9 (1.1)	
Sex: F/M	312/221	52/70	140/173	
ADNI2				
Age: $\mu(\sigma)$	75.8 (7)	76.2(7.6)	74.6 (7.9)	
MMSE: $\mu(\sigma)$	29.3 (0.7)	21.1(4.3)	27.8 (1.1)	
Sex: F/M	110/94	120/163	151/203	
AIBL				
Age: $\mu(\sigma)$	73.5 (6.4)	75.4 (7.9)	76.6 (6.5)	
MMSE: $\mu(\sigma)$	29.2 (0.8)	19.5 (5.8)	27.2 (1.3)	
Sex: F/M	239/182	51/37	41/62	
	CN	BV	SV	PNFA
FTLDNI				
Age: $\mu(\sigma)$	64.3 (7.1)	62.1 (5.8)	62.7 (6.8)	68.9 (7.7)
MMSE: $\mu(\sigma)$	29.7 (0.5)	22.6 (6.2)	22.5 (5.7)	24.9 (5.5)
Sex: F/M	72/58	23/48	14/23	19/16

2.4.1. Plane and slice selection

After executing the t1-linear pipeline of the Clinica library, we obtained a 3D image for the brain of each participant. Since we used 2D convolutional operations to reduce the CNN parameter space and model complexity, we required a 2D representation of the brain. Thus, we selected only the coronal plane for the present study. We acknowledge that the selection of another axis could have an effect on predictions, however, our main goal was to investigate the application of SSL and to compare it with traditional supervised approaches. Though, each MRI sample had in total 208 coronal slices, we

considered only 120 coronal slices in the middle. Those slices were the input candidates for the ConvNeXt CNN that was trained in a self-supervised way.

2.4.2. Feature learning

For the self-supervised learning of the ConvNeXt CNN, we used the NNCLR [20] method for learning visual representations from the input data by creating a latent space that constitutes semantic perturbations found in the actual data. We applied a pipeline of random augmentations to a randomly selected coronal slice for the creation of two views. Examples of such views are provided in Figure 2. Thus, those two views represent the same slice of the same brain and should be represented close to each other in the feature space, while the views of different human brains or the views of the same participant's brain recorded in different sessions should lie far apart. Since NNCLR is just a framework for learning representations from data, an underlying CNN model, called often a backbone, is an actual feature extractor. We used the ConvNeXt-tiny architecture [17] as our backbone and loaded the pre-trained version from the PyTorch library for the initialization of the backbone's weights.

2.4.3. Classification component

To utilize the CNN model as a feature extractor we considered only the output produced by a $2D$ adaptive average pooling operation after the last convolutional block (Figure 1 left). Thus, the classification component takes as input the representations of the MRI scans provided by the convolutional neural network. The dimension of the extracted feature vector per MRI slice is 768. Our classification component is a simple neural network consisting of a single fully-connected layer preceded by a layer normalization (Figure 1 right).

2.4.4. Feature attribution

There exist various methods to derive scores for the importance of input features with respect to a given prediction. We used the Integrated Gradients (IG) attribution method that estimates importance scores by approximating integrated gradients [22]. Specifically, IG considers a straight path from some baseline to the input and computes the gradients along that path. These accumulated gradients are called integrated gradients. The baseline is of the same dimensions as the input data and can, for example, contain the values

all set to zero. We used the IG implementation provided by the Captum library[4] to calculate attribution maps.

3. Results

3.1. Contrastive learning

We trained a feature extraction model using NNCLR on the ADNI3, ADNI2, and FTLDNI data for three learning trials. For each trial, we randomly divided data into training and test sets. We used the same training and test sets for all further experiments. If more than one MRI recording was available per participant, then we assigned all of the participant's MRI scans only to one set, thus avoiding data leakage. This resulted in 10% of data belonging to a test set. The model was trained for 1000 epochs using a batch size of 180 samples. The size of the NNCLR queue Q was set to 8192. We applied three different data augmentation techniques with a probability of 0.5 to produce views visualized in Figure 2: a horizontal flip (b), cropping and the consecutive resizing to the original input size (c), and an erasing of a randomly selected region in input that sets all values in that region to zero (d). Figure 3 (a) shows the training curve over 1,000 epochs. During the first epochs, the contrastive loss drops rapidly down, while during the last epochs it starts slowly to saturate.

3.2. Diagnostic group separation

The classification component was trained for 100 epochs on the same three training sets that were used to train a feature extractor. We use a batch size of 64 samples and decay a learning rate with cosine annealing after each epoch for every 20 epochs. Using this setting we created 3 models that classify the extracted representations from MRI samples by a ConvNeXt trained previously via NNCLR into 4 (CN, AD, MCI, BV), 3 (CD, AD, MCI), and 2 (AD, CN) groups, respectively. Each model was trained and tested on the three training and test sets. Furthermore, we evaluated our models on the AIBL dataset, which was not used during training and served as independent test dataset to assess the generalizability of our approach. In our proposed architecture only the classification component was trained since ConvNeXt serves as a feature extractor. However, to compare the

[4]Captum: `https://captum.ai/`

Table 2: Classification results of our proposed architecture, consisting of a ConvNeXt feature extractor trained in a self-supervised way and a single-layer classification component. CN: a cognitively normal state, AD: dementia due to Alzheimer's disease, MCI: mild cognitive impairment, BV: behavioural variant of frontotemporal neurodegeneration, MCC: Matthews correlation coefficient.

	Balanced accuracy	MCC
ADNI2/3 & FTLDNI test set		
CN vs AD vs MCI vs BV:	0.59±0.02	0.32±0.01
CN vs AD:	0.81±0.03	0.6±0.05
AIBL (independent sample)		
CN vs AD vs MCI:	0.53±0.01	0.29±0.02
CN vs AD:	0.79±0.01	0.58±0.01

results with a supervised approach, we let ConvNeXt—after loading pre-trained model parameters for ImageNet—to be trained together with the classification module using the same settings. Figure 3 (b) compares the training loss curves the two different settings. The classifier that uses the ConvNeXt feature encoder backbone trained in a self-supervised way on MRI scans shows lower error rate than using the ConvNeXt backbone with initial model parameters pre-trained on ImageNet. Therefore, we used only the ConvNeXt model that was trained on the MRI scans in a self-supervised way in all following experiments.

Table 2 shows the achieved results of our proposed architecture for the identification of neurodegenerative disorders using the ConvNeXt feature extractor trained via NNCLR. To determine if a 3D MRI scan belongs to a specific diagnostic group, we first derive the latent representation vectors for 100 random coronal 2D slices using feature extractor and then make a prediction for each slice using the classifier (Figure 1). We applied simple voting procedure in which the most frequently occurring group label determines the final prediction. In Section 4.1 below, we discuss the achieved results and compare them with the state of the art.

3.3. Model's interpretability

The application of machine learning methods in the medical area requires some interpretability for the model's decision process. Since the patient's treatment depends on the diagnosis, such clinical decision support systems

should provide more than just a prediction. Thus, we used the Integrated Gradients (IG) attribution method that calculates importance scores for the input regions for a specified prediction label. As visible from Figure 4, the derived attributions provide a rather general indication of important input regions throughout the brain including primarily the grey and white matter tissue. More specifically, the model learned to consider widespread input areas from the temporal and frontal lobe, and the basal ganglia. Notably, the model successfully learned to not consider tissue outside of the brain, the skull, or image artifacts in the background around the head.

4. Discussion

4.1. Feature learning

In our proposed method, we rely on signals that are derived from the data itself rather than on external classification target labels to train a feature extractor. In Table 3 we compare our proposed method with the state of the art, that means other studies that report results for AIBL as an independent validation dataset. Training only a single classification layer on top of the frozen ConvNeXt model shows competitive results against other supervised approaches as well as manual expert rating as reported by Qiu et al. [2]. Note that some papers did not report the *balanced accuracy* measure, thus, their "simple" accuracy results might be biased towards the majority class of cognitively normal people who comprise 80% in AIBL for the group comparison CN vs AD. We could only find one study that reported group separation results using a self-supervised framework. For the ADNI test data, Ouyang et al. [11] achieved a balanced accuracy of 81.9% for a multilayer perceptron classifier when freezing the feature encoder network, and 83.6% when further finetuning the feature encoder network parameters during training. With regard to our achieved level of performance, we can conclude that the ConvNeXt model learned generalizable features for the subsequent downstream classification tasks.

In this study, we considered only coronal slices and 2D convolution operations, which can limit the representational power of the extracted features. Notably, 3D CNNs have much more parameters than 2D models and, therefore, is currently a computationally intractable solution for self-supervised learning that relies on a very large data corpus, data augmentation, and many learning iterations as training typically converges much slower than in supervised learning. More specifically, training our models for 1000 epochs

Table 3: Comparison of our proposed method with the state of the art. The results are provided for studies that used the AIBL dataset for independent evaluation and the group comparison CN vs AD. CN: a cognitively normal state, AD: dementia due to Alzheimer's disease. ‡ Self-supervised learning and simple classifier; † supervised learning; * balanced accuracy was not reported, "simple" accuracy is provided instead, which might be biased towards the majority class (=CN).

Methods trained on the ADNI data	Balanced accuracy for the independent AIBL data
‡Our method	0.79±0.01
†Wen et al. 2020 [4] (2D slice-level CNN)	0.756±0.015
†Wen et al. 2020 [4] (3D patch-level CNN)	0.802±0.016
†Wen et al. 2020 [4] (3D subject-level CNN)	0.862±0.016
†Dyrba et al. 2021 [1] (3D subject-level CNN)	0.82±3.0
†Qiu et al. 2020 [2] (3D patch-level CNN)	*0.870±0.022**
†Han et al. 2022 [5] (3D subject-level CNN)	*0.865**
†Han et al. 2022 [5] (3D patch-level CNN)	*0.875**
Qiu et al. 2020 [2] (neurologists)	*0.823±0.094**

on a single NVIDIA Quadro RTX 6000 GPU took on average 1 day and 3 hours. Other recent SSL studies followed similar 2D approaches [13, 14] or instead drastically downscaled the 3D images to a low $64 \times 64 \times 64$ resolution [11, 12]. In our future work, to better incorporate plane information, three CNNs, i.e. each for another plane (axial, coronal, sagittal) could be combined to learn 3D MRI data representations as recently proposed for supervised models [23]. Alternatively, a Transformer model could be designed to efficiently process smaller 3D patches of the brain, which would not require separate CNN models for each patch [2, 4, 5].

4.2. Neural network interpretability

We chose self-supervised learning not only as a method to learn a powerful feature extractor that can be used for further downstream tasks but also to learn features of the brain regions that can correlate with a specific neurodegenerative disorder. Though a comprehensive analysis of the salient features lies out of the scope of our current work, we applied the Integrated Gradients method to interpret the models and provide insights into the significance of input regions to predictions (Figure 4). In our future work, we will explore methods to summarize calculated attributions per brain region to assess if

specific disease patterns emerge or estimate the correlation between regional relevance scores and brain regions' volumes.

4.3. Conclusion

We presented an architecture for the identification of neurodegenerative disorders from MRI data, consisting of a feature extractor and a classification component. The feature extractor used the ConvNeXt architecture that was trained in a contrastive self-supervised way. We showed that a ConvNeXt trained via the NNCLR method on MRI data can successfully be used as a feature extractor for subsequent downstream tasks by training only an additional single-layer neural network component which performs the classification. The conducted experiments showed similar state-of-the-art results CNNs trained in a supervised way. We highlighted the possible future work to increase the performance and interpretability of the proposed methodology. With this presented approach, we provide not only an additional practical application of self-supervised learning to MRI data but also insights into the application of such systems for clinical studies in which the interpretability of the model's decision is crucial.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites

in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data collection and sharing for this project was funded by the Frontotemporal Lobar Degeneration Neuroimaging Initiative (National Institutes of Health Grant R01 AG032306). The study is coordinated through the University of California, San Francisco, Memory and Aging Center. FTLDNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Ethical approval

Data collection of the respective neuroimaging initiatives was approved by the internal review boards of each of the participating study sites. All initiatives met common ethical standards in the collection of the data such as the Declaration of Helsinki. Analysis of the data was approved by the internal review board of the Rostock University Medical Center, reference number A 2020-0182.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant DY151/2-1, project ID 454834942.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability statement

The data used for this study is publicly available and can be obtained from the respective repositories: Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL), and Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI). Our source code will be made publicly available at: https://github.com/VadymV/clinic-net

References

[1] M. Dyrba, M. Hanzig, S. Altenstein, S. Bader, T. Ballarini, F. Brosseron, K. Buerger, D. Cantré, P. Dechent, L. Dobisch, E. Düzel, M. Ewers, K. Fliessbach, W. Glanz, J.-D. Haynes, M. T. Heneka, D. Janowitz, D. B. Keles, I. Kilimann, C. Laske, F. Maier, C. D. Metzger, M. H. Munk, R. Perneczky, O. Peters, L. Preis, J. Priller, B. Rauchmann, N. Roy, K. Scheffler, A. Schneider, B. H. Schott, A. Spottke, E. J. Spruth, M.-A. Weber, B. Ertl-Wagner, M. Wagner, J. Wiltfang, F. Jessen, S. J. Teipel, Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease, Alzheimer's research & therapy 13 (2021) 191. doi:10.1186/s13195-021-00924-2.

[2] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, M. Kaku, Y. Zhou, Y. J. Alderazi, A. Swaminathan, S. Kedar, M.-H. Saint-Hilaire, S. H. Auerbach, J. Yuan, E. A. Sartor, R. Au, V. B. Kolachalama, Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (2020) 1920–1933. doi:10.1093/brain/awaa137.

[3] F. Eitel, M.-A. Schulz, M. Seiler, H. Walter, K. Ritter, Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research, Experimental Neurology 339 (2021) 113608. doi:10.1016/j.expneurol.2021.113608.

[4] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, Medical Image Analysis 63 (2020) 101694. doi:10.1016/j.media.2020.101694.

[5] K. Han, M. He, F. Yang, Y. Zhang, Multi-task multi-level feature adversarial network for joint Alzheimer's disease diagnosis and atrophy localization using sMRI, Physics in medicine and biology 67 (2022). doi:10.1088/1361-6560/ac5ed5.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly,

J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986.

[8] A. E. Orhan, V. V. Gupta, B. M. Lake, Self-supervised learning through the eyes of a child, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[9] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, C. Lippert, 3D self-supervised methods for medical imaging, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/d2dc6368837861b42020ee72b0896182-Abstract.html.

[10] C. Hu, C. Li, H. Wang, Q. Liu, H. Zheng, S. Wang, Self-supervised learning for MRI reconstruction with a parallel network training framework, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VI, volume 12906 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 382–391. doi:10.1007/978-3-030-87231-1_37.

[11] J. Ouyang, Q. Zhao, E. Adeli, E. V. Sullivan, A. Pfefferbaum, G. Zaharchuk, K. M. Pohl, Self-supervised longitudinal neighbourhood embedding, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer

Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 80–89. doi:10.1007/978-3-030-87196-3_8.

[12] A. Fedorov, L. Wu, T. Sylvain, M. Luck, T. P. DeRamus, D. Bleklov, S. M. Plis, V. D. Calhoun, On self-supervised multimodal representation learning: An application to Alzheimer's disease, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 1548–1552. doi:10.1109/ISBI48211.2021.9434103.

[13] R. Couronné, P. Vernhet, S. Durrleman, Longitudinal self-supervision to disentangle inter-patient variability from disease progression, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 231–241.

[14] S. Dadsetan, M. Hejrati, S. Wu, S. Hashemifar, Cross-domain self-supervised deep learning for robust Alzheimer's disease progression modeling, 2022. URL: https://arxiv.org/abs/2211.08559. doi:10.48550/ARXIV.2211.08559.

[15] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 9525–9536. URL: https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

[16] L. Sixt, M. Granz, T. Landgraf, When explanations lie: Why many modified BP attributions fail, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 9046–9057. URL: http://proceedings.mlr.press/v119/sixt20a.html.

[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: 2022 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[18] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - A new approach to self-supervised learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

[19] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607. URL: http://proceedings.mlr.press/v119/chen20j.html.

[20] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, A. Zisserman, With a little help from my friends: Nearest-neighbor contrastive learning of visual representations, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 9568–9577. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00945. doi:10.1109/ICCV48922.2021.00945.

[21] A. Routier, N. Burgos, M. Diaz-Melo, M. Bacci, S. Bottani, O. El-Rifai, S. Fontanella, P. Gori, J. Guillon, A. Guyot, R. Hassanaly, T. Jacquemont, P. Lu, A. Marcoux, T. Moreau, J. Samper-González, M. Teichmann, E. Thibeau-Sutre, G. Vaillant, J. Wen, A. Wild, M. O. Habert, S. Durrleman, O. Colliot, Clinica: An open-source software platform for reproducible clinical neuroscience studies, Frontiers Neuroinformatics 15 (2021) 689675. doi:10.3389/fninf.2021.689675.

[22] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney,

NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.

[23] H. Qiao, L. Chen, F. Zhu, A fusion of multi-view 2D and 3D convolution neural network based MRI for Alzheimer's disease diagnosis, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3317–3321. doi:10.1109/EMBC46164.2021.9629923.

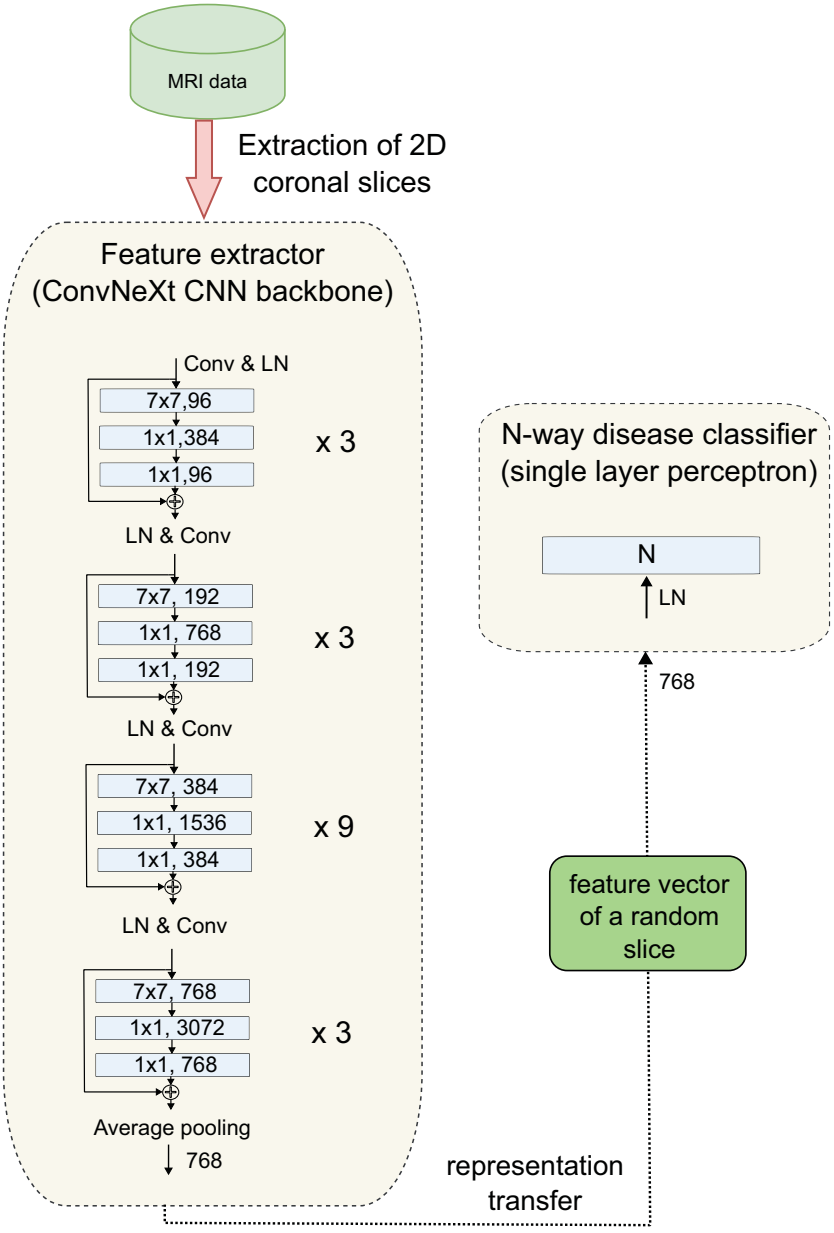
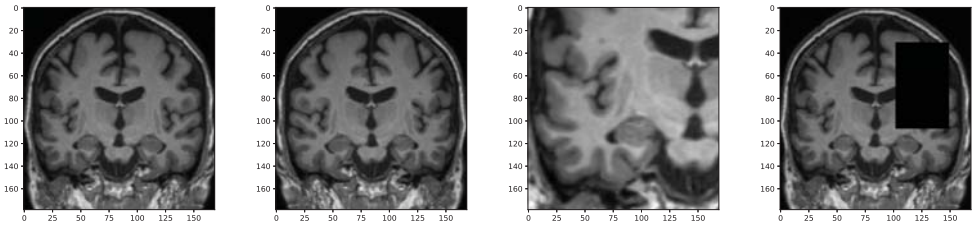
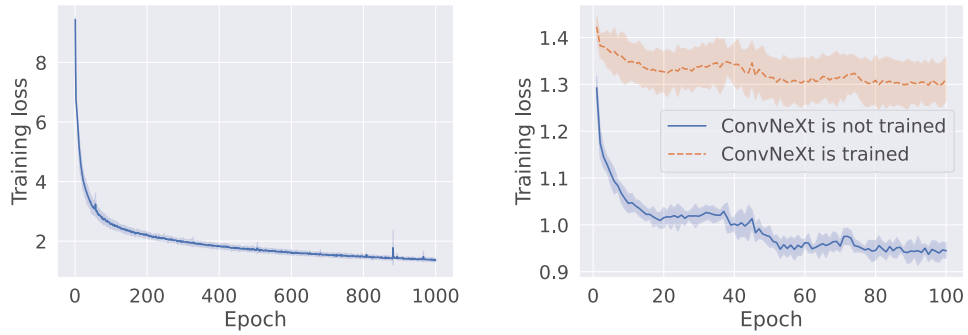


Figure 1: Illustration of the proposed architecture. A CNN model, trained in a self-supervised way, extracts features from coronal slices. The classification component learns to classify neurodegenerative disorders from the extracted features. CNN: convolutional neural network. LN: layer normalization. Conv: convolutional operation.



(a) Original (b) Horizontal flip (c) Crop and resize (d) Erasing

Figure 2: Randomly applied data augmentations to the input during training.



(a) Self-supervised ConvNeXt training (b) Classification component training

Figure 3: Loss curves of the models trained for three learning trials. Shaded areas denote 1 standard deviation. (a) Self-supervised training of the ConvNeXt feature extractor network via the NNCLR method. (b) Comparison of the classification component training following two strategies: with a ConvNeXt feature encoder backbone that was previously trained on the MRI data using the self-supervised NNCLR method and was not trained further during classifier training (blue solid line); and a ConvNeXt feature encoder backbone with the initial ImageNet model weights loaded and which was trained together with the classification component (supervised training, red dashed line).

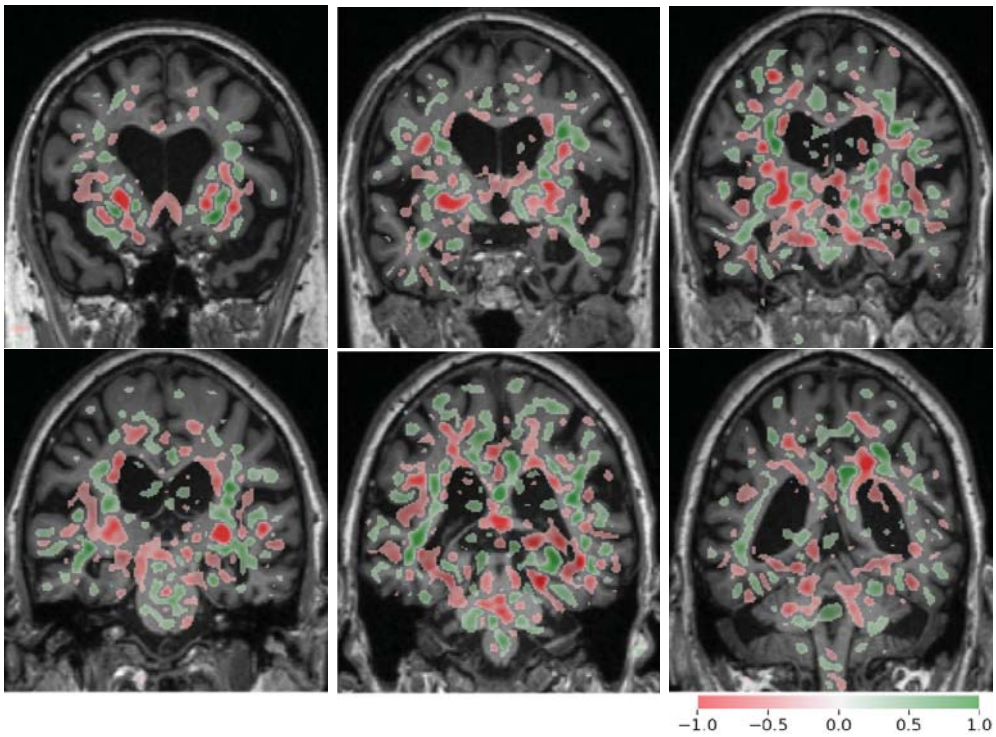


Figure 4: Calculated attributions using the Integrated Gradients method for one arbitrarily selected MRI scan that was correctly classified as Alzheimer's dementia. Green and red colour highlight pixel contributions to the model's prediction. The attribution overlay image was smoothed with a Gaussian kernel $\sigma = 2.5$ and thresholded at $|ig| < 0.2$ to remove small values and improve visualization.

NeuroImage: Clinical 17 (2018) 579–586

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction

Judith Henf[a,b,*], Michel J. Grothe[a], Katharina Brueggen[a], Stefan Teipel[a,b], Martin Dyrba[a,c]

[a] DZNE, German Center for Neurodegenerative Diseases, Rostock, Germany
[b] Department for Psychosomatic Medicine, University Medicine Rostock, Rostock, Germany
[c] MMIS Group, University of Rostock, Rostock, Germany

ARTICLE INFO

Keywords:
Alzheimer's disease
Mild cognitive impairment
Partial volume effects
Diffusion tensor imaging
Mean diffusivity
Gray matter

ABSTRACT

Mean diffusivity (MD) measured by diffusion tensor imaging can reflect microstructural alterations of the brain's gray matter (GM). Therefore, GM MD may be a sensitive marker of neurodegeneration related to Alzheimer's Disease (AD). However, due to partial volume effects (PVE), differences in MD may be overestimated because of a higher degree of brain atrophy in AD patients and in cases with mild cognitive impairment (MCI). Here, we evaluated GM MD changes in AD and MCI compared with healthy controls, and the effect of partial volume correction (PVC) on diagnostic utility of MD.

We determined region of interest (ROI) and voxel-wise group differences and diagnostic accuracy of MD and volume measures between matched samples of 39 AD, 39 MCI and 39 healthy subjects before and after PVC. Additionally, we assessed whether effects of GM MD values on diagnosis were mediated by volume.

ROI and voxel-wise group differences were reduced after PVC. When using these ROIs for predicting group separation in logistic models, both PVE corrected and uncorrected GM MD values yielded a poorer diagnostic accuracy in single predictor models than regional volume. For the discrimination of AD patients and healthy controls, the effect of GM MD on diagnosis was significantly mediated by volume of hippocampus and posterior cingulate ROIs.

Our results suggest that GM MD measurements are strongly confounded by PVE in the presence of brain atrophy, underlining the necessity of PVC when using these measurements as specific metrics of microstructural tissue degeneration. Independently of PVC, regional MD was not superior to regional volume in separating prodromal and clinical stages of AD from healthy controls.

1. Introduction

Diffusion Tensor Imaging (DTI) allows the examination of microstructural cell damage (Uluğ et al., 1999). During the course of Alzheimer's Disease (AD), these changes are assumed to predate the macrostructural changes that are observable using volumetric magnetic resonance imaging (Weston et al., 2015). Therefore, DTI measures could be used as sensitive markers of AD-related neurodegeneration and help to identify individuals with prodromal AD. Mean diffusivity (MD) reflects the average degree of diffusion of water molecules in all directions. Fractional anisotropy (FA), on the other hand, reflects directionality of diffusion. As opposed to FA that is mostly used for assessing white matter (WM) fiber tract integrity, MD may also be used to assess microstructural alterations of the brain's gray matter (GM). An increase of MD in GM is assumed to reflect the breakdown of microstructural barriers to diffusion which would predate volumetric changes (Weston et al., 2015). Therefore, GM MD could be a valuable measure of early GM cell damage in AD.

Previous studies on GM MD in AD reported increases of hippocampal MD and whole brain GM MD in subjects with mild cognitive impairment (MCI) as compared to healthy controls (Cherubini et al., 2010; Eustache et al., 2016; Fellgiebel et al., 2004; Müller et al., 2005; Scola et al., 2010), as well as in MCI patients that converted to AD compared to MCI subjects that remained stable over a period of several years (Douaud et al., 2013; Fellgiebel et al., 2006; Scola et al., 2010; van Uden et al., 2016). In some studies, hippocampal diffusivity even demonstrated a higher diagnostic and prognostic accuracy than hippocampal volume (Fellgiebel et al., 2006; Kantarci et al., 2005; Müller

Abbreviations: AD, Alzheimer's Disease; DTI, diffusion tensor imaging; FLAIR, fluid-attenuated inversion recovery; GM, gray matter; MCI, mild cognitive impairment; MD, mean diffusivity; PCC, posterior cingulate cortex; PVC, partial volume correction; PVE, partial volume effects; ROI, region of interest

* Corresponding author at: DZNE German Center for Neurodegenerative Diseases, Rostock, c/o Zentrum für Nervenheilkunde, Gehlsheimer Str. 20, D-18147 Rostock, Germany.
E-mail address: judith.henf@med.uni-rostock.de (J. Henf).

http://dx.doi.org/10.1016/j.nicl.2017.10.005
Received 20 February 2017; Received in revised form 6 September 2017; Accepted 3 October 2017
Available online 04 October 2017

et al., 2007). Considering MCI a possibly prodromal phase of AD, these findings corroborate the potential utility of GM MD as an earlier biomarker than volumetric changes. However, longitudinal evidence for the assumed succession of microstructural and macrostructural GM alterations is still lacking, even though microstructural alterations have been shown to precede macrostructural changes in the white matter (Ly et al., 2014).

Regarding the findings on GM MD changes in MCI subjects, one needs to acknowledge that extensive hippocampus atrophy is already present in the MCI stage of AD (Shi et al., 2009), probably confounding the measurement of GM MD values due to partial volume effects (PVE). Partial volume effects arise when the signal within a cortical voxel does not purely represent the GM signal at this location, but is confounded by intermixing signals of surrounding cerebrospinal fluid (CSF) or WM tissue. In older people and AD patients, the probability of GM voxels to also contain CSF signal increases due to brain atrophy (Jeon et al., 2012). Thus, given the high MD signal in CSF, PVE may lead to overestimation of differences in GM MD in patients with AD as compared to healthy controls. In this case, MD would represent the joint effect of microstructural changes and macrostructural brain atrophy. However, most of the studies using DTI data of healthy older and AD participants did not consider PVE at all (Cherubini et al., 2010; Douaud et al., 2013; Eustache et al., 2016; Müller et al., 2007). One study stated that PVC was not necessary as the regions of interest (ROIs) were not located at the surface of the brain (Eustache et al., 2016).

Several approaches for correcting PVE in DTI data have been proposed. Kantarci et al. (2001, 2010) used a dedicated DTI sequence based on fluid-attenuated inversion recovery (FLAIR) to suppress the CSF signal during data acquisition. Salminen et al. (2016a, 2016b) recomputed the MD for each region of interest (ROI) using a customized diffusion weighted imaging sequence with multiple non-zero *b*-values and an extended model to fit the tensor. They could show that the decay of signal intensity was mono-exponential for *b*-values ≥ 680, indicating successful suppression of CSF signal after removal of $b \sim 0$ data. Both approaches by Kantarci et al. (2001, 2010) and by Salminen et al. (2016a, 2016b) are well-grounded, but they require additional MR sequences that go beyond current clinical standards, either a FLAIR DTI sequence or a DTI sequence with more than one non-zero *b*-value. So far, three approaches were proposed for post-acquisition partial volume correction (PVC) of single-shell DTI as used in clinical practice. First, Rose et al. (2008) proposed to exclude all voxels exceeding a certain threshold of MD (1300×10^{-6} mm^2/s). This fixed threshold, however, leads to an altered distribution of MD values (Weston et al., 2015). Second, Pasternak et al. (2009) introduced the method of free water elimination, consisting of sub-voxel modeling of two tensors: a tissue and a CSF tensor. However, this method suffers from computational challenges that require defining regularization parameters and constraints on tensor estimates to obtain a unique solution for model fitting. Although Pasternak's approach has been found to reduce CSF contamination and to improve diffusion measures in white matter fiber tracts (Berlot et al., 2014; Metzler-Baddeley et al., 2012; Pasternak et al., 2009), a formal validation of this model is still missing. Additionally, differences of the permeability of the cell bodies in gray matter in comparison to the myelin sheets of the fiber bundles in white matter may make it necessary to adjust some of the model parameters when applying this model to gray matter (Pasternak et al., 2009). Third, Koo et al. (2009) proposed the CSF contamination model. They found a high consistency of their PVC approach with FLAIR DTI, corroborating the validity of this post-hoc data correction procedure. Jeon et al. (2012) successfully applied this CSF contamination model to GM MD data of young and older participants and participants with AD to correct for partial volume effects. They found that PVC stronger reduced GM MD in elder and AD participants than in younger participants, but they did not examine the potential diagnostic value of corrected GM MD for AD.

The aim of the current study was to evaluate the effect of PVC on cortical and subcortical gray matter MD changes and their potential diagnostic utility in AD and MCI. Therefore, we assessed regional and voxel-wise GM MD changes in AD and MCI compared to cognitively healthy older controls (HC). To correct the MD values for PVE, we chose the CSF contamination model proposed by Koo et al. (2009) because it can be applied on single-shell DTI data post acquisition and it was previously used for PVC of GM (Jeon et al., 2012). We hypothesized that people with AD dementia and people with MCI would have higher MD values than matched healthy controls. Following previous literature (e.g. Fellgiebel et al., 2004), this effect should be most pronounced in the left hippocampus. As we assumed that MD alterations were partly confounded by increased brain atrophy in AD subjects, we expected that PVC would decrease group differences and diagnostic accuracy of MD and reduce the correlation between regional MD and volume measures.

2. Methods

2.1. Sample

We included 117 subjects from the database of the German Center for Neurodegenerative Diseases in Rostock. Among those subjects, there were 39 cognitively healthy controls (HC), 39 individuals with AD and 39 subjects with MCI. The subjects were matched according to age, gender, years of education and imaging protocol, given that two different sequences were used for MRI and DTI acquisition (Table S1 of the Supplementary material). All diagnoses were made in an interdisciplinary team of an experienced neurologist, psychiatrist and neuropsychologist. Diagnosis of AD was made in accordance with the NINCDS-ADRCA criteria (McKhann et al., 2011) and MCI was diagnosed according to the Mayo criteria (Petersen et al., 1999).

All subjects had to fulfill the following criteria: MRI scans that passed the quality control (see Section 2.2 for details), time interval between MRI and neuropsychological assessment of at maximum 3 months, no significant neurological, psychiatric or medical condition (except for MCI or AD), in particular cerebrovascular apoplexy, vascular dementia, subclinical hypothyroidism or substance abuse. Healthy controls were free of cognitive complaints and scored within one standard deviation of the age and education adjusted norm in all subtests of the Consortium to Establish a Registry of Alzheimer's Disease (CERAD) cognitive battery (Morris et al., 1987).

All subjects or their representatives had given informed consent according to the declaration of Helsinki. The study was approved by the ethics committee of the University Medical Center Rostock (HV 2009–0010).

2.2. Image acquisition

All images were acquired using a 3 T scanner (SIEMENS MAGNETOM Verio). Two different protocols were used for the T1-weighted MRI and DTI sequences. Detailed imaging parameters can be found in Table S1 of the Supplementary material. Only scans were included that passed quality control by a trained expert who evaluated ghosting effects, blurring, motion and susceptibility artifacts.

2.3. Processing of MRI and DTI data

Deformation-based analysis of the T1-weighted scans was performed using SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm/) implemented in Matlab 2013b (Mathworks, Natwick). First, MRI scans were segmented into GM, WM and CSF partitions using the VBM8 toolbox. Then, the images were normalized to an aging and AD-specific reference template from a previous study (Grothe et al., 2013) using the Diffeomorphic Anatomical Registration Through Exponentiated Lie (DARTEL) algebra algorithm (Ashburner, 2007) with modulation for non-linear

transformation components only. Finally, gray matter maps were smoothed using an 8 mm full width at half maximum (FWHM) kernel. To verify the sensitivity of our results to the choice of segmentation and normalization algorithm, we also applied an alternative segmentation and normalization algorithm on our imaging data using FSL (Version 5.0.9, FMRIB, Oxford, UK, http://www.fmrib.ox.ac.uk/fsl/). All further analysis steps were identical in both processing pipelines.

DTI data were preprocessed using the diffusion toolbox of FSL. Automated batch processing of DTI data in FSL was performed using in-house software. Data was corrected for eddy currents and head motion. Then, skull stripping was performed using the Brain Extraction Tool. Diffusion Tensors were fitted to the data with DTIfit, and the resulting FA and MD maps were coregistered to the T1-weighted scans.

To correct for PVE caused by CSF in GM voxels, the CSF contamination model (Koo et al., 2009) was used as adapted by Jeon et al. (2012):

$$exp(-b\overline{D}(k)) = \lambda_{app-GM}(k) \bullet exp(-b\overline{D}_{GM}(k)) + \lambda_{app-CSF}(k) \bullet exp(-b\overline{D}_{CSF}(k)) \quad (1)$$

with $\overline{D}(k)$ being the measured mean diffusivity at a specific voxel k, $\lambda_{app-GM}(k)$ and $\lambda_{app-CSF}(k)$ being the apparent signal fraction weightings of GM and CSF compartments, b being 1000 s/mm^2 and $\overline{D}_{csf}$ being defined as a constant, 3.0×10^{-3} mm^2/s (Pasternak et al., 2009). The tissue probability maps obtained from the segmentation algorithm of the T1-weighted scans provided the apparent signal fraction weightings $\lambda_{app-GM}(k)$ and $\lambda_{app-CSF}(k)$ (Jeon et al., 2012). Eq. (1) was solved for $\overline{D}_{GM}$ and applied to the coregistered MD maps. Potential WM partial volume effects were excluded by eliminating those voxels with a FA > 0.2.

Coregistered PVE corrected and uncorrected MD maps were normalized to the aforementioned aging and AD-specific template by applying the deformation fields obtained for the T1-weighted scans. For voxel-wise analyses, preprocessed MD and GM maps were spatially smoothed using an 8 mm FWHM kernel.

2.4. Extraction of gray matter mean diffusivity and volume in regions of interest

Left and right hippocampi as well as the posterior cingulate cortex (PCC) were chosen as regions of interest following the literature on peak differences in regional MD and volumes in AD (Cherubini et al., 2010; Eustache et al., 2016; Rose et al., 2008). The Harvard-Oxford structural atlas (Desikan et al., 2006) was warped from MNI space to the aging and AD-specific reference space using non-linear DARTEL transformation. This normalized atlas was used to provide anatomical labels for the examined ROIs. Only those voxels were included that had a gray matter probability of 50% or higher according to the individual normalized tissue probability map of the respective participant. Average regional MD values and volumes were then obtained for each subject using the normalized data.

2.5. Statistical modeling

Whole-brain voxel-wise comparisons were made using analysis of covariance (ANCOVA) models in SPM8, with diagnosis as factor and MRI protocol, age, gender and years of education as covariates. Voxel-wise effects were assessed at a statistical threshold of $p < 0.001$, uncorrected. ROI based analysis of MD and volume was conducted using linear models in SPSS 21, with diagnosis as main outcome and the aforementioned covariates (ANCOVA). To determine the effect size of each factor in the ANCOVA, a partial eta^2 was computed. Then, post-hoc *t*-tests were performed.

To compare the accuracy of group separation based on corrected and uncorrected regional MD and regional volume, logistic regression models including MRI protocol, age, gender and years of education as covariates were calculated using the *glm* function in R version 3.3.2 (R Development Core Team, 2008). The binary outcome to predict was the diagnosis (MCI vs. HC or AD vs. HC). Models were arranged to either contain volume, PVE corrected MD or uncorrected MD. To evaluate the diagnostic utility of the predictors in the logistic regression models, we calculated areas under the receiver operator characteristic curve (AUC) for each model and compared the AUC values using bootstrapped CIs ($N = 5000$, *pROC* package in R).

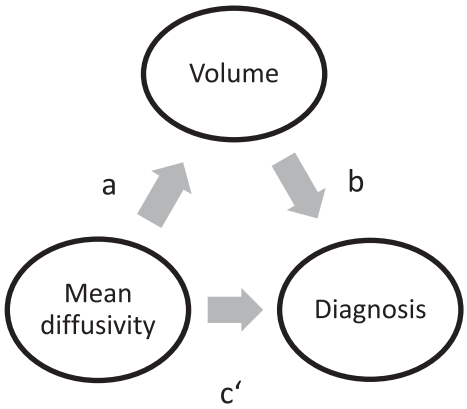


Fig. 1. Mediation model.

We used partial correlations calculated in SPSS to determine the correlations between regional MD and regional volumes, controlled for the covariates MRI protocol, age, gender and years of education. To compare the resulting correlations, the Williams' test (Williams, 1959) for dependent correlations was used as implemented in the *Psych* package in R. A mediation analysis was performed to determine whether the relationship of GM MD and diagnosis was mediated by regional volume (Fig. 1). Separate mediation models were estimated for the diagnostic subgroups (MCI vs. HC or AD vs. HC) and for PVE corrected and uncorrected MD values. Age, years of education, gender and MRI protocol were included as covariates. Direct (c′) and indirect effects (a ∗ b) were calculated using the PROCESS tool for SPSS by Hayes (2013), in which logistic regression models were applied for the prediction of diagnostic subgroup. Statistical significance of the direct and indirect effects was determined using bias-corrected bootstrapped confidence intervals (CI) with 10,000 repetitions and a CI of 95%.

To assess the sensitivity of our imaging variables (MD and volume) to the choice of segmentation and normalization algorithm, we correlated the imaging measures resulting from the SPM processing pipeline with the imaging measures resulting from the FSL pipeline. Also, we calculated the Pearson correlation of the AUCs resulting from logistic regression using the *SPM* imaging values with those resulting from logistic regression using the *FSL* imaging values. Thus, we could also determine the sensitivity of our results to the segmentation and normalization algorithm.

3. Results

3.1. Sample

The diagnostic subgroups did not differ in age and gender but they did slightly differ in years of education (see Table 1). As expected, the groups differed in MMSE scores with AD patients having the lowest scores and healthy controls having the highest scores.

3.2. Whole brain voxel-wise analysis

Without PVC, MD was elevated across wide parts of the cortex in AD patients compared to healthy controls (Fig. 2). When comparing MCI cases with healthy controls, increased MD was found in both hippocampi. After correcting for partial volume effects, MCI cases and

Table 1
Sample characteristics.

Diagnosis	AD	MCI	Healthy controls (HC)
No. of participants (women)	39 (21)	39 (21)	39 (21)
Age, mean (SD)[1]	74.3 (5.1)	74.9 (5.5)	73.3 (5.3)
Years of education, mean (SD) [2]	11.8 (2.7)	12.5 (2.8)	13.2 (2.8)
MMSE, mean (SD)[3]	22.56 (4.8)	26.92 (1.7)	28.26 (1.0)

[1] $F(2,114) = 0.92$, $p = 0.4$, [2] $F(2,114) = 2.61$, $p = 0.08$, [3] $F(2,114) = 39.18$, $p < 0.001$.

healthy controls did not differ significantly in GM MD. The comparison of AD patients with healthy controls still yielded significant results, with elevated MD in both hippocampi, as well as in the middle and posterior cingulate cortex (Fig. 2).

3.3. Analysis of covariance (ROI based)

Main effects for diagnosis on MD were found in all three ROIs, regardless of PVC (Table 2). Effect sizes were larger in uncorrected MD than in PVE corrected MD data. Post hoc tests revealed that the main effect of diagnosis was driven by the significantly increased MD in the AD subgroup compared to both healthy controls and MCI subjects. Even though mean MD values of the MCI group were numerically greater than in the HC group for all ROIs, none of these differences were statistically significant.

Main effects for regional volume were found in all ROIs and with similar effect sizes across ROIs. Post hoc tests showed that this effect was also mainly driven by decreased volumes in the AD group. However, a significant difference in the MCI group compared to the healthy controls was found for the left hippocampus volume (Table 2).

3.4. Logistic regression models for ROIs

The results of the logistic regression models are shown in Table 3. When discriminating MCI from HC, the models with bilateral hippocampal MD or volume yielded AUCs above chance, whereas the AUCs of the models containing MD or volume of the PCC did not exceed chance level. When comparing the AUCs of the different models per ROI, the differences of AUCs between those models containing uncorrected MD, corrected MD or volume were non-significant ($p > 0.05$) in the MCI/HC subgroup.

When discriminating AD from HC, all models yielded significant AUCs. Comparing the AUCs in this subgroup, the volume models yielded greater AUCs than the models with PVE corrected MD. Also, the PVE uncorrected MD models reached significantly greater AUCs than PVE corrected MD models.

In all diagnostic group comparisons, the AUCs were numerically greater for the models containing volume as a predictor than those with MD as a predictor, although the AUCs mostly did not differ significantly.

3.5. Correlation analysis (ROI based)

Partial correlations of regional MD values and volumes are shown in Fig. 3. In each diagnostic group and for all ROIs, the correlation between GM MD and GM volume was significantly higher in uncorrected data than in PVE corrected data (all $p < 0.001$, Williams' Test). In some cases, the correlations even became non-significant after PVC was applied. In the PCC, the differences in correlations before and after PVC were numerically smaller than the differences in both hippocampi (range of difference: 0.02 to 0.08 in the PCC and 0.16 to 0.39 in the hippocampi), indicating more severe PVEs in the hippocampus compared to the PCC. In the right hippocampus, the change in correlation was considerably higher in the MCI subgroup than in the AD and HC groups. However, this result could not be reproduced when a different segmentation and normalization algorithm was used (Fig. S2 in the Supplementary material).

3.6. Mediation analysis

The results of the mediation analysis can be found in Table 4. As indicated by the bootstrapped CIs, the indirect effect of regional volume was significant for the discrimination of AD vs. HC, whereas none of the direct effects of MD were significantly different from zero. When separating MCI from HC, a trend (90% CI) for an indirect effect could be found in the left hippocampus, equally for PVE corrected and uncorrected MD. Numerically, the indirect effects were greater for uncorrected MD than for corrected MD in the bilateral hippocampi, although the CIs clearly overlapped.

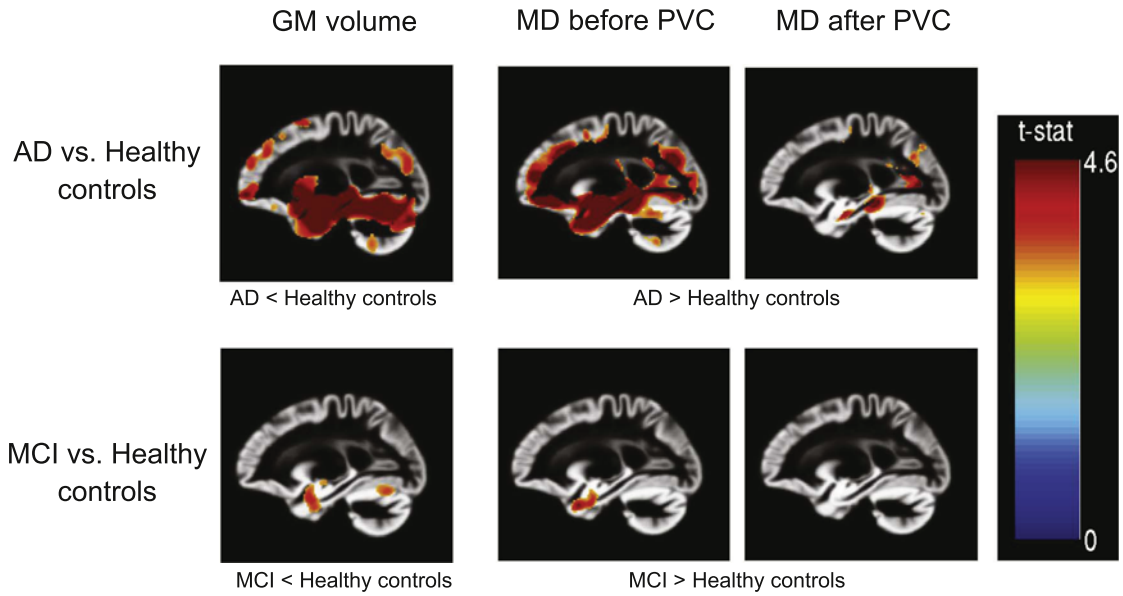


Fig. 2. T-maps of voxel-wise comparisons between the diagnosis subgroups, controlled for age, gender, years of education and MRI protocol. $p < 0.001$ (uncorrected), x = −22.

Table 2
ROI Analysis of uncorrected and corrected MD values and volumes, controlled for age, gender, years of education and MRI protocol.

		Mean (SD) MD in $mm^2/s \times 10^{-3}$ Volume in ml			ANCOVA			
		AD	MCI	HC	F	p (unc.)	Partial eta^2	Post hoc t-tests, $p < 0.05$
Left hippocampus	MD without PVC	1.22 (0.13)	1.15 (0.08)	1.10 (0.08)	14.61	< 0.001	0.21	AD > MCI, AD > HC
	MD with PVC	1.05 (0.10)	1.02 (0.072)	0.99 (0.07)	5.59	< 0.01	0.09	AD > HC AD > MCI†
	Volume	3.77 (0.67)	4.44 (0.64)	4.82 (0.51)	30.16	< 0.001	0.35	AD < MCI < HC
Right hippocampus	MD without PVC	1.26 (0.13)	1.19 (0.07)	1.15 (0.08)	13.98	< 0.001	0.23	AD > MCI, AD > HC
	MD with PVC	1.10 (0.09)	1.07 (0.06)	1.05 (0.06)	4.9	< 0.01	0.08	AD > MCI† AD > HC
	Volume	2.40 (0.48)	2.85 (0.38)	3.02 (0.32)	26.62	< 0.001	0.33	AD < MCI AD < HC
PCC	MD without PVC	1.24 (0.16)	1.14 (0.10)	1.11 (0.09)	19.68	< 0.001	0.26	AD > MCI, AD > HC
	MD with PVC	1.05 (0.15)	0.98 (0.10)	0.95 (0.09)	15.46	< 0.001	0.22	AD > MCI, AD > HC
	Volume	11. 93 (1.16)	13.51 (1.11)	13.89 (1.29)	28.38	< 0.001	0.34	AD < MCI AD < HC

† Post hoc comparisons revealed a trend ($p < 0.1$).

3.7. Alternative segmentation and normalization (FSL)

Detailed results can be found in the Supplementary material (Tables S2, S3 and Fig. S2). In sum, the Pearson correlations of the MD and volume values were moderate to strong ($r = 0.54$ to 0.85) with the correlation being the lowest in the PVE corrected MD values (Table 5).

The Pearson correlation of the AUCs resulting from the respective processing pipelines (FSL vs. SPM) was $r = 0.86$, indicating a high overall consistency of the results of the logistic regression.

4. Discussion

In the current study, we evaluated GM MD changes and their diagnostic utility in AD and MCI when considering partial volume effects. In line with our hypothesis, our results showed that partial volume correction led to a decreased size and number of significant clusters in voxel-wise comparisons when compared to uncorrected MD values. Similarly, in the ROI-based analyses, PVE corrected MD was significantly inferior in separating diagnostic groups compared to uncorrected MD and volume. Thirdly, we showed that the relationship of regional GM MD and diagnosis was significantly mediated by regional volume in AD and HC, and that the correlations between regional MD values and volumes in bilateral hippocampi and PCC were significantly reduced by PVE correction.

The higher correlations of volume with uncorrected MD compared to PVE corrected MD suggest that uncorrected MD values are confounded by atrophy effects resulting in CSF contamination of measured GM MD values. This finding underscores the necessity to correct for PVE if microstructural effects independently of atrophy are the outcome of interest. Our results are in line with a previous study by Jeon et al. (2012), analyzing the regional correlations between cortical thickness and MD in AD, older HC and young HC. Cortical thickness is a measure that is sensitive to changes in gray matter volume, especially cortical thinning (Hutton et al., 2009). Jeon et al. (2012) found significant correlations of MD and cortical thickness in uncorrected data in all subgroups. These correlations decreased significantly in cognitively healthy older and AD participants when applying PVC. We found a similar pattern when examining correlations between regional GM MD

Table 3
Results of the logistic regression analysis, controlled for age, gender, years of education and MRI protocol.

Diagnosis	ROI	PVC	MD		Volume	
			β	AUC [95% CI]	β	AUC [95% CI]
MCI vs. HC	Left hippocampus	With PVC	0.55†	0.67 [0.54, 0.78]	− 0.85**	0.72 [0.60, 0.83]
		Without PVC	0.65*	0.69 [0.56, 0.81]		
	Right hippocampus	With PVC	0.46	0.64 [0.51, 0.76]	− 0.62*	0.68° [0.56, 0.80]
		Without PVC	0.73*	0.71 [0.59, 0.82]		
	Posterior cingulate cortex	With PVC	0.39	0.63 [0.50, 0.75]	− 0.32	0.62 [0.49, 0.74]
		Without PVC	0.39	0.63 [0.50, 0.75]		
AD vs. HC	Left hippocampus	With PVC	1.07**	0.72 [0.61, 0.83]	− 2.64***	0.90°° [0.81, 0.96]
		Without PVC	1.89***	0.82°° [0.72, 0.90]		
	Right hippocampus	With PVC	0.92**	0.71 [0.59, 0.82]	− 2.22***	0.87°° [0.78, 0.94]
		Without PVC	1.65***	0.81°°° [0.70, 0.90]		
	Posterior cingulate cortex	With PVC	2.27***	0.81 [0.70, 0.90]	− 2.23***	0.88° [0.80, 0.95]
		Without PVC	2.54***	0.84°° [0.74, 0.92]		
AD vs. MCI	Left hippocampus	With PVC	0.45	0.61 [0.48, 0.73]	− 1.40***	0.79°° [0.69, 0.89]
		Without PVC	1.07**	0.72°°° [0.60, 0.83]		
	Right hippocampus	With PVC	0.41	0.57 [0.44, 0.70]	− 1.29***	0.78°° [0.67, 0.88]
		Without PVC	1.06**	0.71°°° [0.59, 0.82]		
	Posterior cingulate cortex	With PVC	1.81***	0.75 [0.64, 0.86]	− 1.73***	0.84° [0.75, 0.92]
		Without PVC	2.05***	0.79° [0.68, 0.89]		

Standardized regression weights (β) resulting from logistic regression analysis. The confidence intervals of the area under the receiver operator characteristic curve (AUC) were calculated using bootstrapping.
Asterisks (*) indicate significance of the standardized regression weights (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $^{\dagger}p < 0.1$). Circles (°) indicate significantly larger AUCs compared to the AUC of the MD_{PVC} model, determined using bootstrapping (°$p < 0.1$, °°$p < 0.01$, °°°$p < 0.001$). None of the other AUCs differed significantly.

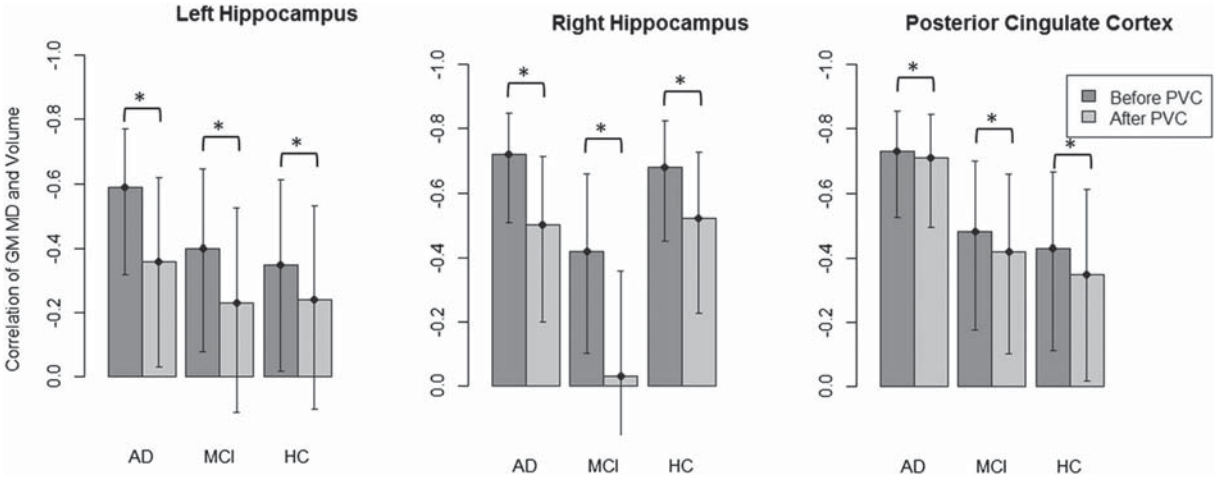


Fig. 3. Partial correlations of uncorrected and corrected GM MD values with GM volume, controlled for MRI protocol, age, gender and years of education.
The bars on the left always represent the MD values without PVC, bars on the right represent MD values with PVC.
*indicates significant differences of the partial correlations according to the Williams' test, error bars show confidence intervals (95%) of correlation coefficients.

and volume. However, other studies on GM MD in MCI that did not use PVE correction found no significant correlations between regional volumes and MD values (Eustache et al., 2016; Müller et al., 2005). Unfortunately, these studies did not report the correlation sizes because of nonsignificant *p*-values. Also, only zero order correlations were calculated without adjustment for age, gender or education. In our study, in the left hippocampus as well as in the PCC, the correlations between MD and volume were numerically - albeit not significantly - higher in AD subjects than in MCI subjects and higher in MCI than in HC subjects. This is in line with the findings that partial volume effects are more important in brains with stronger atrophy (Jeon et al., 2012). However, the correlations in the right hippocampus were lowest in MCI subjects. As this effect could not be reproduced in the repeated analyses using a different segmentation and normalization pipeline (Fig. S2 in the Supplementary material), we assume that it is an artefact of the segmentation and normalization algorithm (VBM 8).

This is further supported by the fact that in the mediation analysis following the approach recommended by Hayes (2013), volume was a significant mediator in the relationship between MD and diagnosis of AD vs. HC, regardless of PVC. In the left hippocampus, there was also a trend for this mediation on the group separation of MCI vs. HC. The findings of the mediation analysis, but also the strong correlation of GM MD and GM volume in healthy older subjects and subjects with MCI and AD suggest that changes in uncorrected MD values are largely explained by changes in brain volume in these populations. A strong correlation can be expected in MCI and AD, assuming that alterations of both MD and volume are caused by the same neurodegenerative process. However, the fact that the mediating effect of volume persisted after PVC had been applied could be due to brain atrophy that was already too advanced in these subjects to allow for a sufficient PVE correction. This would mean that MD value estimation would still partly be driven by CSF contamination.

Table 5
Pearson correlations of the imaging values (MD and volume) resulting from the FSL and SPM processing pipelines.

		Left hippocampus	Right hippocampus	PCC
MD	Without PVC	0.74**	0.77**	0.85**
	With PVC	0.54**	0.59**	0.79**
Volume		0.82**	0.82**	0.53**

** $p < 0.01$.

Table 4
Results of the mediation analysis, controlled for age, gender, years of education and MRI protocol.

Diagnosis	ROI	PVC	Direct Effect (c′) β [95% CI]	Indirect effect (a ∗ b) β [95% CI]
MCI vs. HC	Left hippocampus	With PVC	− 0.45 [− 1.17, 0.28]	− 0.23† [− 0.73, 0.002]
		Without PVC	− 0.53 [− 1.44, 0.38]	− 0.38† [− 1.09, 0.01]
	Right hippocampus	With PVC	− 0.46 [− 1.24, 0.31]	− 0.17 [− 0.56, 0.04]
		Without PVC	− 0.71 [− 1.16, 1.06]	− 0.21 [− 1.13, 0.31]
	Posterior cingulate cortex	With PVC	− 0.32 [− 1.43, 0.79]	− 0.22 [− 0.98,0.24]
		Without PVC	− 0.35 [− 1.45, 0.74]	− 0.22 [− 1.01, 0.33]
AD vs. HC	Left hippocampus	With PVC	− 0.47 [− 1.47, 0.54]	− 1.28* [− 2.43, − 0.44]
		Without PVC	− 0.68 [− 1.9, 0.54]	− 1.75*[− 3.26, − 0.63]
	Right hippocampus	With PVC	0.35 [− 0.55, 1.24]	− 1.44* [− 2.64, − 0.65]
		Without PVC	− 0.06 [− 1.22, 1.10]	− 1.74* [− 3.72, − 0.42]
	Posterior cingulate cortex	With PVC	− 0.69 [− 2.15, 0.76]	− 1.83* [− 3.32, − 0.66]
		Without PVC	− 1.04 [− 2.52, 0.44]	− 1.52* [− 2.95, − 0.32]
AD vs. MCI	Left hippocampus	With PVC	− 0.02 [− 0.65, 0.62]	− 0.51* [− 1.07, − 0.14]
		Without PVC	− 0.37 [− 1.18, 0.44]	− 0.74* [− 1.41, − 0.18]
	Right hippocampus	With PVC	0.05 [− 0.52, 0.61]	− 0.48* [− 0.97, − 0.15]
		Without PVC	− 0.27 [− 1.07, 0.54]	− 0.77* [− 1.54, − 0.1]
	Posterior cingulate cortex	With PVC	− 0.23 [− 1.52, 1.07]	− 1.61* [− 3.23, − 0.53]
		Without PVC	− 0.60 [− 1.9, 0.70]	− 1.3* [− 2.73, − 0.33]

* The 95% CI does not include zero.
† the corresponding 90% CI does not include zero.

Comparing the diagnostic accuracy of MD and volume, regional MD was not a better indicator for separating AD from HC than regional volume, as indicated by consistently larger AUCs for volume as a predictor. This finding contradicts some previous studies on regional gray matter MD and volume changes in MCI and AD subjects (Fellgiebel et al., 2006; Kantarci et al., 2005; Müller et al., 2007), but is in line with recent findings in an independent sample (Brueggen et al., 2015). As argued there, those studies that found hippocampal MD to be a better predictor for diagnosis or conversion than hippocampal volume did not control for gender, age and education. Also, the sample sizes were smaller than in this recent (Brueggen et al., 2015) and our current study. Considering the different ROIs, we found that the AUCs for both volume and MD of the left hippocampus were greater than those of the right hippocampus and of the PCC. This agrees with a meta-analysis that showed hippocampal asymmetry in MCI and AD cases (Shi et al., 2009).

From our findings, we would conclude that GM MD is not a more useful marker than volume of the hippocampus for detecting prodromal and clinical stages of AD. We showed that the changes in uncorrected GM MD values in MCI and AD patients largely reflect the macrostructural changes in regional brain volume. Although this resulted in a better diagnostic accuracy compared to the PVE corrected MD values, only PVE corrected GM MD values constitute a specific marker of microstructural changes over the course of AD. In contrast to later AD stages, however, one may expect that MD is a useful marker of a neurodegenerative process at very early disease stages that are not yet characterized by overt macrostructural brain atrophy, such as preclinical AD, i.e. cortical amyloidosis without clinical symptoms and limited GM atrophy (Dubois et al., 2014). Also, findings in healthy older people support the assumption that MD may be more sensitive in healthy individuals without significant brain atrophy. Carlesimo et al. (2010) could show that left hippocampal MD was correlated with cognitive performance in cognitively normal older people. Although not correcting for PVE, they observed that hippocampal MD predicted cognitive performance more sensitively than hippocampal volume. Subsequently, it would be interesting to analyze the relationship of PVE corrected hippocampal MD and amyloid accumulation in preclinical AD.

As a limitation, the MCI subjects of our sample were not uniquely amnestic MCI. Thus, our sample is more heterogeneous than the MCI subgroups in other studies (Cherubini et al., 2010; Fellgiebel et al., 2006; Müller et al., 2007; Scola et al., 2010). Second, in our study, we applied the previously validated CSF contamination model by Koo et al. (2009). However, if FLAIR DTI or DTI with more than one non-zero *b*-value had been available for the sample, DTI maps with suppressed CSF signal could have been obtained so that additional PVC would have been unnecessary (Kantarci et al., 2005; Salminen et al., 2016a). Future studies should assess the impact of different PVC methods on the MD estimates. The approach by Salminen et al. (2016a, 2016b) may be especially promising for clinical practice as it does not require additional scanning time.

In general, as the PVC method by Koo et al. directly uses the segmented maps resulting from VBM8, our results are potentially sensitive to the choice of segmentation and normalization algorithm. However, there was a high consistency of findings in PVE corrected MD data across segmentation/normalization approaches, underscoring the overall relevance of findings irrespective of the normalization or segmentation pipeline. Future studies should systematically evaluate the sensitivity of GM MD estimates to the choice of segmentation and normalization.

5. Conclusion

Evaluating the diagnostic utility of GM MD in AD and MCI, we found that the effects of MD are being overestimated without PVC when using the PVC proposed by Koo et al. (2009). Therefore, when comparing groups with different levels of atrophy, correction for PVE is indispensable. However, when comparing the diagnostic utility of corrected and uncorrected regional MD to regional volume, regional MD was not superior to volume in separating the potentially prodromal and the clinical stages of AD from matched healthy controls. Future studies need to test whether PVE-corrected MD may be more useful at the preclinical stage of AD.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary Material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2017.10.005.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113.

Berlot, R., Metzler-Baddeley, C., Jones, D.K., Sullivan, M.J.O., 2014. NeuroImage CSF contamination contributes to apparent microstructural alterations in mild cognitive impairment. NeuroImage 92, 27–35. http://dx.doi.org/10.1016/j.neuroimage.2014.01.031.

Brueggen, K., Dyrba, M., Barkhof, F., Hausner, L., Filippi, M., Nestor, P.J., Hauenstein, K., Klöppel, S., Grothe, M.J., Kasper, E., Teipel, S.J., 2015. 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. 48, 197–204. http://dx.doi.org/10.3233/JAD-150063.

Carlesimo, G.A., Cherubini, A., Caltagirone, C., Spalletta, G., 2010. Hippocampal mean diffusivity and memory in healthy elderly individuals A cross-sectional study. Neurology 74, 194–200.

Cherubini, A., Peran, P., Spoletini, I., Di Paola, M., Di Iulio, F., Hagberg, G.E., Sancesario, G., Gianni, W., Bossu, P., Caltagirone, C., Sabatini, U., Spalletta, G., 2010. Combined volumetry and DTI in subcortical structures of mild cognitive impairment and Alzheimer's disease patients. J. Alzheimers Dis. 19, 1273–1282. http://dx.doi.org/10.3233/JAD-2010-091186.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

Douaud, G., Menke, R.A.L., Gass, A., Monsch, A.U., Rao, A., Whitcher, B., Zamboni, G., Matthews, P.M., Sollberger, M., Smith, S., 2013. Brain microstructure reveals early abnormalities more than two years prior to clinical progression from mild cognitive impairment to Alzheimer's disease. J. Neurosci. 33, 2147–2155. http://dx.doi.org/10.1523/JNEUROSCI.4437-12.2013.

Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., Blennow, K., DeKosky, S.T., Gauthier, S., Selkoe, D., Bateman, R., et al., 2014. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. 13, 614–629.

Eustache, P., Nemmi, F., Saint-Aubert, L., Pariente, J., Peran, P., 2016. Multimodal magnetic resonance imaging in Alzheimer's disease patients at prodromal stage. J. Alzheimers Dis. 50, 1035–1050. http://dx.doi.org/10.3233/JAD-150353.

Fellgiebel, A., Wille, P., Müller, M.J., Winterer, G., Scheurich, A., Vucurevic, G., Schmidt, L.G., Stoeter, P., 2004. Ultrastructural hippocampal and white matter alterations in mild cognitive impairment: a diffusion tensor imaging study. Dement. Geriatr. Cogn. Disord. 18, 101–108. http://dx.doi.org/10.1159/000077817.

Fellgiebel, A., Dellani, P.R., Greverus, D., Scheurich, A., Stoeter, P., Müller, M.J., 2006. Predicting conversion to dementia in mild cognitive impairment by volumetric and diffusivity measurements of the hippocampus. Psychiatry Res. 146, 283–287. http://dx.doi.org/10.1016/j.pscychresns.2006.01.006.

Grothe, M., Heinsen, H., Teipel, S., 2013. Longitudinal measures of cholinergic forebrain atrophy in the transition from healthy aging to Alzheimer's disease. Neurobiol. Aging 34, 1210–1220. http://dx.doi.org/10.1016/j.neurobiolaging.2012.10.018.

Hayes, A.F., 2013. Introduction to Mediation, Moderation, and Conditional Process Analysis: a Regression-Based Approach. Guilford Press.

Hutton, C., Draganski, B., Ashburner, J., Weiskopf, N., 2009. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. NeuroImage 48, 371–380. http://dx.doi.org/10.1016/j.neuroimage.2009.06.043.

Jeon, T., Mishra, V., Uh, J., Weiner, M., Hatanpaa, K.J., White 3rd, C.L., Zhao, Y.D., Lu, H., Diaz-Arrastia, R., Huang, H., 2012. Regional changes of cortical mean diffusivities with aging after correction of partial volume effects. NeuroImage 62, 1705–1716. http://dx.doi.org/10.1016/j.neuroimage.2012.05.082.

Kantarci, K., CR, J.J., Xu, Y., Campeau, N., O'Brien, P., Smith, G., Ivnik, R., Boeve, B., Kokmen, E., Tangalos, E., Petersen, R., 2001. Mild cognitive impairment and Alzheimer disease: regional diffusivity of water. Radiology 101–107. http://dx.doi.org/10.1148/radiology.219.1.r01ap14101.

Kantarci, K., Petersen, R.C., Boeve, B.F., Knopman, D.S., Weigand, S.D., O'Brien, P.C.,

Shiung, M.M., Smith, G.E., Ivnik, R.J., Tangalos, E.G., et al., 2005. DWI predicts future progression to Alzheimer disease in amnestic mild cognitive impairment. Neurology 64, 902–904.

Kantarci, K., Avula, R., Senjem, M.L., Samikoglu, A.R., Zhang, B., Weigand, S.D., Przybelski, S.A., Edmonson, H.A., Vemuri, P., Knopman, D.S., Ferman, T.J., Boeve, B.F., Petersen, R.C., Jack, C.R., 2010. Dementia with Lewy bodies and Alzheimer disease: neurodegenerative patterns characterized by DTI. Neurology 74, 1814–1821. http://dx.doi.org/10.1212/WNL.0b013e3181e0f7cf.

Koo, B.B., Hua, N., Choi, C.H., Ronen, I., Lee, J.M., Kim, D.S., 2009. A framework to analyze partial volume effect on gray matter mean diffusivity measurements. NeuroImage 44, 136–144. http://dx.doi.org/10.1016/j.neuroimage.2008.07.064.

Ly, M., Canu, E., Xu, G., Oh, J., Mclaren, D.G., Dowling, N.M., Alexander, A.L., Sager, M.A., Johnson, S.C., Bendlin, B.B., 2014. Midlife Measurements of White Matter Microstructure Predict Subsequent Regional White Matter Atrophy in Healthy Adults. 2054. pp. 2044–2054. http://dx.doi.org/10.1002/hbm.22311.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dement. J. Alzheimer's Assoc. 7, 263–269. http://dx.doi.org/10.1016/j.jalz.2011.03.005.

Metzler-Baddeley, C., Sullivan, M.J.O., Bells, S., Pasternak, O., Jones, D.K., 2012. NeuroImage how and how not to correct for CSF-contamination in diffusion MRI. NeuroImage 59, 1394–1403. http://dx.doi.org/10.1016/j.neuroimage.2011.08.043.

Morris, J.C., Mohs, R.C., Rogers, H., Fillenbaum, G., Heyman, A., 1987. Consortium to establish a registry for Alzheimer's disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. Psychopharmacol. Bull. 24, 641–652.

Müller, M.J., Greverus, D., Dellani, P.R., Weibrich, C., Wille, P.R., Scheurich, A., Stoeter, P., Fellgiebel, A., 2005. Functional implications of hippocampal volume and diffusivity in mild cognitive impairment. NeuroImage 28, 1033–1042. http://dx.doi.org/10.1016/j.neuroimage.2005.06.029.

Müller, M.J., Greverus, D., Weibrich, C., Dellani, P.R., Scheurich, A., Stoeter, P., Fellgiebel, A., 2007. Diagnostic utility of hippocampal size and mean diffusivity in amnestic MCI. Neurobiol. Aging 28, 398–403.

Pasternak, O., Sochen, N., Gur, Y., Intrator, N., Assaf, Y., 2009. Free water elimination and mapping from diffusion MRI. Magn. Reson. Med. 62, 717–730. http://dx.doi.org/10.1002/mrm.22055.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56, 303–308.

R Development Core Team, 2008. R: a Language and Environment for Statistical Computing.

Rose, S.E., Janke, A.L., Chalk, J.B., 2008. Gray and white matter changes in Alzheimer's disease: a diffusion tensor imaging study. J. Magn. Reson. Imaging 27, 20–26. http://dx.doi.org/10.1002/jmri.21231.

Salminen, L.E., Conturo, T.E., Bolzenius, J.D., Cabeen, R.P., Akbudak, E., Paul, R.H., 2016a. Reducing CSF Partial Volume Effects to Enhance Diffusion Tensor Imaging Metrics of Brain Microstructure. 18. pp. 5–20.

Salminen, L.E., Conturo, T.E., Laidlaw, D.H., Cabeen, R.P., Akbudak, E., Lane, E.M., Heaps, J.M., Bolzenius, J.D., Baker, L.M., Cooley, S., Scott, S., Cagle, L.M., Phillips, S., Paul, R.H., 2016b. Regional age differences in gray matter diffusivity among healthy older adults. Brain Imaging Behav. 10, 203–211. http://dx.doi.org/10.1007/s11682-015-9383-7.

Scola, E., Bozzali, M., Agosta, F., Magnani, G., Franceschi, M., Sormani, M.P., Cercignani, M., Pagani, E., Falautano, M., Filippi, M., Falini, A., 2010. A diffusion tensor MRI study of patients with MCI and AD with a 2-year clinical follow-up. J. Neurol. Neurosurg. Psychiatry 81, 798–805. http://dx.doi.org/10.1136/jnnp.2009.189639.

Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: meta-analyses of MRI studies. Hippocampus 19, 1055–1064. http://dx.doi.org/10.1002/hipo.20573.

van Uden, I.W.M., Tuladhar, A.M., van der Holst, H.M., van Leijsen, E.M.C., van Norden, A.G.W., de Laat, K.F., Rutten-Jacobs, L.C.A., Norris, D.G., Claassen, J.A.H.R., van Dijk, E.J., Kessels, R.P.C., de Leeuw, F.-E., 2016. Diffusion tensor imaging of the hippocampus predicts the risk of dementia; the RUN DMC study. Hum. Brain Mapp. 37, 327–337. http://dx.doi.org/10.1002/hbm.23029.

Uluğ, A.M., Moore, D.F., Bojko, A.S., Zimmerman, R.D., 1999. Clinical use of diffusion-tensor imaging for diseases causing neuronal and axonal damage. Am. J. Neuroradiol. 20, 1044–1048.

Weston, P.S.J., Simpson, I.J.A., Ryan, N.S., Ourselin, S., Fox, N.C., 2015. Diffusion imaging changes in grey matter in Alzheimer's disease: a potential marker of early neurodegeneration. Alzheimers Res. Ther. 7, 47. http://dx.doi.org/10.1186/s13195-015-0132-3.

Williams, E.J., 1959. Regression Analysis. WILEY, New York.

Supplement to Paper

1. Methods

Table S1. Acquisition parameters of DTI and T1-weighted MRI scans.

	No. of subjects	Voxel Size (mm)	Repetition time (ms)	Echo time (ms)	Flip angle	Slice thickness (mm)	Inversion Time (ms)	Field of view (mm)	b-values (s/mm²)	Gradients	Parallel Acquisition Technique (PAT)	Repetitions	Gap (mm)	Acquisition time (min)
MPRAGE (old)	18	1x1x1	1900	2.52	9°	1	900	250	-	-	2	1	0.5 (50%)	4:18
MPRAGE (new)	21	1x1x1	2500	4.82	7°	1	1100	256	-	-	-	1	0.5 (50%)	9:20
DTI (old)	18	2x2x2	8200	93	-	2	-	250	1000	20	3	3	0.4 (20%)	9:19
DTI (new)	21	2x2x2	12700	81	-	2	-	256	1000	30	3	2	0	14:11

2. Results

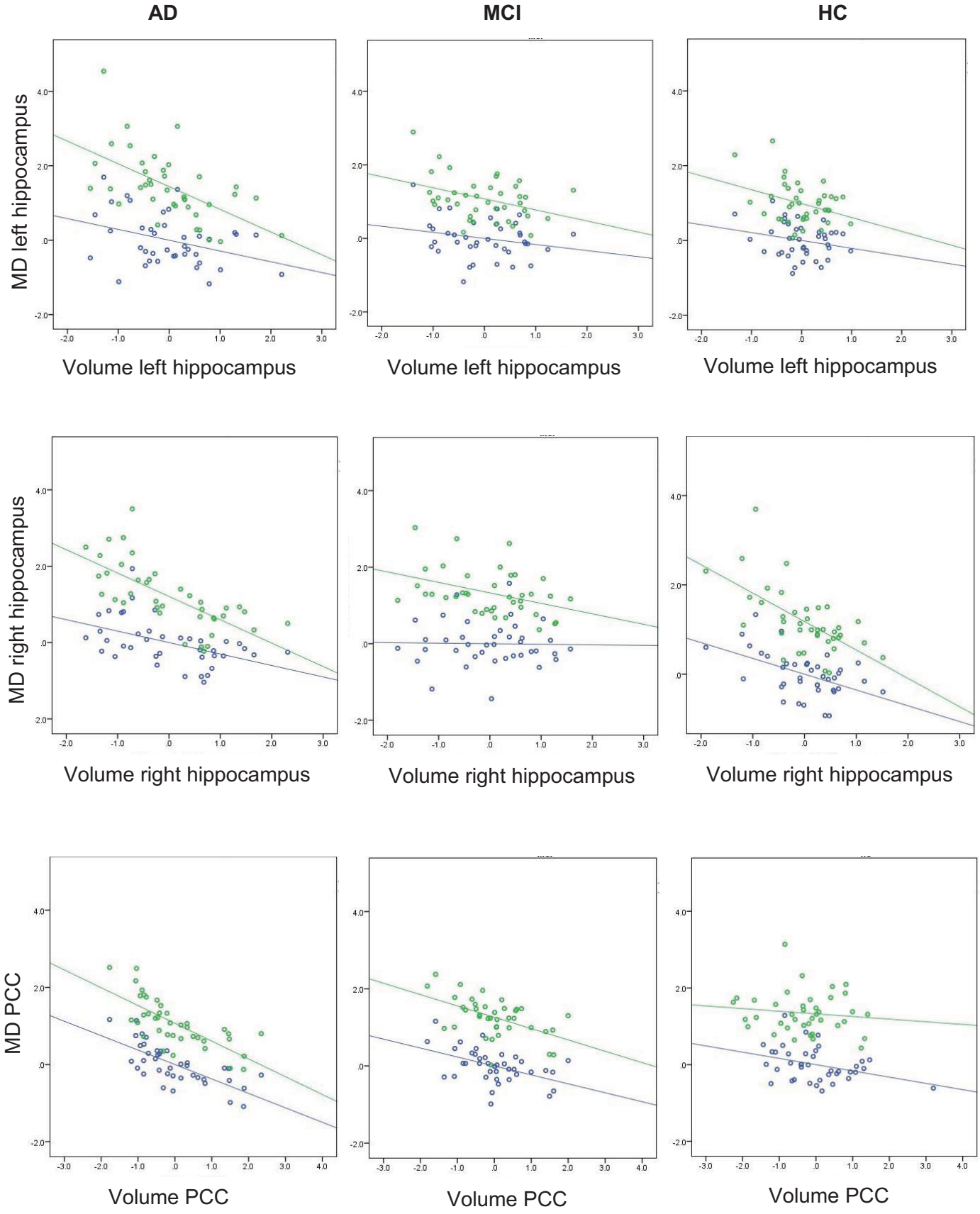


Figure S1. Partial correlation plots for each ROI and diagnosis subgroup. In these plots, "volume" is short for "standardized residuals from regressing volume on covariates (age, gender, years of education and MRI protocol) and "MD" is short for "standardized residuals from regressing MD on covariates".

3. Results after Segmentation and Normalization using FSL

Quality control revealed incorrect normalization (FNIRT) in 3 subjects, so that sample size for the following results is N=38 for each diagnostic subgroup.

Table S2. ROI Analysis of uncorrected and corrected MD values and volumes resulting from segmentation and normalization using FSL. Results are controlled for age, gender, years of education and MRI protocol (N=38 for each subgroup)

		Mean (SD) MD in $mm^2/s \times 10^{-3}$ Volume in ml			ANCOVA			
		AD	MCI	HC	F	*p* (unc.)	partial eta^2	Post hoc t-tests, p<.05
Left Hippocampus	MD without PVC	1.00 (.12)	.89 (.11)	.85 (.09)	20.01	<.001	.27	AD>MCI, AD>HC
	MD with PVC	.76 (.11)	.72 (08)	.72 (.09)	2.87	.061	.05	
	Volume	3.31 (.66)	4.00 (.60)	4.33 (0.46)	34.59	<.001	.39	HC>MCI>AD
Right Hippocampus	MD without PVC	1.07 (.15)	.93 (.11)	.89 (.09)	22.81	<.001	.30	AD>MCI AD>HC
	MD with PVC	.81 (.12)	.77 (.08)	.75 (.08)	3.8	<.05	.07	
	Volume	2.88 (.68)	3.60 (.63)	3.85 (0.53)	29.11	<.001	.35	AD<HC AD<MCI
PCC	MD without PVC	1.00 (.13)	.90 (.08)	.88 (.08)	14.92	<.001	.22	AD>MCI AD>HC
	MD with PVC	.75 (.12)	.69 (.07)	.69 (.08)	6	<.01	.10	AD>HC AD>MCI
	Volume	13.54 (1.45)	14.03 (1.01)	14.13 (1.08)	3.27	<.05	.06	AD<HC AD<MCI[†]

[†] post hoc comparisons revealed a trend (p<.01).

Table S3. Results of the logistic regression analysis using the volume and MD values that resulted from an alternative segmentation and normalization algorithm (FSL). Covariates were age, gender, years of education and MRI protocol (N=38 for each subgroup).

Diagnosis	ROI	PVC	MD		Volume	
			β	AUC [95% CI]	β	AUC [95% CI]
MCI vs. HC	Left Hippocampus	With PVC	.01	.62 [.49, .74]	-.70	.67* [.54, .79]
		Without PVC	.39	.63 [.50, 75]		
	Right Hippocampus	With PVC	.06	.61 [.48, .74]	-.41	.63 [.49, .75]
		Without PVC	.41	.64 [.52, .77]		
	Posterior Cingulate Cortex	With PVC	.02	.62 [.49, .74]	-.15	.63 [.49, .74]
		Without PVC	.15	.62 [.49, .74]		
AD vs. HC	Left Hippocampus	With PVC	.65*	.72 [.60, .83]	-3.34***	.93 [.86, .98]
		Without PVC	2.09***	.88 [.80, .95]		
	Right Hippocampus	With PVC	.74*	.73 [.61, .84]	-2.88***	.92 [.84, .98]
		Without PVC	2.09***	.87 [.79, .95]		
	Posterior Cingulate Cortex	With PVC	.92**	.73 [.62, .84]	.62*	.71 [.59, .82]
		Without PVC	1.61***	.82 [.71, .91]		

Standardized regression weights (β) resulting from logistic regression analysis. The confidence intervals of the area under the receiver operator characteristic curve (AUC) were calculated using bootstrapping.
Asterisks (*) indicate significance of the standardized regression weights (*** p<.001, **p<.01, *p<.05, † p<.1).

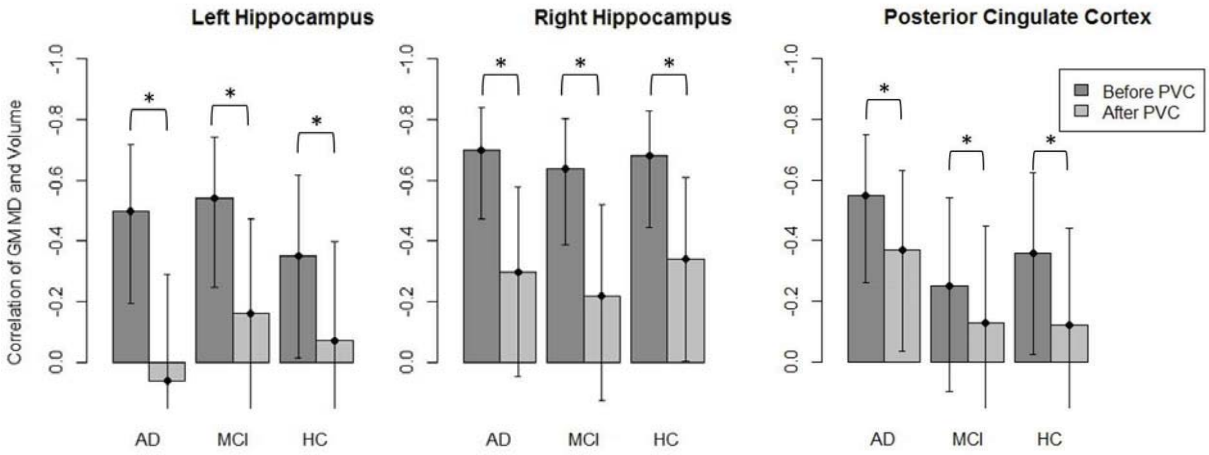


Figure S2. Partial correlations of uncorrected and corrected GM MD values with GM volume using imaging measures resulting from an alternative segmentation and normalization algorithm (FSL). The partial correlations were controlled for MRI protocol, age, gender and years of education (N=38 for each subgroup).
The bars on the left always represent the MD values without PVC, bars on the right represent MD values with PVC
*** indicates significant differences of the partial correlations according to the Williams' test, error bars show confidence intervals (95%) of correlation coefficients.**

NeuroImage: Clinical 14 (2017) 183–194

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI

Stefan J. Teipel[a,b,*], Alexandra Wohlert[a], Coraline Metzger[c], Timo Grimmer[d], Christian Sorg[e], Michael Ewers[f], Eva Meisenzahl[g], Stefan Klöppel[h,k], Viola Borchardt[i,j], Michel J. Grothe[b], Martin Walter[i,j], Martin Dyrba[b]

[a]Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany
[b]DZNE, German Center for Neurodegenerative Diseases, Rostock, Germany
[c]Institute of Cognitive Neurology and Dementia Research (IKND), Department of Psychiatry and Psychotherapy, Otto von Guericke University, Germany and German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
[d]Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technische Universität München, Munich, Germany
[e]Department of Neuroradiology of Klinikum rechts der Isar, Technische Universität München, Department of Psychiatry of Klinikum rechts der Isar, TUM-Neuroimaging Center, Einsteinstr. 1, 81675 Munich, Germany
[f]Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians-Universität LMU, Munich, Germany
[g]Department of Psychiatry, Klinikum der Universität München, Ludwig-Maximilians-Universität LMU, Munich, Germany
[h]Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Faculty of Medicine, University of Freiburg, Germany
[i]Leibniz Institute for Neurobiology, Magdeburg, Germany
[j]Department of Psychiatry, University Tübingen, Germany
[k]University Hospital of Old Age Psychiatry, Bern, Switzerland

ARTICLE INFO

Article history:
Received 6 June 2016
Received in revised form 30 November 2016
Accepted 17 January 2017
Available online 18 January 2017

ABSTRACT

Background: In monocentric studies, patients with mild cognitive impairment (MCI) and Alzheimer's disease (AD) dementia exhibited alterations of functional cortical connectivity in resting-state functional MRI (rs-fMRI) analyses. Multicenter studies provide access to large sample sizes, but rs-fMRI may be particularly sensitive to multiscanner effects.
Methods: We used data from five centers of the "German resting-state initiative for diagnostic biomarkers" (psymri.org), comprising 367 cases, including AD patients, MCI patients and healthy older controls, to assess the influence of the distributed acquisition on the group effects. We calculated accuracy of group discrimination based on whole brain functional connectivity of the posterior cingulate cortex (PCC) using pooled samples as well as second-level analyses across site-specific group contrast maps.
Results: We found decreased functional connectivity in AD patients vs. controls, including clusters in the precuneus, inferior parietal cortex, lateral temporal cortex and medial prefrontal cortex. MCI subjects showed spatially similar, but less pronounced, differences in PCC connectivity when compared to controls. Group discrimination accuracy for AD vs. controls (MCI vs. controls) in the test data was below 76% (72%) based on the pooled analysis, and even lower based on the second level analysis stratified according to scanner. Only a subset of quality measures was useful to detect relevant scanner effects.
Conclusions: Multicenter rs-fMRI analysis needs to employ strict quality measures, including visual inspection of all the data, to avoid seriously confounded group effects. While pending further confirmation in biomarker stratified samples, these findings suggest that multicenter acquisition limits the use of rs-fMRI in AD and MCI diagnosis.

1. Introduction

Criteria for prodromal Alzheimer's disease (AD) (Albert et al., 2011; Dubois et al., 2010; Dubois et al., 2007; Dubois et al., 2014) and AD dementia (McKhann et al., 2011) diagnosis include structural imaging markers, such as MRI based hippocampus volumetry, molecular imaging markers, such as amyloid PET, and functional imaging markers, such as ^{18}FDG-PET. All these imaging markers have already been evaluated in large multicenter cohorts, such as ADNI, EDSD, NEST-DD and others (Doraiswamy et al., 2014; Herholz, 2010; Kilimann et al., 2014; Risacher et al., 2009).

Particularly, FDG-PET has proven a precise predictor of imminent conversion from mild cognitive impairment (MCI) to AD dementia (Ito et al., 2015). At the same time, PET imaging is relatively expensive

* Corresponding author at: Department of Psychosomatic Medicine, University Medicine Rostock, DZNE, Gehlsheimer Str. 20, 18147 Rostock, Germany.
E-mail address: stefan.teipel@med.uni-rostock.de (S.J. Teipel).

http://dx.doi.org/10.1016/j.nicl.2017.01.018

and availability of PET scanners is limited. Resting state fMRI (rs-fMRI) has been discussed as a functional imaging alternative for ^{18}FDG-PET (Teipel et al., 2015). Decline of default mode network connectivity, a brain network encompassing key regions of AD pathology such as posterior cingulate, precuneus, inferior parietal lobes, prefrontal cortex and medial temporal lobes (Fox et al., 2005), has been shown in AD dementia and MCI patients compared to age matched controls in a range of studies (Chhatwal et al., 2013; Greicius et al., 2004; Thomas et al., 2014). Results on diagnostic accuracies are mixed, ranging from 62% to >90% group separation of MCI or AD dementia cases from healthy control cases in monocenter studies (Dyrba et al., 2015b; Koch et al., 2012). Such variation across studies likely not only reflects differences in the cohorts, but also variation in acquisition parameters of rs-fMRI sequences between studies. High variability of group discrimination across sites, however, would severely diminish the value of rs-fMRI as an imaging biomarker of AD.

Multicenter studies in healthy people revealed high variability of task related functional MRI image properties, such as transient signals, smoothness and the shape of the hemodynamic response function, even when multicenter data stemmed from the same brand and model of scanners (Zou et al., 2005). Consistent with these findings, test-retest reliability studies of rs-fMRI suggest high intra-individual variability of resting state connectivity even in healthy people repeatedly scanned at the same scanner (Chen et al., 2015; Jovicich et al., 2016; Lin et al., 2015; Meindl et al., 2010; Orban et al., 2015; Shirer et al., 2015), including long-term evaluation after more than 12 months (Blautzik et al., 2013; Chou et al., 2012; Guo et al., 2012). Multiscanner evaluation suggests high variability of signal-to-noise and contrast-to-noise ratios, particularly when using field strengths of 3T and higher (Jovicich et al., 2016; Lin et al., 2015; Magnotta et al., 2006). In an explicit linear model, center accounted for a large amount of variance across voxel-wise resting state connectivity (Suckling et al., 2012).

Several rs-fMRI multicenter studies have investigated alterations of functional connectivity in AD and other neuropsychiatric conditions, but without consideration of multiscanner effects (Esslinger et al., 2011) even though protocols differed between sites in some studies (Chhatwal et al., 2013; Demertzi et al., 2015; Martucci et al., 2015; Sripada et al., 2014; Thomas et al., 2014). Some of these studies used the same scanner type across sites (Demertzi et al., 2015; Esslinger et al., 2011; Thomas et al., 2014), but some did not (Chhatwal et al., 2013; Martucci et al., 2015; Sripada et al., 2014).

Several studies reported techniques to reduce inter-scanner variability, mostly, however, in data from healthy people. One study probed a wide range of processing steps to reduce test-retest variability (Shirer et al., 2015). Another study comparing different connectivity metrics found most stable results for cross-correlation as compared to cross-coherence and partial cross-correlation (Fiecas et al., 2013). Two studies in healthy adults and young people at risk of psychosis, respectively, used scanner as a covariate in a second-level linear ANOVA model (Anticevic et al., 2015; Biswal et al., 2010), another study in healthy adults used conjunction analysis across scanners (Long et al., 2008). Only one previous study explicitly modelled center effects in healthy older people and MCI cases using a meta-analysis of between group effects across four different cohorts (Tam et al., 2015).

Here, we used rs-fMRI data of people with AD dementia, MCI and healthy older controls from the "German resting-state initiative for diagnostic biomarkers" (www.psymri.org) collected at five sites to compare previously employed measures of scan quality across sites (Jenkinson et al., 2002; Yan et al., 2013a), determine the effect of multicenter acquisition on between group effects, and assess diagnostic accuracies from different univariate analysis approaches. We expected to find large heterogeneity of between group effects that would likely impair the use of multicenter rs-fMRI data as diagnostic biomarker for AD. We used the widely established structural measure of hippocampus volume that has been found to be stable against multicenter effects (Ewers et al., 2006) as an internal benchmark for the functional connectivity metric.

Table 1
Demographic characteristics, all sites

	AD	MCI	Controls
No. cases (women)[1]	84 (46)	115 (59)	151 (82)
Age (SD) [years][2]	72.0 (9.0)	72.6 (8.0)	69.0 (7.8)
MMSE (SD), number[3]	22.4 (4.4), 84	26.7 (1.8), 115	28.9 (1.0) 115
MoCA (SD), number[4]	–	22.7 (3.0), 22	26.4 (2.1), 19
education (SD) [years][5]	10.9 (2.4)	12.4 (3.3)	12.9 (3.1)

MMSE – Mini Mental State Examination (Folstein et al., 1975)
MoCA – Montreal Cognitive Assessment (Nasreddine et al., 2005)
[1] Not significantly different between groups, $\chi^2 = 0.315$, 2 df, $p = 0.85$.
[2] Significantly different between groups, $F(2, 347) = 7.5$, $p < 0.001$.
[3] Significantly different between groups, Kruskal Wallis Test, $p < 0.001$.
[4] Significantly different between groups, Kruskal Wallis Test, $p < 0.001$.
[5] Significantly different between groups, $F(2, 323) = 11.4$, $p < 0.001$.

2. Material and methods

The original data set consisted of 367 rs-fMRI scans that have been retrieved retrospectively from five sites within the framework of the "German resting-state initiative for diagnostic biomarkers" (www.psymri.org). After a first round of visual quality check, 350 rs-fMRI data were retained, whereas 17 scans were dropped due to severe problems with scan quality, incomplete scans or scans covering only parts of the brain. From the remaining 350 scans, all 100 scans from one site (site V) were rated as borderline quality due to severe susceptibility effects and subsequent analyses were conducted both with and without the scans from this site. Distribution of demographic characteristics of participants across sites is summarized in Tables 1 and 2; the number of participants per scanner is shown in supplementary table 1.

The retained data included scans from 84 patients with clinically probable AD according to NINCDS-ADRCA criteria (McKhann et al., 1984), 115 individuals fulfilling Mayo criteria of amnestic MCI (Petersen et al., 1999) and 151 healthy elderly control individuals. All participants were free of any significant neurological, psychiatric, or medical condition (except for AD or MCI in patients), in particular cerebrovascular apoplexy, vascular dementia, depression, or subclinical hypothyroidism, as well as substance abuse. Healthy controls were required to have no cognitive complaints and scored within one standard deviation of the age and education adjusted norm in all subtests of the Consortium to Establish a Registry of Alzheimer's Disease (CERAD) cognitive battery (Morris et al., 1989).

Written informed consent was provided by all subjects, or their representatives. The study was approved by local ethics committees at each of the participating centers, and has been conducted in accord with the Helsinki Declaration of 1975.

Table 2
Demographic characteristics, one site excluded.

	AD	MCI	Controls
No. cases (women)[1]	53 (31)	79 (43)	118 (61)
Age (SD) [years][2]	72.4 (8.8)	74.8 (6.0)	70.4 (6.2)
MMSE (SD), number[3]	22.5 (4.4), 53	26.5 (1.8), 79	28.8 (1.0) 97
MoCA (SD), number[4]	–	22.7 (3.0), 22	26.4 (2.1), 19
education (SD) [years][5]	11.4 (2.1)	13.0 (3.4)	13.6 (3.1)

MMSE – Mini Mental State Examination (Folstein et al., 1975)
MoCA – Montreal Cognitive Assessment (Nasreddine et al., 2005)
[1] Not significantly different between groups, $\chi^2 = 0.689$, 2 df, $p = 0.71$.
[2] Significantly different between groups, $F(2, 247) = 9.8$, $p < 0.001$.
[3] Significantly different between groups, Kruskal Wallis Test, $p < 0.001$.
[4] Significantly different between groups, Kruskal Wallis Test, $p < 0.001$.
[5] Significantly different between groups, $F(2, 246) = 9.73$, $p < 0.001$.

2.1. Imaging and data acquisition

Data were obtained from five different 3.0 Tesla MRI scanners. Acquisition parameters for the rs-fMRI sequences are given in Table 3. In one center (site I), the subjects were instructed to keep their eyes open, whereas in the remaining centers (sites II-V) all subjects were requested to close their eyes, relax, but not to fall asleep. Functional MRI was based on echo-planar imaging using scan durations between 6 and 9 min for the rs-fMRI sequence. The number of acquired time points was between 120 and 240 with a voxel size ranging from $2 \times 2 \times 2.6$ up to $3.28 \times 3.28 \times 4.4$ mm^3 (Table 3). Anatomical scans were obtained from all scanners with an isotropic resolution of 1 mm^3 during the same session.

2.2. MR processing

The **anatomical T_1-weighted image** for each participant was segmented into gray matter, white matter, and cerebrospinal fluid (CSF) partitions of 1.5 mm isotropic voxel-size using the tissue prior free segmentation routine of the VBM8-toolbox (Gaser et al., 1999) that extends Statistical Parametric Mapping (SPM8) (Friston et al., 2007). The Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm (Ashburner, 2007) was applied to normalize the T_1-weighted gray matter and white matter partitions to the Montreal Neurological Institute (MNI) reference coordinate system using the default brain template included in VBM8. Individual flow-fields resulting from the DARTEL registration to the reference template were used to warp the gray matter segments. Voxel values of the warped gray matter segments were only modulated for the non-linear component of the deformation field, thus accounting at this step for differences in head size which are modeled by the affine component of the transformation.

Functional MRI data processing was carried out using Data Processing Assistant for Resting-State fMRI (DPARSF 3.2) (Chao-Gan and Yu-Feng, 2010), considering the recommendations from a recent systematic evaluation of processing alternatives (Shirer et al., 2015). After the removal of the first six images to account for gradient field stabilization, the rs-fMRI data was slice time corrected and realigned to the temporal mean image. Slice time correction addresses the problem that, for functional MRI, the 3D image of one time point is typically obtained by acquiring a series of 2D slices, with each slice being acquired one after another within the full period of one repetition time, for instance three seconds (Table 3). Thus, different slices of one 3D image measure the brain activity at a slightly different moment in time (Sladky et al., 2011). Slice time correction compensates for phase shifts in the time series signal using the cardinal sine interpolation based on the fast Fourier transform (Sladky et al., 2011). Sladky et al. found that this correction step improved the stability of estimates and magnitude of effects obtained from event-related and block design paradigms in task-based functional MRI (Sladky et al., 2011). It is also commonly applied to rs-fMRI data for both approaches seed-based functional connectivity and independent component analysis (Dyrba et al., 2015b; Koch et al., 2012; Meindl et al., 2010; Power et al., 2014; Yan et al., 2013b), which assess the correlation or homogeneity of the time series signal of remote brain regions or voxels. Controversially, previous studies only found minor, non-significant effects of applying slice time correction to rs-fMRI data (Shirer et al., 2015; Wu et al., 2011). These observations may be due to the subsequent step of bandpass filtering, which eliminates high-frequency components of the data with a wavelength of less than ten seconds and, thus, reduces the influence of slight short-term inaccuracies. The anatomical T_1-weighted image for each participant was coregistered to the mean functional image. The deformation fields generated by DARTEL from the anatomical T_1-weighted images were used to project the functional scans from each subjects' native image space into the MNI reference space. We combined this step with the reslicing of all functional data to an isotropic resolution of 3

Table 3
Scanner characteristics.

Center	Model	Manufacturer	TR [s]	TE [s]	Flip angle [°]	Matrix size	Field of view [mm^3]	Number of volumes	Voxel size [mm^3]	Gap [mm]	Slice thickness [mm]	Spacing between slices [mm]	Slice acquisition order
I	TrioTim	Siemens	2.61	0.030	80		64 × 64 × 42	192 × 192 × 151	200		3 × 3 × 3.6	0.6	3
3.6	Interleaved, ascending												
II	Verio	Siemens	3	0.030	90		96 × 96 × 45	192 × 192 × 117	120		2 × 2 × 2.6	0.6	2
2.6	Contiguous, descending												
III	Verio	Siemens	2.58	0.030	80		64 × 64 × 47	224 × 224 × 165	180		3.5 × 3.5 × 3.5	0	3.5
3.5	Interleaved, ascending												
IV	Trio	Siemens	3	0.030	80		64 × 64 × 28	210 × 210 × 123	120		3.28 × 3.28 × 4.4	0.4	4
4.4	Interleaved, ascending												
V	Achieva	Philips	2	0.043	82		96 × 96 × 32	220 × 220 × 128	240		2.29 × 2.29 × 4	0	4
4	Interleaved, descending												

All centers used an echo-planar imaging (EPI) sequence with axial slice orientation.

mm. The subsequent nuisance regression included covariates of head movement (rotation, translation, and derivatives) and the mean time courses for the global brain signal, the white matter segment signal, and the CSF segment signal. Although global signal regression was found to introduce negative correlations (Murphy et al., 2009; Shirer et al., 2015), studies consistently reported that it effectively reduces the signal-to-noise ratio (Power et al., 2014; Shirer et al., 2015; Yan et al., 2013a). Recently, Shirer et al. evaluated the influence of global signal regression on group separation but only found a minor, non-significant effect (Shirer et al., 2015). Subsequently, the images were band-pass filtered using the frequency band 0.1–0.01 Hz and smoothed using a 6 mm full-width-at-half-maximum (FWHM) Gaussian kernel. Ventral posterior cingulate cortex (PCC) functional connectivity maps were calculated using a spherical seed with 4 mm radius, which was set at MNI position 0,-53,26 (Hedden et al., 2009). Finally, Pearson correlation coefficients of the signal time courses were adjusted to be normally distributed using Fisher's Z-transform (Fisher, 1915): $z = 0.5 \ln [(1 + r)/(1 - r)]$.

2.3. Extraction of hippocampus volumes

A mask for the hippocampus was obtained by manual delineation of the hippocampus in the reference template (Grothe et al., 2012) using the interactive software package Display (McConnell Brain Imaging Centre at the Montreal Neurological Institute) and a previously described protocol for segmentation of the medial temporal lobe (Pruessner et al., 2000). Individual gray matter volumes of the hippocampus were extracted automatically from the warped gray matter segments by summing up the modulated gray matter voxel values within hippocampus ROI in the reference space.

3. Statistics

3.1. Quality control measures for scanner effects

We compared previously employed scan characteristics across scanners and diagnostic groups, including:

- Framewise displacement (FD) – mean and percentage above threshold (0.5 mm) (Jenkinson et al., 2002; Power et al., 2012; Power et al., 2014; Yan et al., 2013a)
- Temporal signal-to-noise ratio (tSNR) (Marcus et al., 2013; Welvaert and Rosseel, 2013)
- Standardized DVARS – root mean square of change in signal intensity from one time point to the next (Power et al., 2012)
- Percentage of outlier voxels (Zuo et al., 2014)
- Foreground to background energy ratio (FBER) (Zuo et al., 2014)
- Fractional amplitude of low frequency fluctuations (fALFF) (Yan et al., 2013b).

Additionally, we compared the regional correlations between PCC and anterior medial prefrontal cortex (aMPFC) time courses between scanners and groups, based on spherical seed regions with a radius of 4 mm at MNI coordinates 0, −53, 26 (PCC) (Hedden et al., 2009) and −6, 52, −2 (aMPFC) (Andrews-Hanna et al., 2010).

To limit the number of measures, we decided not to use some previously employed measures (Zuo et al., 2014), such as entropy focus criterion (EFC) (Atkinson et al., 1997), image smoothness (IS) (Zuo et al., 2014), or ghost-to-signal ratio (GSR). The GSR needs manual interaction for the definition of the area of ghost artifacts in native subject space and was obsolete for the detection of poor scan quality as our scans underwent visual inspection. We excluded EFC and IS which target strong blurring, motion, and noise and become redundant when including tSNR, percentage of outlier voxels, FD, and FBER.

A description of the quality measures can be found in the supplementary material section.

3.2. Spatial pattern of group differences

We determined differences in voxel-wise correlations of PCC activity between AD patients and controls and between MCI patients and controls using two different univariate approaches to take scanner effects into account:

- First, we determined group differences using a fixed effects linear model with diagnosis and scanner as independent factors, henceforth referred to as *pooled analysis with scanner covariate*. Significant clusters were identified with at least 10 voxels passing the uncorrected threshold of $p < 0.01$.
- Secondly, we used a *second level analysis* with linear models of between group differences at the first level and a one-sample t-test of the between group effects across the 3 scanners for AD vs. control comparison and 5 scanners for the MCI vs. control comparison at the second level. Significant clusters were identified with at least 10 voxels passing the uncorrected threshold of $p < 0.01$.

Additionally, we assessed the spatial coherence of voxel-wise group differences between single scanners using conjunction analysis (Friston et al., 2005). Conjunction analysis resembles an ANOVA model for detecting group effects for more than two groups, but allows setting a threshold k to define the minimum number of effects, so that a second level group effect is considered to be present in a given voxel if a significant group difference had been found for at least k individual scanners (Friston et al., 2005). With our data, the value of k could range from 1, indicating an effect for at least one single scanner, to 5, indicating that a group effect must be present for each of the five scanners.

3.3. Accuracy of group discrimination

We defined regions of interest (ROI) as those brain regions that showed significant group differences in the voxel-based comparisons of AD or MCI and healthy control subjects. Specifically, we binarized the statistical maps thresholded at $p < 0.01$ as described above for each statistical approach (i.e., pooled analysis, and second level analysis) yielding 2 (statistical approach) × 2 (AD vs. controls and MCI vs. controls) = 4 different ROIs. For each of these ROIs, we extracted averaged Fisher's Z-transformed correlation coefficients. To this end, the individual voxel-wise correlation maps in MNI standard space were multiplied by the thresholded binary ROIs, and the voxel values within each ROI were averaged for each individual scan, yielding scalar markers as predictors in linear logistic regression analyses.

To obtain an estimate of the accuracy of group discrimination for each modality and analysis technique, we used block-wise cross validation with repeated random sampling, based on Gaussian-distributed random numbers generated in R. We repeatedly split the data set into 63.2% of training data and 36.8% of test data. For each of the repeatedly drawn training samples, the logistic regression parameters were estimated and subsequently applied to the remaining test data set. Classification accuracy, sensitivity, and specificity as well as area under the receiver operating characteristic curves were recorded for each test data set. The entire cross-validation process was iterated 1000 times to determine the variability of the estimates of accuracy across runs. We determined nonparametric bootstrap confidence intervals with the 2.5 and 97.5 percentiles defining the lower and upper limits of the confidence interval ((Efron and Tibshirani, 1993), Chapter 13). Logistic regression analysis was calculated in R, using function glm with the parameter 'family' = binary.

To define a benchmark for the effect size of group discrimination, we repeated the bootstrapped determination of the area under the receiver operating characteristic curves for the widely established measure of hippocampus volume, averaged across left and right hemispheres.

3.4. Scanner effects

We employed variance component analysis using libraries "nlme" and "ape" in R with the function "varcomp" to determine the effect of scanner on functional connectivity, with diagnosis as fixed effect covariate and scanner as random effect covariate. We determined the proportion of variance attributable to scanner relative to the variance attributable to error. Variances were scaled to sum to 1.

4. Results

4.1. Quality control measures for scanner effects

Framewise displacement showed comparable displacements across sites, both in mean values as well as in percentage of framewise displacement >0.5 mm. Similarly, the foreground-to-background energy ratio, the fractional amplitude of low frequency fluctuations in PCC, and the mean functional connectivity between PCC and anterior medial prefrontal cortex indicated no outlying center (Supplementary Fig. 1 to 5). When looking at single sites, differences between diagnostic groups showed a general trend in the expected direction that only occasionally reached statistical significance. For instance, cognitively impaired patients showed slightly more head motion than controls (Supplementary Fig. 1) and lower temporal signal-to-noise ratio (supplementary figure 6). Mean whole brain temporal signal-to-noise ratio, the mean percentage of outlier voxels, and standardized DVARS identified an outlier in the site V data, with significantly decreased tSNR and standardized DVARS, and increased number of outlier voxels in the healthy control group compared to the MCI and AD group (Supplementary Figs. 6 to 8), one-sided Wilcoxon tests, $p < 0.01$. Site II showed a significantly reduced tSNR compared to the other sites (Supplementary Fig. 6), two-sided t-test, $p < 0.001$; but this systematic bias was evenly distributed across all subject groups.

4.2. Spatial pattern of group differences

We found group differences between AD patients and controls and between MCI patients and controls both in the pooled data analysis as well as the second level analysis only at an uncorrected level of significance of $p < 0.01$, but no effects at an uncorrected p-value of 0.001. Functional connectivity of the PCC was smaller in AD and MCI cases compared to controls when the data of site V were removed from the analysis. Peak areas of group effects were located in the mid temporal cortex, anterior cingulum and inferior parietal cortex (including angular gyrus) for the AD vs. control comparison, and in the precuneus, middle cingulate cortex, insula cortex, fusiform gyrus and medial temporal lobes (including amygdala and parahippocampal cortex) for the MCI vs. controls comparison (Figs. 1 and 2). The conjunction analysis revealed small clusters in only few regions when setting the minimum number of effects to $k = 2$, i.e. when group effects were significant in data from at least two scanners (data not shown). For the MCI vs. controls comparison, no cluster survived when the number of effects k was >2. When the data of site V were included in the analysis, effects were in the opposite direction with larger functional connectivity in the AD and MCI cases compared to controls (data not shown).

4.3. Accuracy of group discrimination

These analyses were only conducted in the sample without including the data of site V. The distribution of MCI and control cases was relatively well balanced across the four sites. Since the distribution of AD cases was imbalanced across sites, the analyses for the AD vs. control comparisons were repeated across all sites and across the only two sites with a balanced number of AD cases and controls.

For the AD vs. controls comparison, mean AUCs in the test data ranged from 74% for the second level data to 82% for the pooled data, and accuracies ranged from 69% for the second level data to 76% for

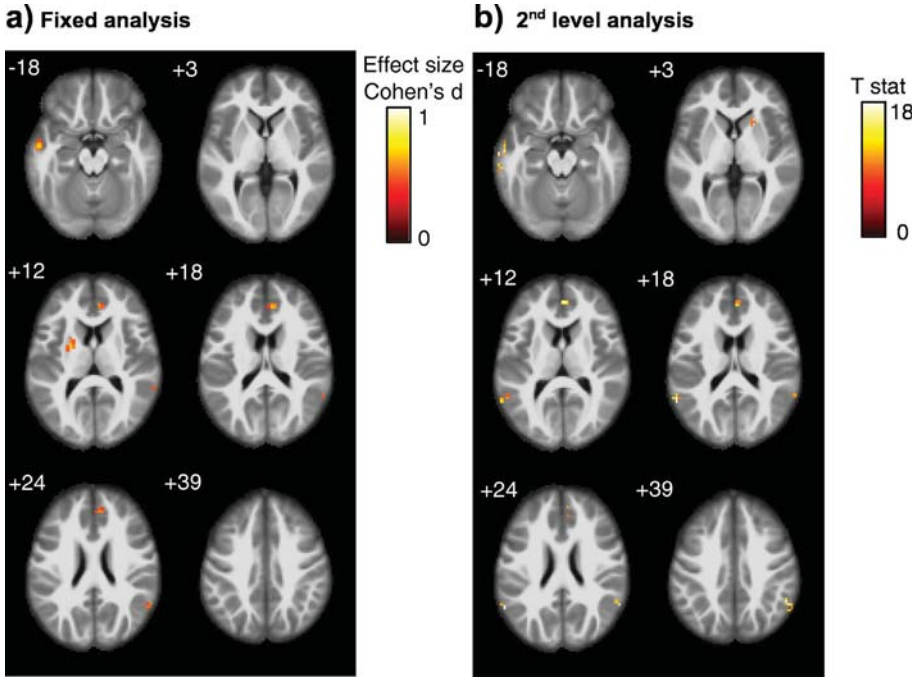


Fig. 1. AD vs. control comparison Group effects of PCC functional connectivity differences between AD patients and controls, using (Panel a) a fixed effect analysis pooling all scans across scanners with scanner as covariate, and (Panel b) a second level analysis with scanner as second level factor. Significant cluster of at least 10 voxel passing an uncorrected threshold of significance of $p < 0.01$, are projected onto an anatomical MRI scan in MNI space. Numbers in the upper left corner of each image slice indicate the MNI z-coordinate, i.e. the axial section in MNI space. Color bars represent color coding for Cohen's d effect size estimates (Cohen, 1977) for the pooled analysis, and T values for the second level analysis, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

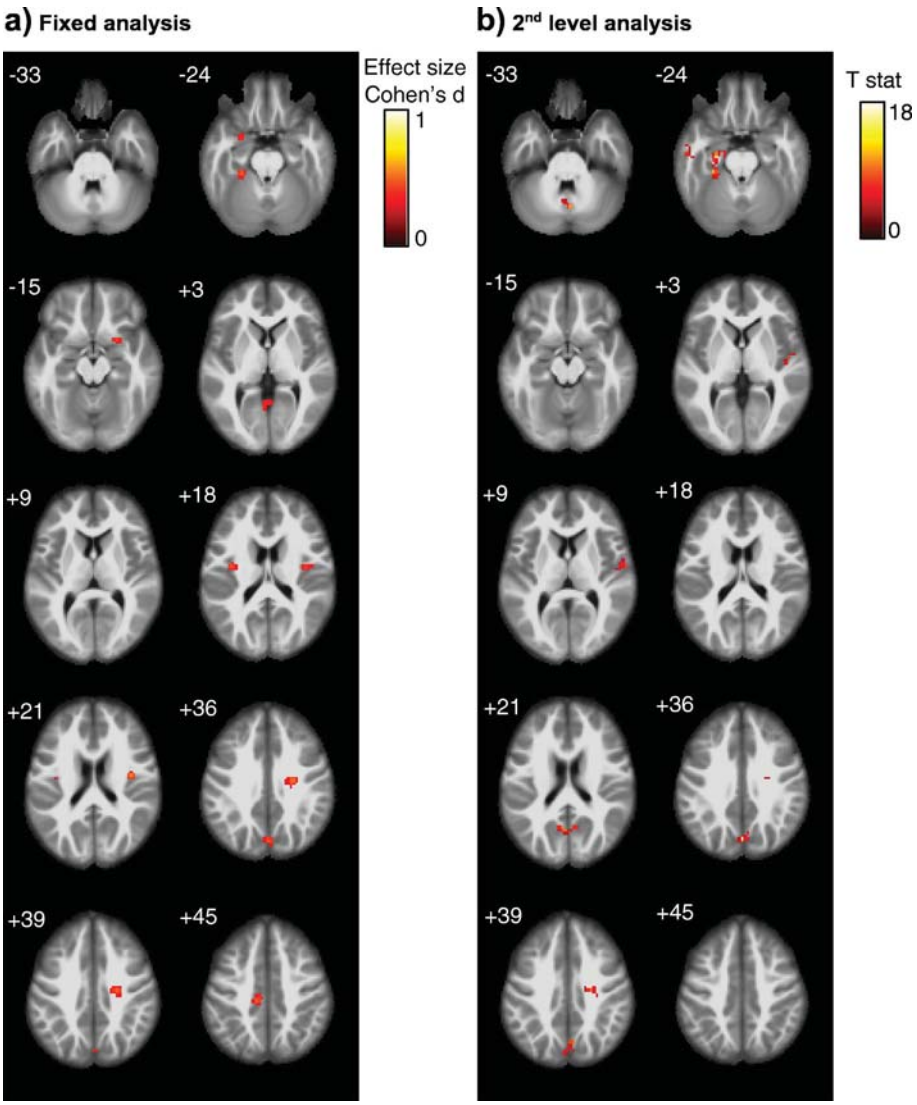


Fig. 2. MCI vs. control comparison Group effects of PCC functional connectivity differences between MCI patients and controls, using (Panel a) a fixed effect analysis pooling all scans across scanners with scanner as covariate, and (Panel b) a second level analysis with scanner as second level factor. Significant cluster of at least 10 voxel passing an uncorrected threshold of significance of $p < 0.01$, are projected onto an anatomical MRI scan in MNI space. Numbers in the upper left corner of each image slice indicate the MNI z-coordinate, i.e. the axial section in MNI space. Color bars represent color coding for Cohen's d effect size estimates (Cohen, 1977) for the pooled analysis, and T values for the second level analysis, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the pooled data (Table 4 and Figure 3). For the MCI vs. controls comparison, AUCs (accuracies) ranged from 71% (66%) for the second level data to 81% (72%) for the pooled data (Table 4 and Figure 4).

For comparison, the AUCs for left and right averaged hippocampus volume were 86% [2.5/97.5th percentile confidence interval 77%/95%] for the AD vs. controls comparison, and 74% [2.5/97.5th percentile confidence interval 65%/84%] for the MCI vs. controls comparison.

To determine if levels of accuracy measures (AUC and overall accuracy) differed significantly between the values derived from the pooled vs. the 2nd level data, we used the degree of overlap between confidence intervals of the bootstrapped cross-validation data following Afshartous' rule (Afshartous and Preston, 2010). This rule considers the correlation of accuracy measures between samples and the ratio of the standard errors of the accuracy measures of both samples. Following this approach, neither AUCs nor overall accuracies were significantly different between the pooled and the 2nd level test data for the AD vs. controls and the MCI vs. controls comparisons, respectively, at a two tailed significance level of $p < 0.05$.

4.4. Scanner effects

Excluding site V, for the AD vs. control comparison, the proportion of variance attributable to scanner relative to the error variance was 6.6% across all sites and 6.3% for the two sites with balanced group distribution, and was 5.1% for the MCI vs. control comparison.

5. Discussion

In a relatively large multicenter data set of retrospectively pooled rs-fMRI data we found spatially restricted differences in PCC whole brain functional connectivity between AD patients and controls and MCI patients and controls, both in a pooled analysis and a second level analysis stratified according to scanner. The effects were in the expected direction with connectivity smaller in AD/MCI than in controls when removing the data of one site that had failed on visual data inspection, the quality assessments for tSNR and the standardized DVARS, but met all other quality assessments employed. Our findings lead to two major conclusions: Multicenter rs-fMRI using seed based functional

Table 4
Group discrimination in the test sample

AD vs. controls				
	AUC pooled all	AUC pooled sub	AUC 2nd level all	AUC 2nd level sub
Mean	0.816	0.822	0.739	0.757
Lower CI	0.73	0.721	0.63	0.635
Upper CI	0.898	0.918	0.836	0.87
	Ac pooled all	Ac pooled sub	Ac 2nd level all	Ac 2nd level sub
Mean	0.761	0.738	0.708	0.688
Lower CI	0.667	0.634	0.619	0.585
Upper CI	0.841	0.854	0.794	0.805
MCI vs. controls				
	AUC pooled all		AUC 2nd level all	
Mean	0.805		0.713	
Lower CI	0.719		0.617	
Upper CI	0.885		0.803	
	Ac pooled all		Ac 2nd level all	
Mean	0.72		0.662	
Lower CI	0.644		0.575	
Upper CI	0.808		0.74	

Ac – Accuracy.
AUC – Area under the ROC curve.
sub - subsample from two scanners with matched numbers of AD patients and controls.

connectivity has limited accuracy in the discrimination of AD and MCI cases from controls, and requires careful data quality checks beyond evaluation of global quality metrics, including visual inspection of all the data.

The regional distribution of diagnostic group effects in PCC connectivity found in the subset of data passing the visual quality check resembles the results in previous monocenter studies that reported reduced connectivity in MCI or AD within the posterior cingulate gyrus, inferior parietal lobes and medial temporal lobes (Balthazar et al., 2014; Binnewijzend et al., 2012; Chhatwal et al., 2013; Greicius et al., 2004; Koch et al., 2012; Thomas et al., 2014). Overall, the effect sizes of group differences were small, with regional effects passing only an uncorrected level of $p < 0.01$. This finding suggests that multiscanner variability decreases between group effects in functional connectivity. This interpretation is supported by the contribution of 5.2% to 6.6% of the overall variability by scanner related variance in the variance component analysis. In addition, the poor overlap of between group effects across scanners in the conjunction analysis indicates major confounding of group differences by multiscanner variability.

Levels of diagnostic accuracy ranged between 69% based on second level analysis and 76% based on pooled analysis for the AD vs. control comparisons and 66% and 72% for the MCI vs. control comparisons, respectively, in our study. These values are at the lower range of those previously reported from monocenter studies that involved small samples and failed to employ a cross-validation analysis (Balthazar et al., 2014; Koch et al., 2012).They are, however, close to previous estimates from the test data of cross-validated monocenter studies (Dyrba et al., 2015b). It is important to note that the benchmark for assessing performance of a technique is the cross-validated accuracy in the test data, not the accuracy in the training data. According to this benchmark, our multicenter study is at the level of accuracy of monocenter studies. Thus, although the use of multicenter data increases the degrees of freedom of the test statistics it did not increase the power of group discrimination due to confounding inter-scanner variance. One has to consider, however, that the identification of the peak areas that were included in the accuracy estimation was not part of the cross-validation so that effects may be slightly overestimated.

The levels of accuracy for functional connectivity in our study were below the levels of accuracy for hippocampus volume, one of the best established imaging markers of AD to date (for review see (Teipel et al., 2013)), reaching 86% and 74% AUC for the AD vs. controls comparison, respectively. For the MCI vs. controls comparison, the mean AUC for pooled data functional connectivity (81%) was numerically higher than the AUC for hippocampus (74%). The confidence interval of the hippocampus AUC, however, was largely contained within the confidence interval of the functional connectivity measures, suggesting that the functional connectivity measures were not significantly more accurate for the MCI vs. controls discrimination than the easily accessible hippocampus volumetry.

The results were clearly sensitive to scan quality. When we included the large data set of site V that had severe susceptibility artifacts, the direction of the group differences was inverted. When we considered the global scan quality measures, the tSNR (Marcus et al., 2013) and standardized DVARS (Power et al., 2012) suggested that insufficient signal in the healthy control group was driving this effect. Interestingly, other metrics employed in other multiscanner data pooling activities, including the intrinsic functional connectivity for two key areas of the DMN (Zuo et al., 2014), fractional ALFF (Yan et al., 2013b; Zuo et al., 2014), foreground to background energy ratio (Zuo et al., 2014), or subject head motion (Jenkinson et al., 2002; Power et al., 2012; Power et al., 2014; Yan et al., 2013a), were inconspicuous for these data, suggesting that determining tSNR (Marcus et al., 2013), standardized DVARS (Power et al., 2012), and visual inspection of all the data are indispensable for multiscanner data pooling. This is relevant since large scale data pooling efforts such as the PCP Quality Assessment Protocol (preprocessed-connectomes-project.org/quality-assessment-protocol/index.html) and the 1000 functional connectomes project (Yan et al., 2013b; Zuo et al., 2014) focus on the detection and correction of spatial displacements and head motion that were inconspicuous with the site V data.

Despite the high relevance of multiscanner variability of rs-fMRI data (Jovicich et al., 2016; Lin et al., 2015; Magnotta et al., 2006), the large majority of studies on multicenter rs-fMRI in neuropsychiatric diseases did not take multiscanner effects into account, even if protocols differed between sites (Chhatwal et al., 2013; Demertzi et al., 2015; Esslinger et al., 2011; Martucci et al., 2015; Sripada et al., 2014; Thomas et al., 2014). Regional effects of group differences strongly overlapped between a fixed effect analysis, including scanner as covariate, and a second level analysis stratified according to scanner, but were more extended for the pooled analysis. Numerically, group discrimination was smaller based on the second level analysis compared to the pooled analysis, albeit this difference was not statistically significant. A second level analysis of voxel-wise functional connectivity using Fisher's z-transformed correlation coefficients resembles a center-wise voxel-based meta-analysis (Teipel et al., 2012) that determines voxel-wise effect size within sites and then assesses the confidence level of the voxel-wise effect size estimates across sites. Such an approach has been used in one previous study across four cohorts of 129 MCI cases and 99 controls (Tam et al., 2015). The main outcome of this previous study were Cohen's d (Cohen, 1977) effect size estimates that ranged between 0.10 to 0.48, representing moderate effect sizes of MCI vs. control differences in regions of interests that were empirically derived from proximity metrics without a priori region selection. These moderate effect sizes agree with the effect sizes of group differences below Cohen's $d = 1$ in our peak voxel analysis (figures 1a and 2a).

An interesting question is the effect of multicenter acquisition on between subjects variability in trajectories of intra-individual change from longitudinal studies. Evidence here is still very limited. One recent study evaluated reproducibility of rs-fMRI connectivity across 13 different scanners at baseline and 7 to 60 days of follow-up in five healthy people per site (Jovicich et al., 2016), including different scanner types and vendors. In this study site differences in test-retest-variability of PCC connectivity were marginally not significant ($p < 0.06$). This finding suggests that multicenter acquisition not only introduces higher variability of between group differences as shown in our current study,

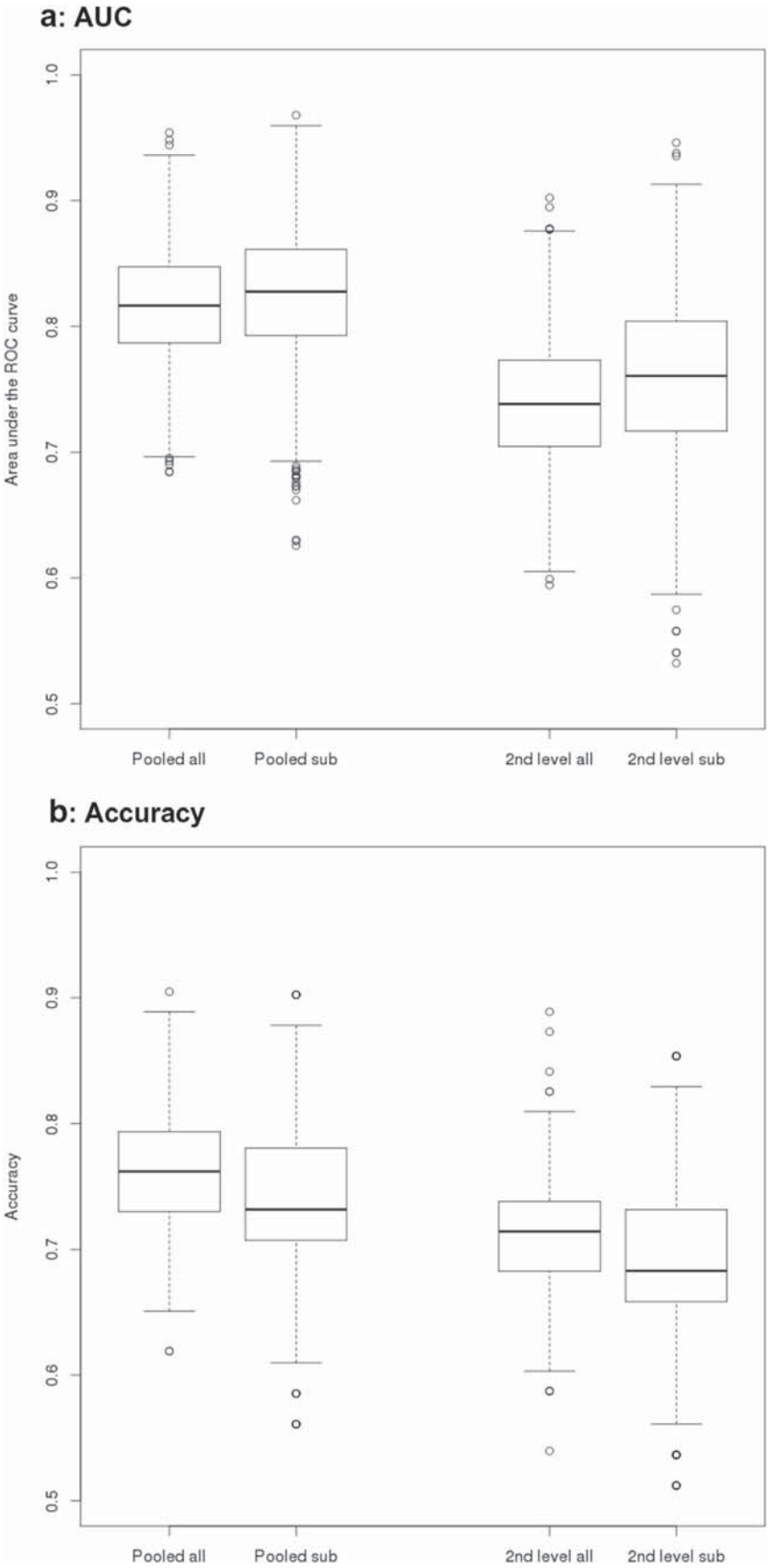


Fig. 3. Areas under ROC and accuracy for AD vs. control comparisons Box plots of AUC and accuracy levels from cross-validation logistic regression. Levels of AUC (Panel a) and accuracy (Panel b) were determined using bootstrapped logistic regression models on the discrimination between AD patients and controls following a pooled analysis with center covariate ("pooled"), and a second level analysis with center as second level factor ("2nd level"), respectively. Analyses were repeated, using all AD and control data ("all") as well as only data from a subset of centers where number of AD cases and controls was matched between centers ("sub").

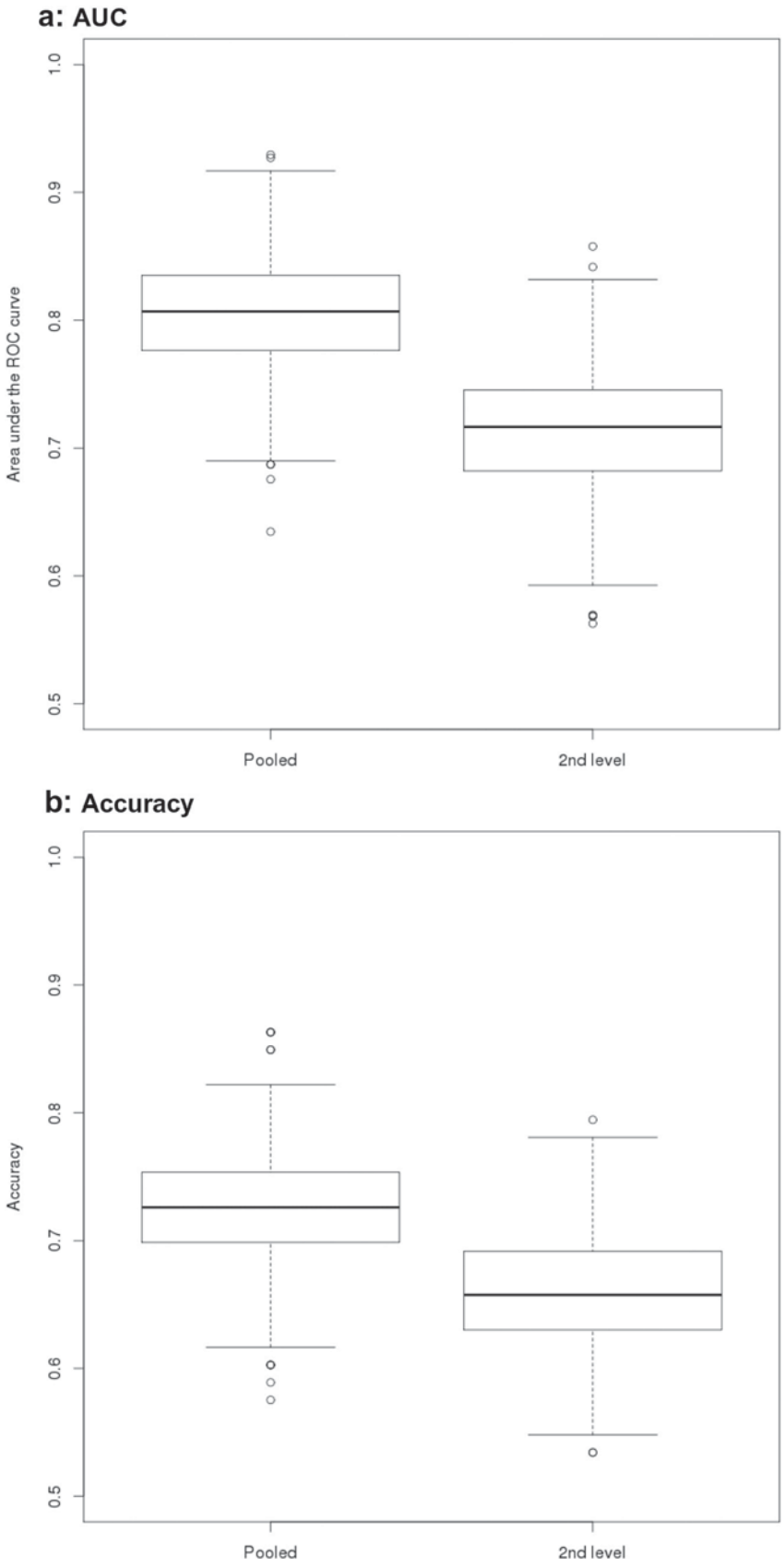


Fig. 4. Areas under ROC and accuracy for MCI vs. control comparisons Box plots of AUC and accuracy levels from cross-validation logistic regression. Levels of AUC (Panel a) and accuracy (Panel b) were determined using bootstrapped logistic regression models on the discrimination between MCI patients and controls following a pooled analysis with center covariate ("pooled"), and a second level analysis with center as second level factor ("2nd level"), respectively.

but may introduce additional noise into the assessment of trajectories of intra-individual change.

We need to consider several limitations of our study. First, the scan protocols were different between scanners. This is not the case in a prospectively planned cohort study with a unified protocol. Still, even in absence of a harmonized protocol previous studies have pooled rs-fMRI data in studying neuropsychiatric diseases so that these findings are pertinent to the present state of research. More homogeneous acquisition parameters may amend some of the scanner effects but at the same time limit the usefulness of multicenter acquisition in routine care, where differences in scanner type and manufacturer will not allow perfect alignment of scanning parameters across sites. We carefully checked the image quality of each single scan by visual inspection. As a result, we excluded the data of site V. The remaining data had high image quality upon visual inspection, consistent with the results of the quality metrics employed. Still, the combination of data from different scanning protocols and scanner resulted in high inter-scanner variability despite sufficient intra-scanner scan quality. In future, we plan to determine the effects of multiscanner acquisition from an ongoing prospective multicenter study in MCI, AD, and healthy controls that employs a harmonized rs-fMRI protocol across sites. Although we expect that the multicenter effects may be smaller in such a harmonized study, we still anticipate that multiscanner effects will limit accuracy of group discrimination. Secondly, different preprocessing protocols may be useful to reduce multiscanner variation. Here, we employed a preprocessing protocol that was oriented on the recommendations from a systematic evaluation of processing steps (Shirer et al., 2015), and used cross-correlation as connectivity metric that has been found more stable than other connectivity metrics, such as cross-coherence and partial cross-correlation, in a previous study (Fiecas et al., 2013). We did, however, not systematically explore other processing steps and connectivity metrics. Thirdly, group discrimination accuracy can never perform better than the reference standard. The reference standard in our study for AD and MCI definition lacked CSF or PET biomarker evidence for most cases, but data came from expert centers experienced in the early diagnosis of AD and MCI. Still, a final judgment of the added value of rs-fMRI for AD diagnosis must await systematic evaluation of diagnostic accuracy in multicenter data from biomarker stratified cases.

In summary, we found spatially restricted group differences in resting state functional connectivity in AD patients and MCI patients compared to controls, limited by high multiscanner variability. The accuracy of group discrimination resembled findings from previous monocenter studies using a training/test data set approach, encouraging the conclusion that rs-fMRI at least when using seed based functional connectivity metrics may play a limited role in early diagnosis of AD or MCI. The discrimination accuracy in the test data did not reach the internal benchmark set by the established marker of hippocampus volumetry. This conclusion needs further corroboration in biomarker qualified multicenter cohorts. From a practical viewpoint, studies pooling multicenter rs-fMRI data should employ careful data quality checks that need to include tSNR, standardized DVARS, and visual inspection of all the data besides other established global metrics, and should use explicit modelling of scanner effects such as provided by second level models or center-based meta-analysis when focusing on univariate approaches. Potential usefulness of multivariate non-linear approaches such as provided by machine learning algorithms that were successfully employed in reducing multiscanner effects for structural connectivity data (Dyrba et al., 2015a; Dyrba et al., 2013) is another open area of research.

Acknowledgment

SJT received support by a grant of the Federal Ministry of Research (BMBF) (AgeGain, 1GQ1425B). ME received support by grants from the Alzheimer Forschung Initiative (AFI), LMUexcellent and the European Commission (PCIG12-GA-2012-334259).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2017.01.018.

References

Afshartous, D., Preston, R.A., 2010. Confidence intervals for dependent data: equating non-overlap with statistical significance. Comput. Stat. Data Anal. 54, 2296–2305.
Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. 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. 7, 270–279.
Andrews-Hanna, J.R., Reidler, J.S., Sepulcre, J., Poulin, R., Buckner, R.L., 2010. Functional-anatomic fractionation of the brain's default network. Neuron 65, 550–562.
Anticevic, A., Haut, K., Murray, J.D., Repovs, G., Yang, G.J., Diehl, C., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Tsuang, M.T., van Erp, T.G., Walker, E.F., Hamann, S., Woods, S.W., Qiu, M., Cannon, T.D., 2015. Association of thalamic dysconnectivity and conversion to psychosis in youth and young adults at elevated clinical risk. JAMA Psychiatry 72, 882–891.
Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113.
Atkinson, D., Hill, D.L.G., Stoyle, P.N.R., Summers, P.E., Keevil, S.F., 1997. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. IEEE Trans. Med. Imaging 16, 903–910.
Balthazar, M.L., de Campos, B.M., Franco, A.R., Damasceno, B.P., Cendes, F., 2014. Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease. Psychiatry Res. 221, 37–42.
Binnewijzend, M.A., Schoonheim, M.M., Sanz-Arigita, E., Wink, A.M., van der Flier, W.M., Tolboom, N., Adriaanse, S.M., Damoiseaux, J.S., Scheltens, P., van Berckel, B.N., Barkhof, F., 2012. Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment. Neurobiol. Aging 33, 2018–2028.
Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kotter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. U.S.A. 107, 4734–4739.
Blautzik, J., Keeser, D., Berman, A., Paolini, M., Kirsch, V., Mueller, S., Coates, U., Reiser, M., Teipel, S.J., Meindl, T., 2013. Long-term test-retest reliability of resting-state networks in healthy elderly subjects and with amnestic mild cognitive impairment patients. J Alzheimers Dis. 34, 741–754.
Chao-Gan, Y., Yu-Feng, Z., 2010. DPARSF: A MATLAB Toolbox for "Pipeline" data analysis of resting-state fMRI. Front. Syst. Neurosci. 4, 13.
Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., Dong, H.M., Yang, Z., Zang, Y.F., Zuo, X.N., Weng, X.C., 2015. Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. PLoS One 10, e0144963.
Chhatwal, J.P., Schultz, A.P., Johnson, K., Benzinger, T.L., Jack Jr., C., Ances, B.M., Sullivan, C.A., Salloway, S.P., Ringman, J.M., Koeppe, R.A., Marcus, D.S., Thompson, P., Saykin, A.J., Correia, S., Schofield, P.R., Rowe, C.C., Fox, N.C., Brickman, A.M., Mayeux, R., McDade, E., Bateman, R., Fagan, A.M., Goate, A.M., Xiong, C., Buckles, V.D., Morris, J.C., Sperling, R.A., 2013. Impaired default network functional connectivity in autosomal dominant Alzheimer disease. Neurology 81, 736–744.
Chou, Y.H., Panych, L.P., Dickey, C.C., Petrella, J.R., Chen, N.K., 2012. Investigation of long-term reproducibility of intrinsic connectivity network mapping: a resting-state fMRI study. AJNR Am. J. Neuroradiol. 33, 833–838.
Cohen, J., 1977. Statistical Power Analysis for the Behavioural Sciences. Academic Press, New York.
Demertzi, A., Antonopoulos, G., Heine, L., Voss, H.U., Crone, J.S., de Los Angeles, C., Bahri, M.A., Di Perri, C., Vanhaudenhuyse, A., Charland-Verville, V., Kronbichler, M., Trinka, E., Phillips, C., Gomez, F., Tshibanda, L., Soddu, A., Schiff, N.D., Whitfield-Gabrieli, S., Laureys, S., 2015. Intrinsic functional connectivity differentiates minimally conscious from unresponsive patients. Brain 138, 2619–2631.
Doraiswamy, P.M., Sperling, R.A., Johnson, K., Reiman, E.M., Wong, T.Z., Sabbagh, M.N., Sadowsky, C.H., Fleisher, A.S., Carpenter, A., Joshi, A.D., Lu, M., Grundman, M., Mintun, M.A., Skovronsky, D.M., Pontecorvo, M.J., Group, A.A.S, 2014. Florbetapir F 18 amyloid PET and 36-month cognitive decline: a prospective multicenter study. Mol. Psychiatry 19, 1044–1051.
Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P.J., Scheltens, P., 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 6, 734–746.
Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., Dekosky, S.T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N.C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G.A., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Sarazin, M.,

de Souza, L.C., Stern, Y., Visser, P.J., Scheltens, P., 2010. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 9, 1118–1127.
Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., Blennow, K., DeKosky, S.T., Gauthier, S., Selkoe, D., Bateman, R., Cappa, S., Crutch, S., Engelborghs, S., Frisoni, G.B., Fox, N.C., Galasko, D., Habert, M.O., Jicha, G.A., Nordberg, A., Pasquier, F., Rabinovici, G., Robert, P., Rowe, C., Salloway, S., Sarazin, M., Epelbaum, S., de Souza, L.C., Vellas, B., Visser, P.J., Schneider, L., Stern, Y., Scheltens, P., Cummings, J.L., 2014. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. 13, 614–629.
Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., Meindl, T., Pievani, M., Bokde, A.L., Fellgiebel, A., Filippi, M., Hampel, H., Kloppel, S., Hauenstein, K., Kirste, T., Teipel, S.J., Group, E.S, 2013. Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data. PLoS One 8, e64925.
Dyrba, M., Barkhof, F., Fellgiebel, A., Filippi, M., Hausner, L., Hauenstein, K., Kirste, T., Teipel, S.J., 2015a. Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. J. Neuroimaging.
Dyrba, M., Grothe, M., Kirste, T., Teipel, S.J., 2015b. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Hum. Brain Mapp. 36, 2118–2131.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton.
Esslinger, C., Kirsch, P., Haddad, L., Mier, D., Sauer, C., Erk, S., Schnell, K., Arnold, C., Witt, S.H., Rietschel, M., Cichon, S., Walter, H., Meyer-Lindenberg, A., 2011. Cognitive state and connectivity effects of the genome-wide significant psychosis variant in ZNF804A. NeuroImage 54, 2514–2523.
Ewers, M., Teipel, S.J., Dietrich, O., Schonberg, S.O., Jessen, F., Heun, R., Scheltens, P., van de Pol, L., Freymann, N.R., Moeller, H.J., Hampel, H., 2006. Multicenter assessment of reliability of cranial MRI. Neurobiol. Aging 27, 1051–1059.
Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A., Feng, D., 2013. Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fMRI. NeuroImage 65, 231–241.
Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10, 507.
Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental-state: a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198.
Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U.S.A. 102, 9673–9678.
Friston, K.J., Penny, W.D., Glaser, D.E., 2005. Conjunction revisited. NeuroImage 25, 661–667.
Friston, K.J., Ashburner, J., Kiebel, S., Nichols, T., Penny, W.D., 2007. Statistical Parametric Mapping: The Analysis of Funtional Brain Images. First edition ed. Elsevier/Academic Press, Amsterdam; Boston.
Gaser, C., Volz, H.-P., Kiebel, S., Riehemann, S., Sauer, H., 1999. Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research. NeuroImage 10, 107–113.
Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. U.S.A. 101, 4637–4642.
Grothe, M., Heinsen, H., Teipel, S.J., 2012. Atrophy of the cholinergic Basal forebrain over the adult age range and in early stages of Alzheimer's disease. Biol. Psychiatry 71, 805–813.
Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. NeuroImage 61, 1471–1483.
Hedden, T., Van Dijk, K.R., Becker, J.A., Mehta, A., Sperling, R.A., Johnson, K.A., Buckner, R.L., 2009. Disruption of functional connectivity in clinically normal older adults harboring amyloid burden. J. Neurosci. 29, 12686–12694.
Herholz, K., 2010. Cerebral glucose metabolism in preclinical and prodromal Alzheimer's disease. Expert Rev. Neurother. 10, 1667–1673.
Ito, K., Fukuyama, H., Senda, M., Ishii, K., Maeda, K., Yamamoto, Y., Ouchi, Y., Ishii, K., Okumura, A., Fujiwara, K., Kato, T., Arahata, Y., Washimi, Y., Mitsuyama, Y., Meguro, K., Ikeda, M., Group, S.-J.S, 2015. Prediction of outcomes in mild cognitive impairment by using 18F-FDG-PET: a multicenter study. J Alzheimers Dis 45, 543–552.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartres-Faz, D., Arnold, J., Benninghoff, J., Fiedler, U., Roccatagliata, L., Picco, A., Nobili, F., Blin, O., Bombois, S., Lopes, R., Bordet, R., Sein, J., Ranjeva, J.P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargallo, N., Ferretti, A., Caulo, M., Aiello, M., Cavaliere, C., Soricelli, A., Parnetti, L., Tarducci, R., Floridi, P., Tsolaki, M., Constantinidis, M., Drevelegas, A., Rossini, P.M., Marra, C., Schonknecht, P., Hensch, T., Hoffmann, K.T., Kuijer, J.P., Visser, P.J., Barkhof, F., Frisoni, G.B., PharmaCog, C., 2016. Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fMRI study. NeuroImage 124, 442–454.
Kilimann, I., Grothe, M., Heinsen, H., Alho, E.J., Grinberg, L., Amaro Jr., E., Dos Santos, G.A., da Silva, R.E., Mitchell, A.J., Frisoni, G.B., Bokde, A.L., Fellgiebel, A., Filippi, M., Hampel, H., Kloppel, S., Teipel, S.J., 2014. Subregional basal forebrain atrophy in Alzheimer's disease: a multicenter study. J Alzheimers Dis 40, 687–700.
Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A.L., Hampel, H., Coates, U., Reiser, M., Meindl, T., 2012. Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. Neurobiol. Aging 33, 466–478.
Lin, Q., Dai, Z., Xia, M., Han, Z., Huang, R., Gong, G., Liu, C., Bi, Y., He, Y., 2015. A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. Sci. Data 2, 150056.
Long, X.Y., Zuo, X.N., Kiviniemi, V., Yang, Y., Zou, Q.H., Zhu, C.Z., Jiang, T.Z., Yang, H., Gong, Q.Y., Wang, L., Li, K.C., Xie, S., Zang, Y.F., 2008. Default mode network as revealed with multiple methods for resting-state functional MRI analysis. J. Neurosci. Methods 171, 349–355.
Magnotta, V.A., Friedman, L., First, B., 2006. Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study. J. Digit. Imaging 19, 140–147.
Marcus, D.S., Harms, M.P., Snyder, A.Z., Jenkinson, M., Wilson, J.A., Glasser, M.F., Barch, D.M., Archie, K.A., Burgess, G.C., Ramaratnam, M., Hodge, M., Horton, W., Herrick, R., Olsen, T., McKay, M., House, M., Hileman, M., Reid, E., Harwell, J., Coalson, T., Schindler, J., Elam, J.S., Curtiss, S.W., Van Essen, D.C., Consortium, W.U.-M.H., 2013. Human Connectome Project informatics: quality control, database services, and data visualization. NeuroImage 80, 202–219.
Martucci, K.T., Shirer, W.R., Bagarinao, E., Johnson, K.A., Farmer, M.A., Labus, J.S., Apkarian, A.V., Deutsch, G., Harris, R.E., Mayer, E.A., Clauw, D.J., Greicius, M.D., Mackey, S.C., 2015. The posterior medial cortex in urologic chronic pelvic pain syndrome: detachment from default mode network-a resting-state study from the MAPP Research Network. Pain 156, 1755–1764.
McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34, 939–944.
McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia 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. 7, 263–269.
Meindl, T., Teipel, S., Elmouden, R., Mueller, S., Koch, W., Dietrich, O., Coates, U., Reiser, M., Glaser, C., 2010. Test-retest reproducibility of the default-mode network in healthy individuals. Hum. Brain Mapp. 31, 237–246.
Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G., Mellits, E.D., Clark, C., 1989. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology 39, 1159–1165.
Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? NeuroImage 44, 893–905.
Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H., 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53, 695–699.
Orban, P., Madjar, C., Savard, M., Dansereau, C., Tam, A., Das, S., Evans, A.C., Rosa-Neto, P., Breitner, J.C., Bellec, P., Group, P.-A.R, 2015. Test-retest resting-state fMRI in healthy elderly persons with a family history of Alzheimer's disease. Sci. Data 2, 150043.
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56, 303–308.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage 59, 2142–2154.
Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage 84, 320–341.
Pruessner, J.C., Li, L.M., Serles, W., Pruessner, M., Collins, D.L., Kabani, N., Lupien, S., Evans, A.C., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10, 433–442.
Risacher, S.L., Saykin, A.J., West, J.D., Shen, L., Firpi, H.A., McDonald, B.C., Alzheimer's Disease Neuroimaging, I., 2009. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alzheimer Res. 6, 347–361.
Shirer, W.R., Jiang, H., Price, C.M., Ng, B., Greicius, M.D., 2015. Optimization of rs-fMRI Preprocessing for enhanced signal-noise separation, test-retest reliability, and group discrimination. NeuroImage 117, 67–79.
Sladky, R., Friston, K.J., Tröstl, J., Cunnington, R., Moser, E., Windischberger, C., 2011. Slice-timing effects and their correction in functional MRI. NeuroImage 58, 588–594.
Sripada, C., Kessler, D., Fang, Y., Welsh, R.C., Prem Kumar, K., Angstadt, M., 2014. Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder. Hum. Brain Mapp. 35, 4693–4705.
Suckling, J., Barnes, A., Job, D., Brennan, D., Lymer, K., Dazzan, P., Marques, T.R., MacKay, C., McKie, S., Williams, S.R., Williams, S.C., Deakin, B., Lawrie, S., 2012. The Neuro/PsyGRID calibration experiment: identifying sources of variance and bias in multicenter MRI studies. Hum. Brain Mapp. 33, 373–386.
Tam, A., Dansereau, C., Badhwar, A., Orban, P., Belleville, S., Chertkow, H., Dagher, A., Hanganu, A., Monchi, O., Rosa-Neto, P., Shmuel, A., Wang, S., Breitner, J., Bellec, P., Alzheimer's Disease Neuroimaging, I., 2015. Common effects of amnestic mild cognitive impairment on resting-state connectivity across four independent studies. Front. Aging Neurosci. 7, 242.
Teipel, S.J., Wegrzyn, M., Meindl, T., Frisoni, G., Bokde, A.L., Fellgiebel, A., Filippi, M., Hampel, H., Kloppel, S., Hauenstein, K., Ewers, M., 2012. Anatomical MRI and DTI in the diagnosis of Alzheimer's disease: a European multicenter study. J. Alzheimers Dis. 31 (Suppl. 3), S33–S47.

Teipel, S.J., Grothe, M., Lista, S., Toschi, N., Garaci, F.G., Hampel, H., 2013. Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. Med. Clin. North Am. 97, 399–424.

Teipel, S., Drzezga, A., Grothe, M.J., Barthel, H., Chetelat, G., Schuff, N., Skudlarski, P., Cavedo, E., Frisoni, G.B., Hoffmann, W., Thyrian, J.R., Fox, C., Minoshima, S., Sabri, O., Fellgiebel, A., 2015. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. Lancet Neurol. 14, 1037–1053.

Thomas, J.B., Brier, M.R., Bateman, R.J., Snyder, A.Z., Benzinger, T.L., Xiong, C., Raichle, M., Holtzman, D.M., Sperling, R.A., Mayeux, R., Ghetti, B., Ringman, J.M., Salloway, S., McDade, E., Rossor, M.N., Ourselin, S., Schofield, P.R., Masters, C.L., Martins, R.N., Weiner, M.W., Thompson, P.M., Fox, N.C., Koeppe, R.A., Jack Jr., C.R., Mathis, C.A., Oliver, A., Blazey, T.M., Moulder, K., Buckles, V., Hornbeck, R., Chhatwal, J., Schultz, A.P., Goate, A.M., Fagan, A.M., Cairns, N.J., Marcus, D.S., Morris, J.C., Ances, B.M., 2014. Functional connectivity in autosomal dominant and late-onset Alzheimer disease. JAMA Neurol 71, 1111–1122.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8, e77089.

Wu, C.W., Chen, C.L., Liu, P.Y., Chao, Y.P., Biswal, B.B., Lin, C.P., 2011. Empirical evaluations of slice-timing, smoothing, and normalization effects in seed-based, resting-state functional magnetic resonance imaging analyses. Brain Connect. 1, 401–410.

Yan, C.G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Di Martino, A., Li, Q., Zuo, X.N., Castellanos, F.X., Milham, M.P., 2013a. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. NeuroImage 76, 183–201.

Yan, C.G., Craddock, R.C., Zuo, X.N., Zang, Y.F., Milham, M.P., 2013b. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. NeuroImage 80, 246–262.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., Wells, W.M., Group, F.B.R, 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 237, 781–789.

Zuo, X.-N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C.S., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S.J., Courtney, W., Craddock, R.C., Di Martino, A., Dong, H.-M., Fu, X., Gong, Q., Gorgolewski, K.J., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.-H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S.M., Lainhart, J.E., Lei, X., Li, H.-J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D.S., Mayer, A.R., Meindl, T., Meyerand, M.E., Nan, W., Nielsen, J.A., O'Connor, D., Paulsen, D., Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.-X., Weng, X.-C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.-F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.-T., Milham, M.P., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 1, 140049.

Teipel, Stefan J.; Wohlert, Alexandra; Metzger, Coraline; Grimmer, Timo; Sorg, Christian; Ewers, Michael et al. (2017): *Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI. I*n: *NeuroImage: Clinical.*
DOI: *10.1016/j.nicl.2017.01.018.*

Supplementary material

Supplementary Table 1: Number of subjects per site stratified by diagnosis.

Diagnosis	Site I	Site II	Site III	Site IV	Site V	Total
HC	19	40	41	18	33	151
MCI	22	23	18	16	36	115
AD	–	41	–	12	31	84

Description of quality measures

Several metrics have been used in the literature to quantify functional scan quality. This section provides a brief description of the metrics and their interpretation.

Framewise displacement (FD)
Framewise displacement (FD) measures the amount of head motion from one time point to the next (Jenkinson et al., 2002; Power et al., 2012; Power et al., 2014; Yan et al., 2013a). Several approaches have been proposed to integrate the translation and rotation information (Jenkinson et al., 2002; Power et al., 2012; Van Dijk et al., 2012), see (Yan et al., 2013a) for a detailed comparison. The mean FD provides the information how much head movement is present averaged over the whole acquisition time. In contrast, the percentage above a certain threshold provides the proportion of time points in which (strong) motion occurred. We used Jenkinson's method (Jenkinson et al., 2002) as this metric was reported to provide most similar results compared to the more complex voxel-based estimation of head motion

(Yan et al., 2013a). Power et al showed that significant within-subject changes in correlations are detectable already at FD of 0.2 mm and are very pronounced at a FD of 0.5 mm (Power et al., 2014). They also showed that global signal regression reduces the effect of head motion on functional connectivity by zero-centering of motion-related signal intensity (Power et al., 2014). Patients with mild cognitive impairment or dementia are expected to move more during MRI acquisition than healthy subjects (Haller et al., 2014).

Temporal signal-to-noise ratio (tSNR)
The temporal signal to noise ratio (tSNR) is calculated as the mean signal over time divided by the standard deviation over time in each voxel (Van Dijk et al., 2012; Welvaert and Rosseel, 2013). The average tSNR for the whole brain represents the common variability of the BOLD signal, which increases in case of strong artifacts, for instance arising from head motion (Welvaert and Rosseel, 2013). Healthy subjects are expected to have a higher tSNR than patients with mild cognitive impairment or Alzheimer's dementia due to the higher variability of the BOLD signal in patients and the amount of motion in patients (Haller et al., 2014).

Standardized DVARS
DVARS measures the root mean square of change in signal intensity from one time point to the next (Power et al., 2012). D refers to temporal derivative of timecourses, and VARS refers to root mean square variance over voxels. DVARS is a measure of how much the intensity of a brain image changes in comparison to the previous time point. Nichols proposed a standardized version of DVARS to allow for a comparison across subjects and sites[1]. Standardized DVARS is scaled by the temporal standard deviation and temporal autocorrelation so that it is approximately 1 if there are no artifacts in the functional data[1]. The mean value derived from the voxels of the whole brain is provided. This measure is used to detect scan artifacts that are not necessarily related to head motion (Power et al., 2012; Power et al., 2014).

Percentage of outlier voxels
The mean percentage of outliers quantifies the fraction of outlier voxels within the whole brain (Zuo et al., 2014). Outlier voxels are defined as voxel intensity strongly deviating from the temporal median value[2], for instance due to scanner noise or short-term head movement (Power et al., 2014). Fewer outliers mean better signal quality (Zuo et al., 2014). As motion is increased in patients compared to controls (Haller et al., 2014), more outlier voxels are expected in the patient groups.

Foreground to background energy ratio (FBER)
The foreground to background energy ratio estimates the mean amplitude of the BOLD signal within the whole brain compared to the mean amplitude of the signal outside of the brain, for instance arising from noise signal or ghosting (Zuo et al., 2014). Values also depend on the size of the field of view (FOV) defined for the fMRI sequence. Thus, sites cannot be compared directly. Higher values indicate a clearer signal and fewer ghosting artifacts (Zuo et al., 2014).

[1] Nichols, 2013, Notes on Creating a Standardized Version of DVARS, http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/scripts/fsl/standardizeddvars.pdf
[2] See 3dToutcount in AFNI, https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dToutcount.html

Fractional amplitude of low frequency fluctuations (fALFF)
The fractional amplitude of low-frequency fluctuations (fALFF) measures the sum of amplitudes in the low-frequency range 0.1–0.01 Hz compared to the sum of amplitudes of the whole frequency spectrum (Zou et al., 2008). Thus, it can be used as a measure of intrinsic brain activity within a certain brain region (Han et al., 2011). To be able to compare subjects and sites, absolute fALFF values need to be scaled to Z score using whole brain mean and standard deviation (Yan et al., 2013a). Patients with Alzheimer's dementia and mild cognitive impairment were found to have a lower ALFF in the posterior cingulate cortex compared to healthy controls (Han et al., 2011; Wang et al., 2011).

Mean functional connectivity

Functional connectivity is commonly defined as the Pearson correlation of the signal time course between two segregated brain regions (Greicius et al., 2004). For a statistical analysis, these values are commonly standardized to be normal distributed using Fisher's z-transform (Yan et al., 2013a). The functional connectivity between the posterior cingulate cortex and the anterior medioprefrontal cortex was reported to be significantly reduced in Alzheimer's dementia compared to healthy subjects (Koch et al., 2012). Complementary to voxel-wise statistics, the region-of-interest-based analysis allows the quantification of the variability of the functional connectivity across diagnostic groups and sites.

Supplementary figures legends

Supplementary Figure 1: Mean framewise displacement
The mean framewise displacement measures the amount of head motion during the scan. Jenkinson's method (Jenkinson et al., 2002) was used to combine translation and rotation estimates, as this metric was reported to provide most similar results compared to the more complex voxel-based estimation of head motion (Yan et al., 2013a).

Supplementary Figure 2: Percentage framewise displacement > 0.5 mm
The percentage of framewise displacement greater than 0.5 mm measures the proportion of time points in which head motion is strong enough to introduce pronounced artifacts into the BOLD signal (Power et al., 2014).

Supplementary Figure 3: Foreground to background Energy Ratio
The foreground to background energy ratio estimates the mean amplitude of the BOLD signal within the whole brain compared to the mean amplitude of the signal outside of the brain, for instance arising from noise signal or ghosting. Values depend on the size of the field of view and cannot be directly compared between sites. The field of view for site III was larger than that for the other sites and included parts of the neck. Severe ghosting artifacts were present in site V.

Supplementary Figure 4: Fractional amplitude of low frequency fluctuations in PCC
The fractional amplitude of low-frequency fluctuations (fALFF) measures the sum of amplitudes in the low-frequency range 0.1–0.01 Hz compared to the sum of amplitudes of

the whole frequency spectrum. Fractional ALFF maps were obtained from the normalized and detrended functional scans. To be able to compare subjects and sites, absolute fALFF values were scaled to Z score using whole brain mean and standard deviation (Yan et al., 2013a). The figure shows scaled fALFF of the posterior cingulate cortex (PCC), obtained from a spherical ROI at MNI coordinate 0, -53, 26with a radius of 4 mm (Hedden et al., 2009).

Supplementary Figure 5: Mean functional connectivity between PCC and anterior medioprefrontal cortex

Functional connectivity represents the Fisher's z-transformed Pearson correlation of the signal time course between the posterior cingulate cortex (PCC) and the anterior medioprefrontal cortex (aMPFC). It was obtained using two spherical ROI at MNI coordinates 0, -53, 26 (PCC) (Hedden et al., 2009) and -6, 52, -2 (aMPFC) (Andrews-Hanna et al., 2010), each with a radius of 4 mm.

Supplementary Figure 6: Mean temporal signal to noise ratio

The temporal signal to noise ratio (tSNR) is calculated as the mean signal over time divided by the standard deviation over time in each voxel. The average tSNR for the whole brain represents the common variability of the BOLD signal, which increases in case of strong artifacts, for instance arising from head motion (Welvaert and Rosseel, 2013).

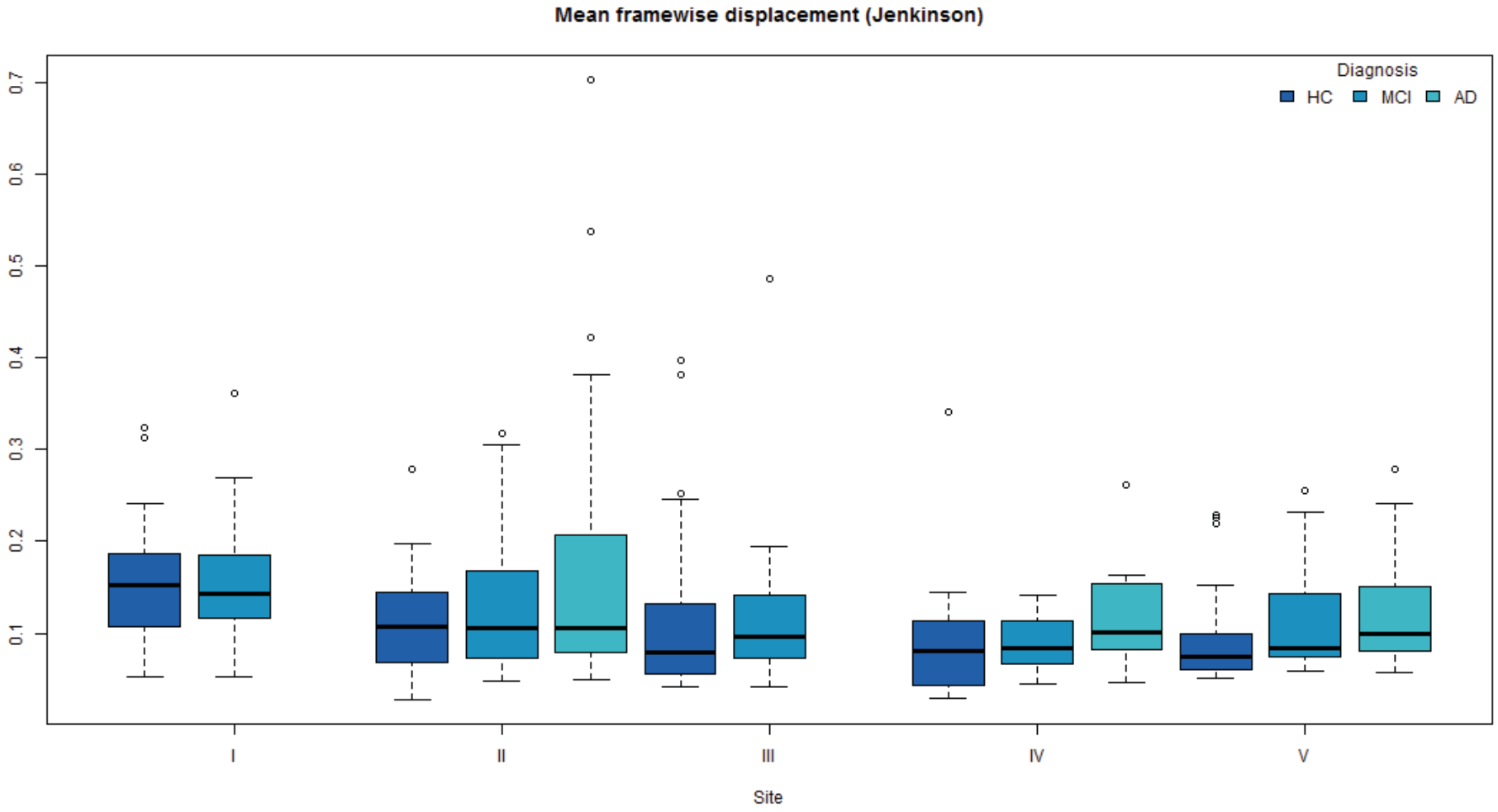
Supplementary Figure 7: Mean percentage of outlier voxels

The mean percentage of outlier voxels was estimated for the whole brain using the program 3dToutcount included in the AFNI software library.

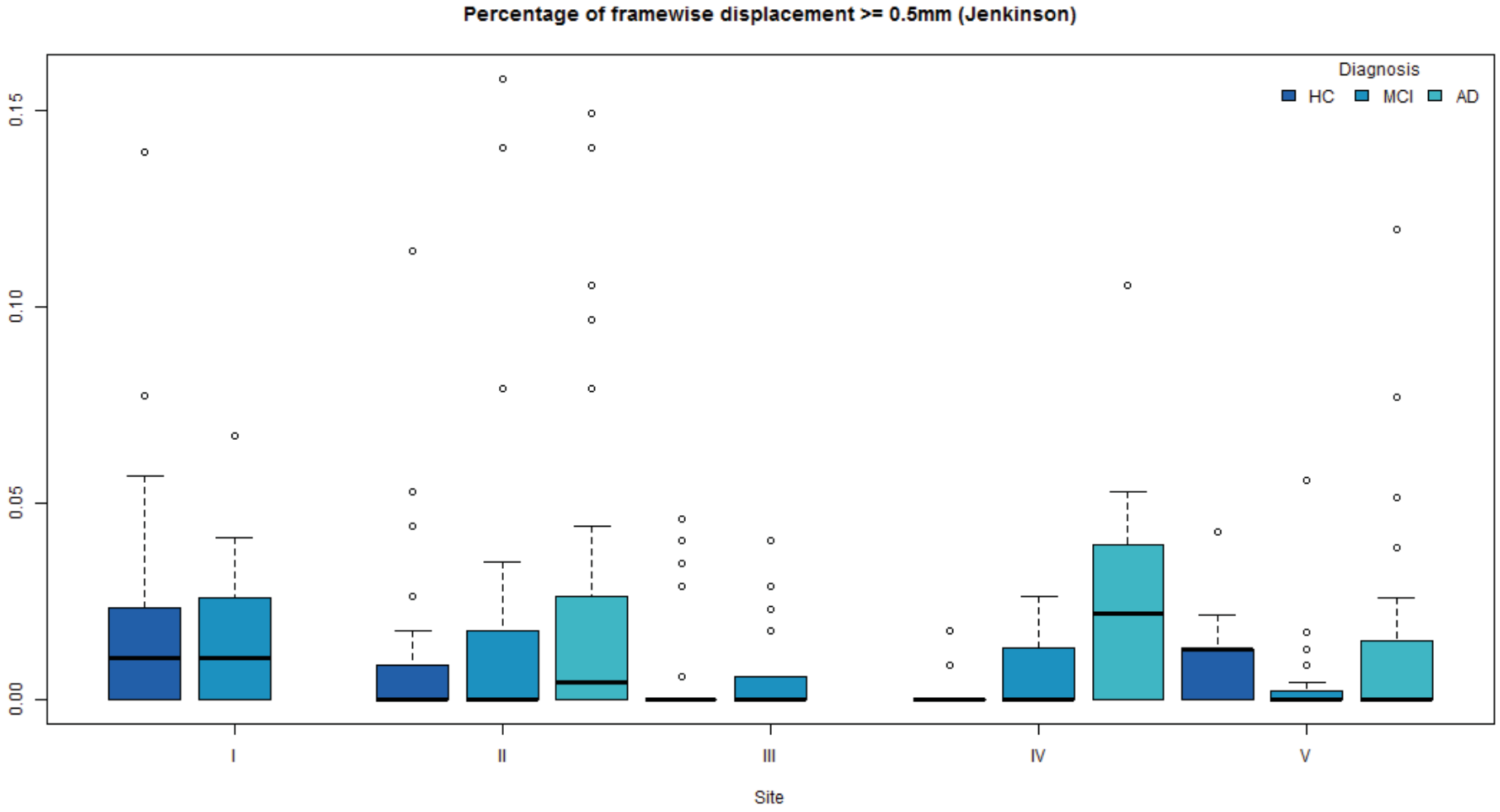
Supplementary Figure 8: Standardized DVARS

DVARS measures the rate of change of the BOLD signal across the entire brain at each frame of data compared to the previous time point. DVARS aims to detect short-scale changes that indicate fMRI quality problems that are not related to motion (Zuo et al., 2014). The standardized version of DVARS is normalized to the temporal standard deviation and autocorrelation in order to allow a comparison between subjects and sites (Zuo et al., 2014). It is approximate one if there are not artifacts in a dataset.

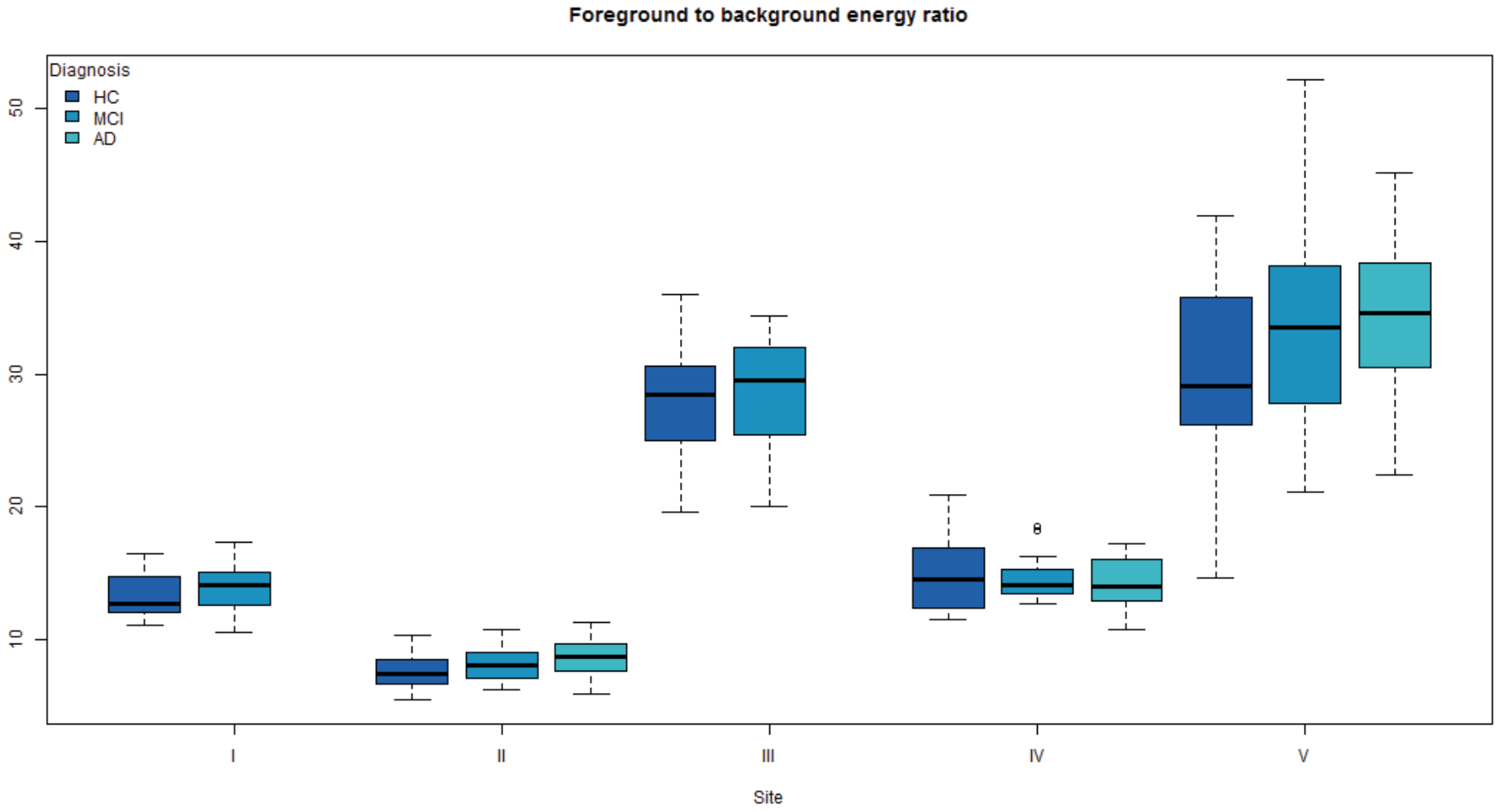
Supplementary Figure 1: Mean framewise displacement



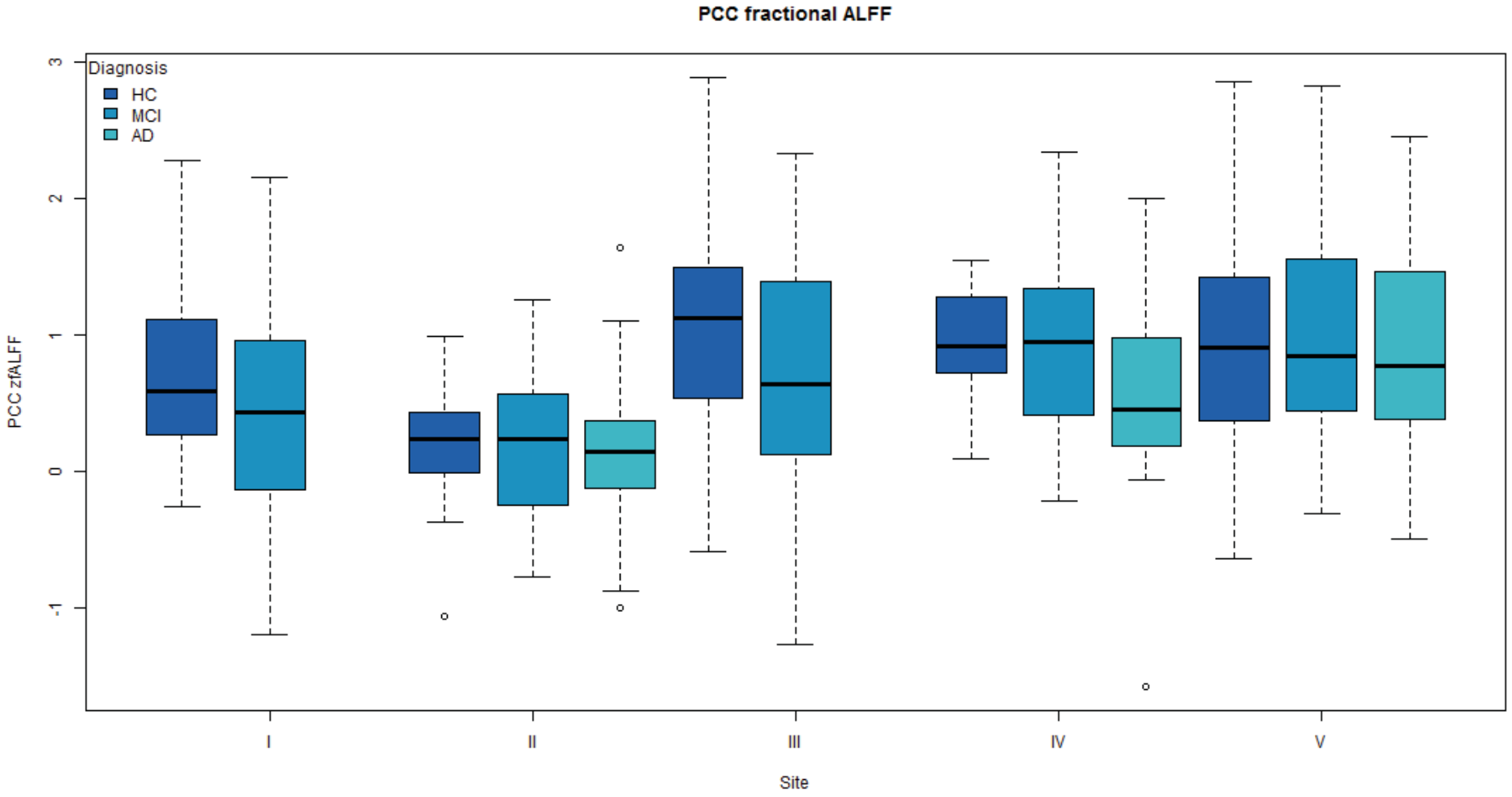
Supplementary Figure 2: Percentage framewise displacement > 0.5 mm



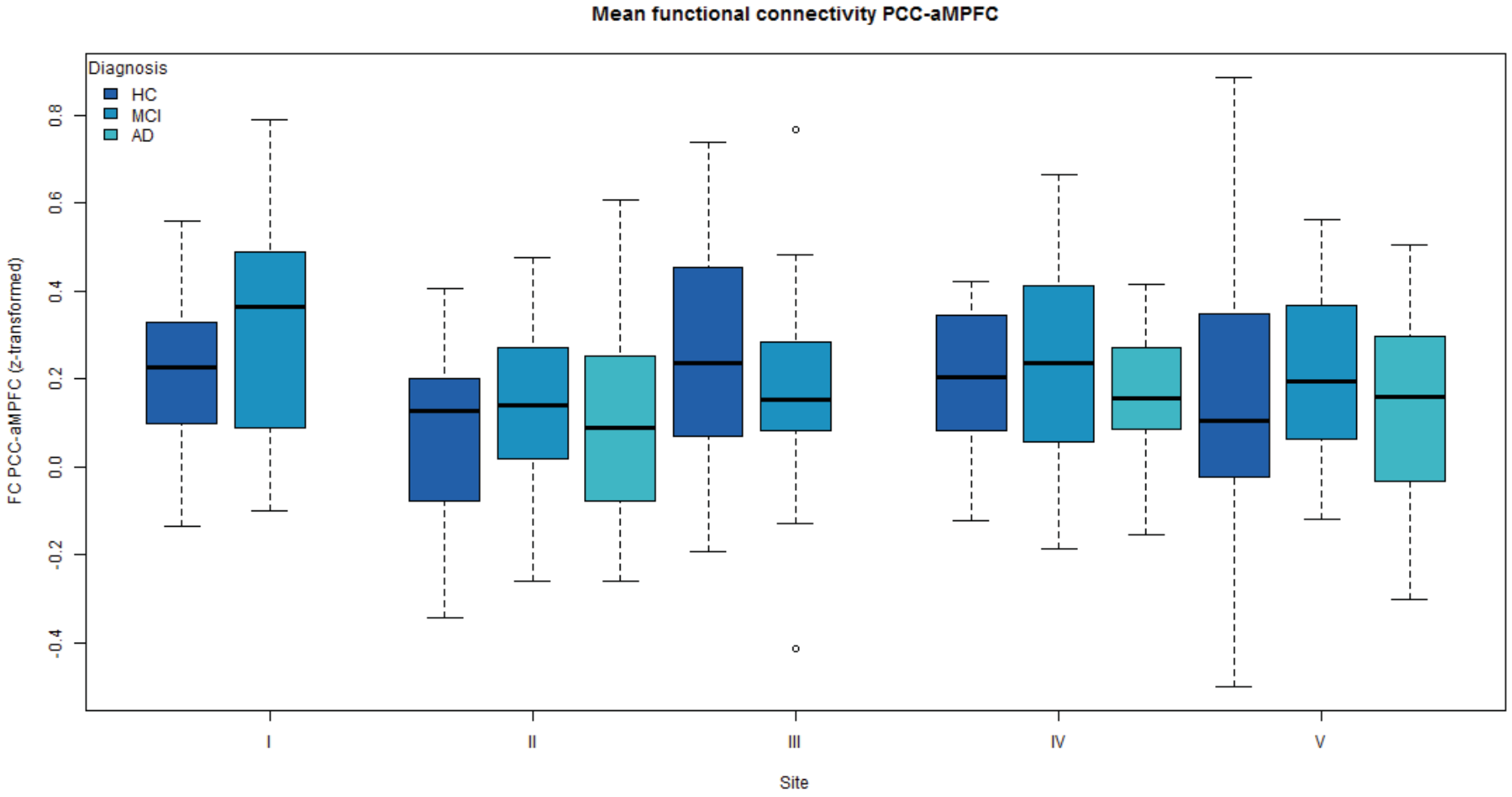
Supplementary Figure 3: Foreground to background Energy Ratio



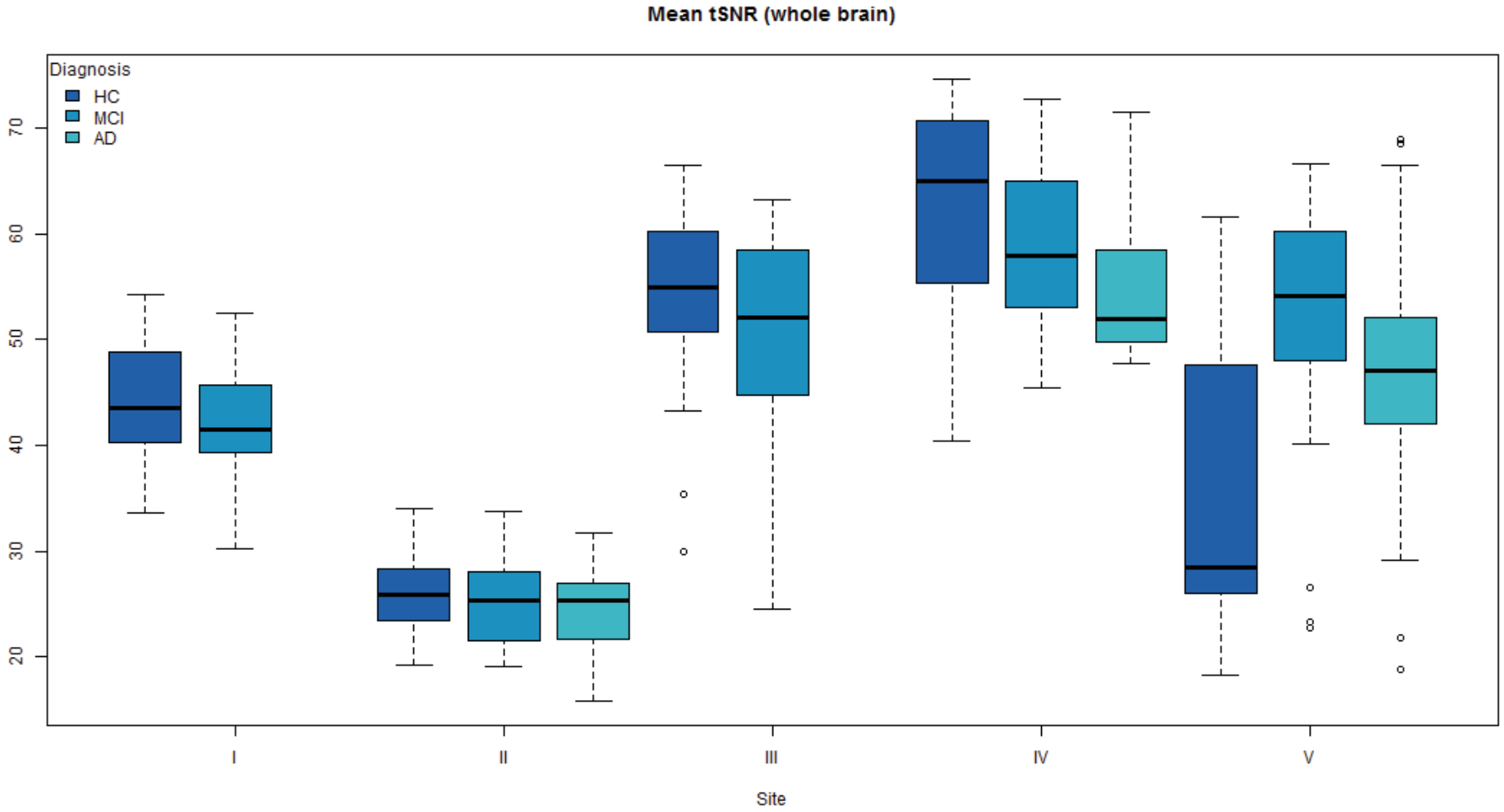
Supplementary Figure 4: Fractional amplitude of low frequency fluctuations in PCC



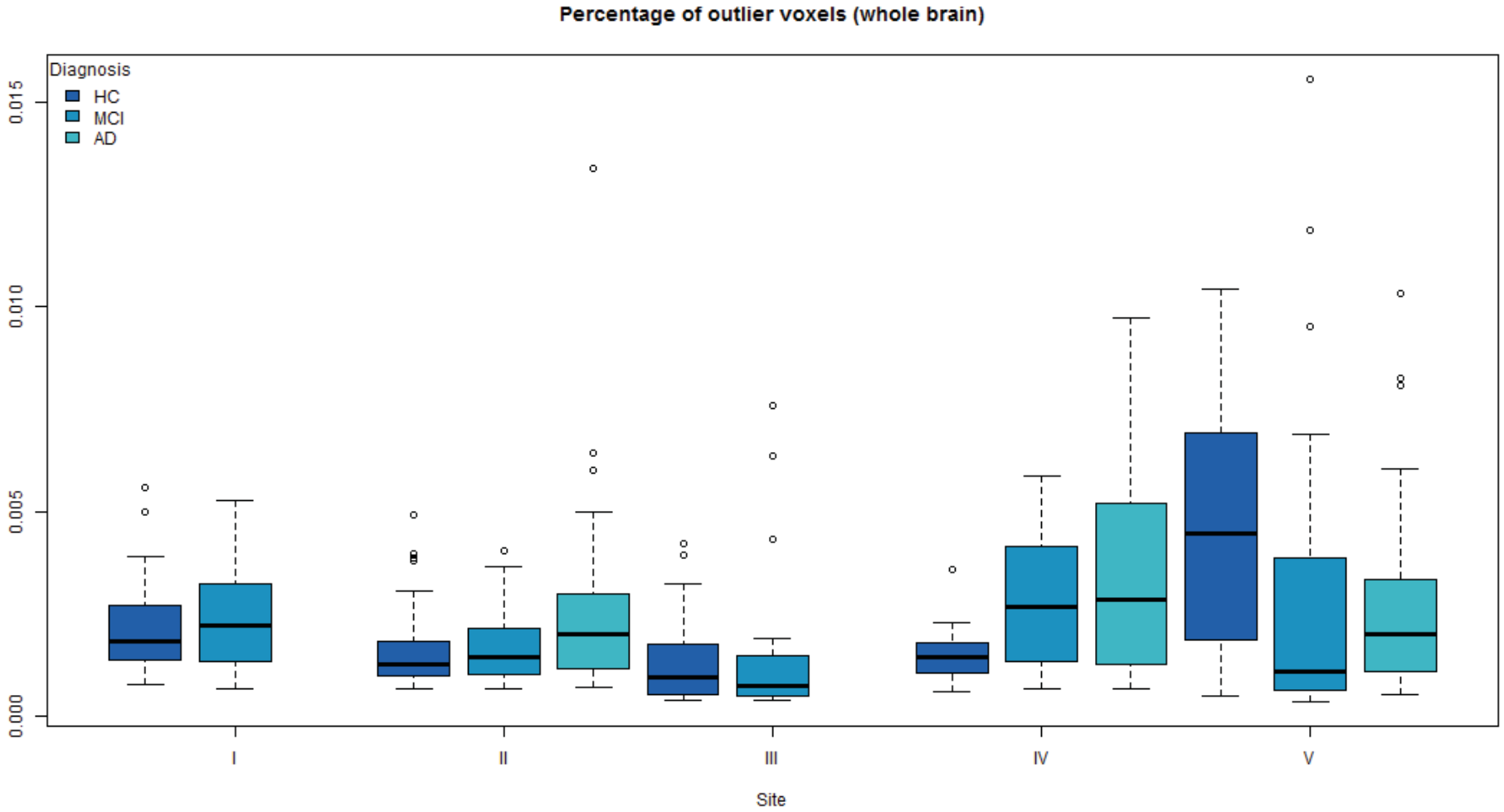
Supplementary Figure 5: Mean functional connectivity between PCC and anterior medioprefrontal cortex



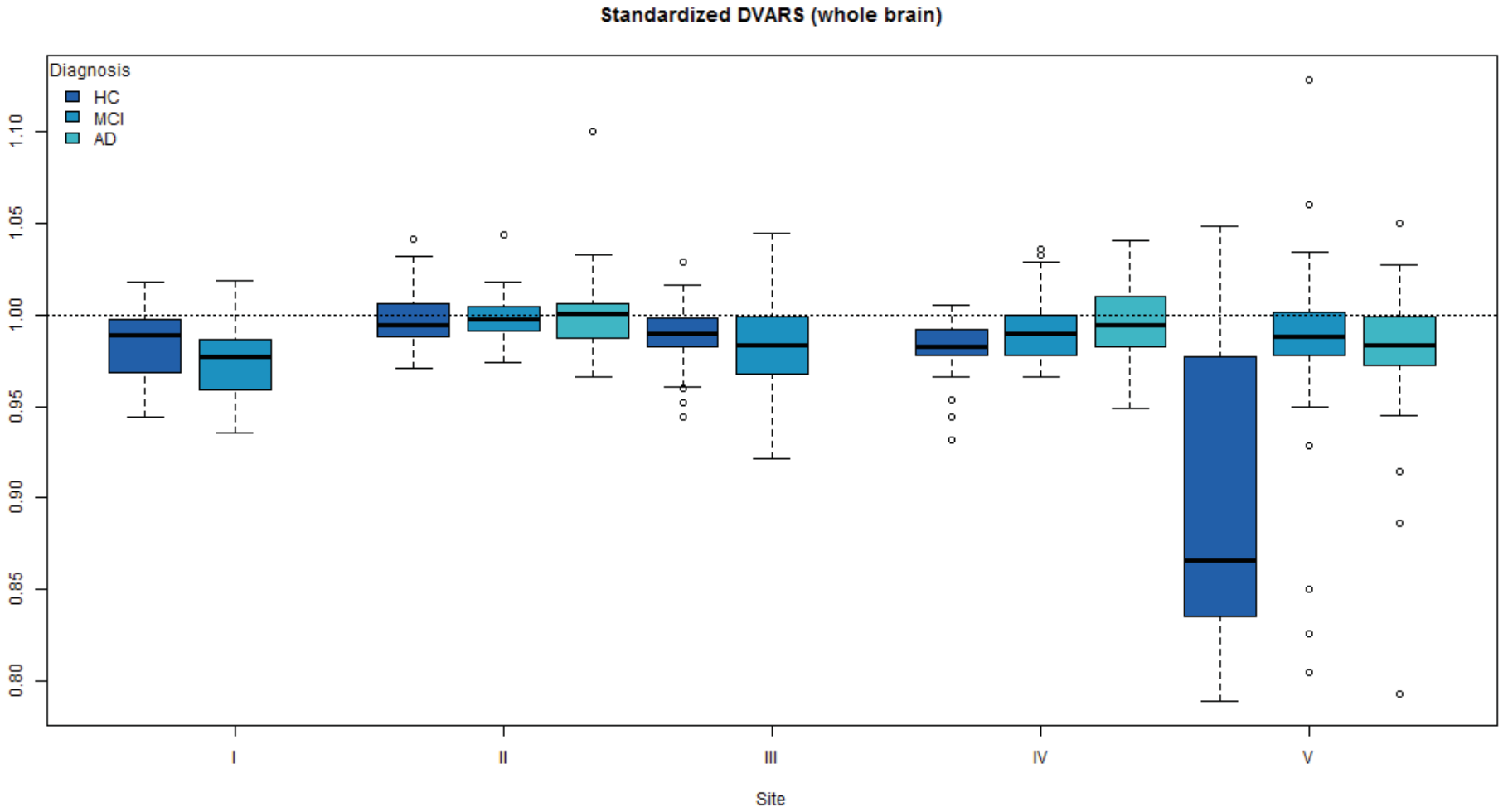
Supplementary Figure 6: Mean temporal signal to noise ratio



Supplementary Figure 7: Mean percentage of outlier voxels



Supplementary Figure 8: Standardized DVARS



frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 04 January 2017
doi: 10.3389/fnagi.2016.00318

Robust Detection of Impaired Resting State Functional Connectivity Networks in Alzheimer's Disease Using Elastic Net Regularized Regression

Stefan J. Teipel[1,2], Michel J. Grothe[2], Coraline D. Metzger[3,4], Timo Grimmer[5], Christian Sorg[6,7,8], Michael Ewers[9], Nicolai Franzmeier[9], Eva Meisenzahl[10], Stefan Klöppel[11,12], Viola Borchardt[13], Martin Walter[13,14] and Martin Dyrba[2]*

[1] Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany, [2] German Center for Neurodegenerative Diseases, Site Rostock/Greifswald, Rostock, Germany, [3] Institute of Cognitive Neurology and Dementia Research and Department of Psychiatry and Psychotherapy, Otto von Guericke University, Magdeburg, Germany, [4] German Center for Neurodegenerative Diseases, Site Magdeburg, Magdeburg, Germany, [5] Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technische Universität München, Munich, Germany, [6] Department of Neuroradiology of Klinikum rechts der Isar, Technische Universität München, Munich, Germany, [7] Department of Psychiatry of Klinikum rechts der Isar, Technische Universität München, Munich, Germany, [8] TUM-Neuroimaging Center, Technische Universität München, Munich, Germany, [9] Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians-Universität, Munich, Germany, [10] Department of Psychiatry, Klinikum der Universität München, Ludwig-Maximilians-Universität, Munich, Germany, [11] Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Faculty of Medicine, University of Freiburg, Freiburg, Germany, [12] University Hospital of Old Age Psychiatry, Bern, Switzerland, [13] Leibniz Institute for Neurobiology, Magdeburg, Germany, [14] Department of Psychiatry, University of Tübingen, Tübingen, Germany

OPEN ACCESS

Edited by:
Pedro Rosa-Neto,
McGill University, Canada

Reviewed by:
Ramesh Kandimalla,
Texas Tech University, USA
Haixian Wang,
Southeast University, China

***Correspondence:**
Stefan J. Teipel
stefan.teipel@med.uni-rostock.de

Received: 21 September 2016
Accepted: 09 December 2016
Published: 04 January 2017

Citation:
Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Klöppel S, Borchardt V, Walter M and Dyrba M (2017) Robust Detection of Impaired Resting State Functional Connectivity Networks in Alzheimer's Disease Using Elastic Net Regularized Regression.
Front. Aging Neurosci. 8:318.
doi: 10.3389/fnagi.2016.00318

The large number of multicollinear regional features that are provided by resting state (rs) fMRI data requires robust feature selection to uncover consistent networks of functional disconnection in Alzheimer's disease (AD). Here, we compared elastic net regularized and classical stepwise logistic regression in respect to consistency of feature selection and diagnostic accuracy using rs-fMRI data from four centers of the "German resting-state initiative for diagnostic biomarkers" (psymri.org), comprising 53 AD patients and 118 age and sex matched healthy controls. Using all possible pairs of correlations between the time series of rs-fMRI signal from 84 functionally defined brain regions as the initial set of predictor variables, we calculated accuracy of group discrimination and consistency of feature selection with bootstrap cross-validation. Mean areas under the receiver operating characteristic curves as measure of diagnostic accuracy were 0.70 in unregularized and 0.80 in regularized regression. Elastic net regression was insensitive to scanner effects and recovered a consistent network of functional connectivity decline in AD that encompassed parts of the dorsal default mode as well as brain regions involved in attention, executive control, and language processing. Stepwise logistic regression found no consistent network of AD related functional connectivity decline. Regularized regression has high potential to increase

diagnostic accuracy and consistency of feature selection from multicollinear functional neuroimaging data in AD. Our findings suggest an extended network of functional alterations in AD, but the diagnostic accuracy of rs-fMRI in this multicenter setting did not reach the benchmark defined for a useful biomarker of AD.

Keywords: regularization, diagnostic imaging, feature selection, functional magnetic resonance imaging (fMRI), Alzheimer's disease

INTRODUCTION

Many studies have identified altered functional connectivity networks in resting state examinations of Alzheimer's disease (AD) patients compared to controls using functional imaging techniques such as FDG-PET or resting state functional MRI (rs-fMRI) (for a recent review see Teipel et al., 2016). Typically AD dementia impairs functional connectivity in the default mode network (DMN; Greicius et al., 2004), but AD pathological changes and ensuing functional disruptions have been shown to extend beyond the regions of the DMN (Agosta et al., 2012; Grothe et al., 2016).

To identify the network characteristics of AD-related changes in functional imaging data, most studies have employed stepwise or multiple linear regression approaches (Agosta et al., 2012; Koch et al., 2012; Sheline and Raichle, 2013). However, features from rs-fMRI and other functional imaging data are often highly collinear across regions, and linear regression approaches are known to be highly sensitive toward collinearity (James et al., 2013; Section 3.3.6). In the presence of a high number of features relative to the number of available observations (Tibshirani, 2011) and when features are collinear (Hoerl and Kennard, 1970; Tibshirani, 1996), regularization techniques have been established for dimension reduction and feature selection. More recently, regularized models, using an elastic net penalty (Zou and Hastie, 2005; Zou and Zhang, 2009), have been applied to multimodal neuroimaging studies to reduce the effect of multicollinearity on feature selection (Trzepacz et al., 2014; Teipel S. J. et al., 2015; Schouten et al., 2016; de Vos et al., 2016).

Here, we used rs-fMRI data from a multicenter study to compare accuracy of group separation, as well as stability of regional feature selection and ensuing identification of cortical networks discriminating AD patients and controls between cross-validated regularized logistic regression with an elastic net penalty and classical stepwise logistic regression. We hypothesized that elastic net logistic regression would lead to more generalizable feature selection and more consistent network identification than classical stepwise logistic regression. Of note, the principles of these methods, except the elastic net penalty, represent textbook knowledge from statistical learning literature, but adoption of these methods to the burning issue of highly collinear features in neuroimaging research is still slow.

MATERIALS AND METHODS

For the current study, we used data from 53 patients with clinically probable AD according to NINCDS-ADRCA criteria (McKhann et al., 1984) and 118 healthy elderly control individuals that have been retrieved retrospectively from four sites within the framework of the "German resting-state initiative for diagnostic biomarkers" (http://www.psymri.org). Distribution of demographic characteristics of participants across sites is summarized in **Table 1**.

All participants were free of any significant neurological, psychiatric, or medical condition (except for AD in patients), in particular cerebrovascular apoplexy, vascular dementia, depression, or subclinical hypothyroidism, as well as substance abuse. Healthy controls were required to have no cognitive complaints and scored within one standard deviation of the age and education adjusted norm in all subtests of the Consortium to Establish a Registry of Alzheimer's Disease (CERAD) cognitive battery (Morris et al., 1989).

Written informed consent was provided by all subjects, or their representatives. The study was approved by local ethics committees at each of the participating centers, and has been conducted in accord with the Helsinki Declaration of 1975.

Imaging and Data Acquisition

The data used in this study were obtained from four different 3.0 Tesla MRI scanners. Acquisition parameters for the rs-fMRI sequences are given in **Table 2**. In one center (site I), the subjects were instructed to keep their eyes open, whereas in the remaining centers (sites II-IV) all subjects were requested to close their eyes, relax, but not to fall asleep. Functional MRI was based on echo-planar imaging using scan durations between 6 and 8.7 min for the rs-fMRI sequence. The number of acquired time points was between 120 and 200 with a voxel size ranging from $2 \times 2 \times 2.6$ up to $3.28 \times 3.28 \times 4.4$ mm^3 (**Table 2**). For anatomical reference, high-resolution T1-weighted gradient echo sequences

TABLE 1 | Demographic characteristics.

	AD	Controls
No. cases (women)[a]	53 (31)	118 (61)
Age (SD) [years][b]	72.4 (8.8)	70.4 (6.2)
MMSE (SD), number[c]	22.5 (4.4), 53	28.8 (1.0) 97
MoCA (SD), number	–	26.4 (2.1), 19
Education (SD) [years][d]	11.4 (2.1)	13.6 (3.1)

MMSE, Mini Mental State Examination (Folstein et al., 1975); MoCA, Montreal Cognitive Assessment (Nasreddine et al., 2005).

[a] *Not significantly different between groups,* $\chi^2 = 0.68$, *1 df,* $p = 0.41$.

[b] *Not significantly different between groups,* $t = 1.67$, *169 df,* $p = 0.96$.

[c] *significantly different between groups, Mann-Whitney U-test,* $p < 0.001$.

[d] *significantly different between groups,* $t = -4.72$, *168 df,* $p < 0.001$.

TABLE 2 | Scanner characteristics.

Center	Model	Manufacturer	TR [s]	TE [s]	Volumes	Voxel size [mm^3]	Gap [mm]
I	TrioTim	Siemens	2.61	0.030	200	3 × 3 × 3.6	0.6
II	Verio	Siemens	3	0.030	120	2 × 2 × 2.6	0.6
III	Verio	Siemens	2.58	0.030	180	3.5 × 3.5 × 3.5	0
IV	Trio	Siemens	3	0.030	120	3.28 × 3.28 × 4.4	0.4

with an isotropic resolution of 1 mm^3 were also obtained from all scanners during the same session.

MR Processing

Functional MRI data processing was carried out using Data Processing Assistant for Resting-State fMRI (DPARSF 3.2) (Chao-Gan and Yu-Feng, 2010), considering the recommendations from a recent systematic evaluation of processing alternatives (Shirer et al., 2015). After the removal of the first six images to account for gradient field stabilization, the rs-fMRI data was slice time corrected and realigned to the temporal mean image. The anatomical T_1-weighted image of each participant was coregistered to the mean functional image and subsequently segmented into gray matter, white matter, and cerebrospinal fluid (CSF) partitions using the Voxel-based Morphometry (VBM8) toolbox (Gaser et al., 1999) that extends Statistical Parametric Mapping (SPM8) (Friston et al., 2007). The Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm (Ashburner, 2007) was applied to normalize the T_1-weighted images to the Montreal Neurological Institute (MNI) reference coordinate system using the default brain template included in VBM8. The deformation fields generated by DARTEL were used to project the functional scans from each subjects' native image space into the MNI reference space. We combined this step with the reslicing of all functional data to an isotropic resolution of 3 mm. The subsequent nuisance regression included covariates of head movement (rotation, translation, and first and second order derivatives) and the mean time courses for the global brain signal, the white matter segment signal, and the CSF segment signal. Although global signal regression was found to introduce negative correlations (Murphy et al., 2009; Shirer et al., 2015), studies consistently reported that it effectively increases the signal-to-noise ratio (Yan et al., 2013; Power et al., 2014; Shirer et al., 2015). Recently, Shirer et al. evaluated the influence of global signal regression on group separation but only found a minor, non-significant effect (Shirer et al., 2015). Subsequently, the images were band-pass filtered using the frequency band 0.1–0.01 Hz. For each individual the time series of signal was extracted for each of the 84 functionally defined regions of the Greicius atlas (Shirer et al., 2012). Pearson's correlation coefficients were computed for the 3486 possible pairs of correlations between these 84 regions (Shirer et al., 2012). Finally, Pearson correlation coefficients of the signal time courses were adjusted to be normally distributed using Fisher's Z-transform (Fisher, 1915): $z = 0.5 \ln [(1+r)/(1-r)]$.

Statistical Analysis

Demographic Characteristics

Baseline demographic characteristics were compared between AD and control cases using parametric and non-parametric tests as required: age and years of education were compared between groups using Student's *t*-test, gender distribution using Chi^2 test, and neuropsychological test results using non-parametric Mann-Whitney *U*-test.

Prediction of Group Membership

We compared two regression models for prediction of group membership (AD vs. controls) in respect to two outcomes, (i) the accuracy of prediction as determined by the area under the receiver operating characteristics curve (AUC), and (ii) the consistency of feature selection.

The two regression models encompassed:

1. bidirectional (backward and forward) stepwise unpenalized logistic regressions using the function *step* in R (The R Foundation for Statistical Computing). The function weights the choices via the Akaike information criterion (AIC), which takes account of the total number of fitted parameters.
2. penalized logistic regression models with an elastic net penalty, as determined using the R package *glmnet* (available at http://cran.r-project.org/web/packages/glmnet/index.html). Elastic net regression is controlled by two parameters, (i) alpha, which sets the degree of mixing between two types of regularized regression, namely ridge regression (regularization by squared L_2 norm; alpha = 0) and the Lasso (Least Absolute Shrinkage and Selection Operator, regularization by L_1 norm; alpha = 1), and (ii) lambda, defining the strength of regularization (Friedman et al., 2010). Alpha was selected to be 0.5, corresponding to a full elastic net penalty, which minimized the partial likelihood deviance of the model (see **Figure 1**). Lambda was determined using grid search with 100-fold cross-validation. The optimal lambda was determined as the mean across 100 iteratively determined lambda values minimizing the deviance of the model. The optimal lambda value was determined for each bootstrap iteration in the training data and applied to the test data as defined below. Details of this method can be found in the appendix.

Both models were determined using strict cross-validation procedures. Random samples of 2/3 of the data were drawn 1000 times to train the prediction models (training data). For both regression models, the prediction accuracy was determined using the remaining 1/3 as test data. Parameter optimization,

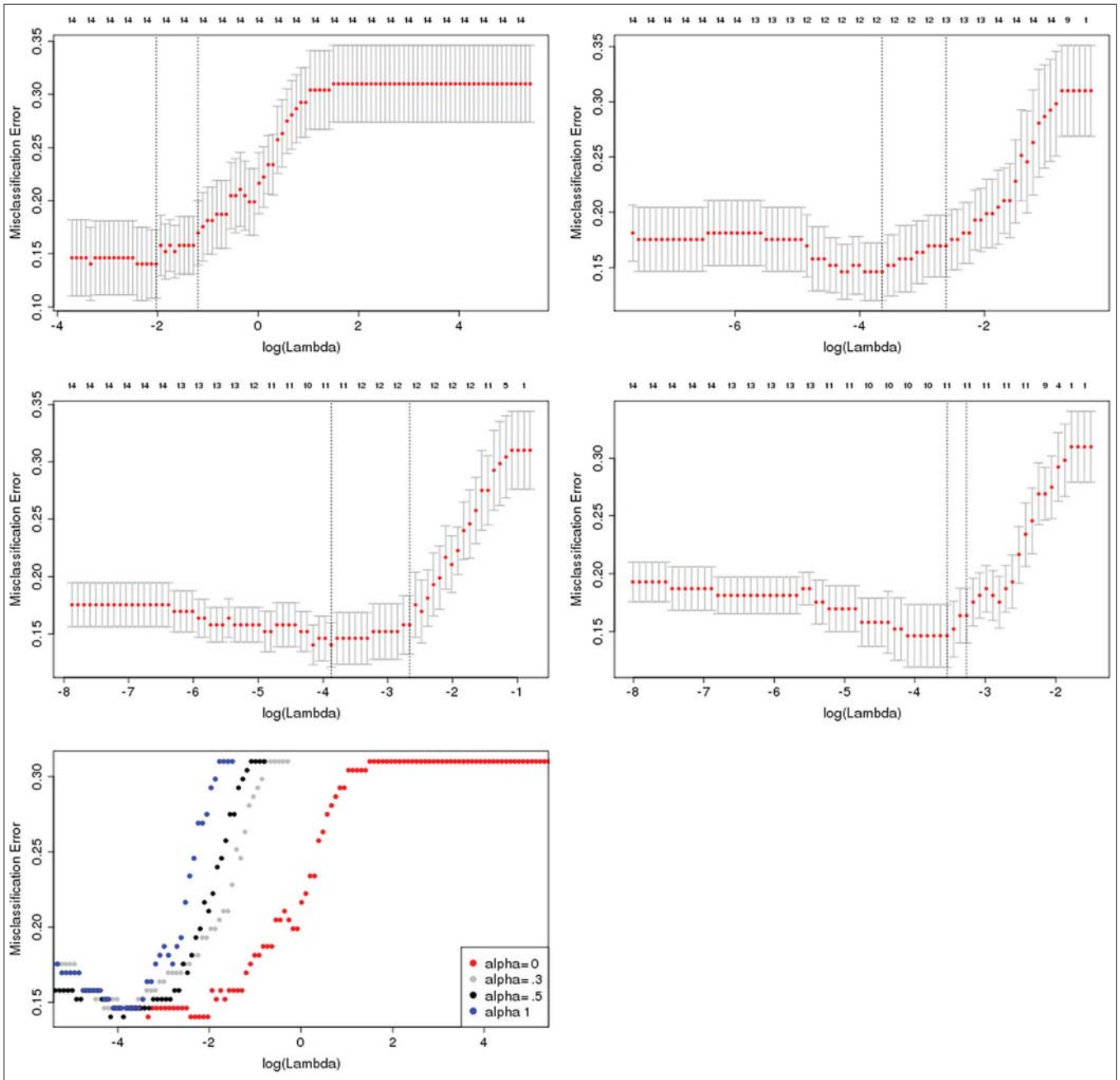


FIGURE 1 | Selection of alpha parameter for penalized logistic regression. Misclassification error plotted against the range of lambda values (plotted on a logarithmic scale) for different values of α for a penalized logistic regression on the rs-fMRI data. Numbers on top of each graph indicate the number of selected variables. Error bars indicate the bootstrapped standard deviation for the misclassification error for each lambda value. The left bottom plot shows the different deviance curves on a unified scale, indicating that $\alpha = 0.5$ yields the lowest deviance together with $\alpha = 0$, corresponding to a ridge regression model.

i.e., selection of optimal lambda and (stepwise) feature selection, was conducted in the training data and subsequently applied to the test data. Prior to model building, the feature space was restricted through determining the set of variables which correlate with diagnosis with a Pearson's correlation coefficient of $|r| > 0.35$ in the training data, resulting in an average number of 36 included predictor variables across the bootstrapped repetitions.

In a second analysis, dummy coded center variables were forced as additional variables into the models to determine the effect of center on model accuracy and feature selection.

To check for multicollinearity of the stepwise logistic regression models, we determined the variance inflation factor (VIF) (Belsley, 1991) for each independent variable on the set of the remaining independent variables using the function vif

in R package "car" (available at https://cran.r-project.org/web/packages/car/index.html).

RESULTS

Demographic Characteristics

Demographic characteristics are summarized in **Table 1**. AD patients and controls were not significantly different in age ($t = 1.67$, 169 df, $p = 0.96$) or sex distribution ($Chi^2 = 0.68$, 1 df, $p = 0.41$). Both groups differed significantly in years of education ($t = -4.72$, 168 df, $p < 0.001$), with less years of education in the AD cases, and, as expected, AD patients scored significantly lower than healthy controls in the MMSE score ($p < 0.001$).

Prediction of Group Membership

The median VIF across all stepwise regression models and variables was 86, indicating a very high collinearity in the large majority of models. Mean area under the ROC curves in the test samples was 70% for the stepwise selection, and 80% for the elastic net regression models for the discrimination between AD cases and controls. The mean AUC and corresponding 2.5/97.5 percentile confidence intervals for both models are shown in **Figure 2**. The selected features are shown in **Table 3** for both models, with seven features selected in at least 50% of 1000 cross-validation repetitions for the elastic net and two features selected for the stepwise logistic regression model. **Figures 3**, **4** show the frequency distribution of feature selection, suggesting that features were more homogeneously and more often selected in the cross-validation repetitions for the elastic net compared to the stepwise logistic regression models, with a median value of 10 features with the stepwise regression and 22 features in the elastic net regression.

When we repeated the analyses with dummy coded center covariates forced into the models, AUC was 81% for the elastic net penalty, and selected features above 50% frequency were unchanged. For the stepwise regression, AUC decreased to 68%, and no feature was selected with a frequency above 45%.

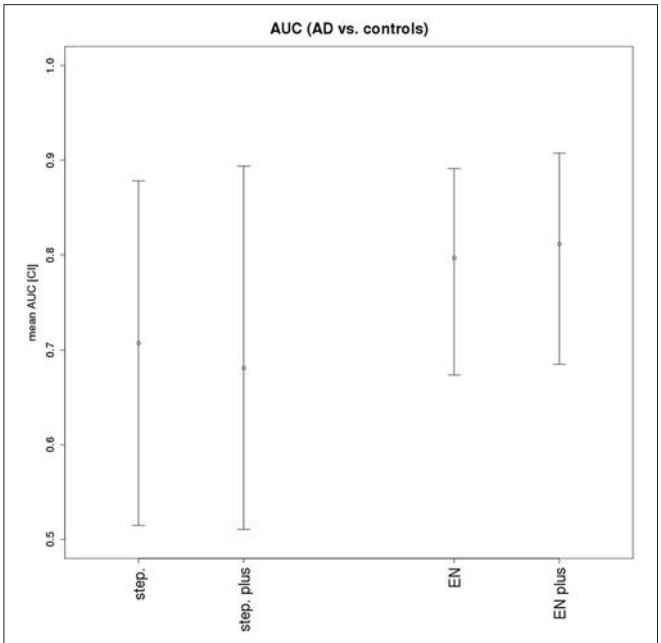


FIGURE 2 | Areas under the ROC curves for stepwise and elastic net logistic regression. AUC and 2.5/97.5 percentile confidence intervals for stepwise logistic regression without scanner (step.) and with scanner forced into the model (step. plus), and for elastic net logistic regression without scanner (EN) and with scanner forced into the model (EN plus).

DISCUSSION

In accordance with our hypothesis, we found more accurate group discrimination between AD dementia cases and controls and more homogeneous feature selection from resting state fMRI data when using regularized logistic regression with an elastic net penalty compared with a classical stepwise logistic regression. These findings support the notion that regularized regression is superior to classical stepwise feature selection for dealing with highly collinear multidimensional functional imaging data. The features retrieved from penalized regression point to alterations of an extended functional network in mild AD dementia, compromising the dorsal DMN, but also key regions for language processing, object recognition and attention.

As illustrated by the high VIF with a median value of 86 (values above five are considered indicative for serious multicollinearity; Belsley, 1991), the regional rs-fMRI values exhibited a high degree of collinearity that compromised unbiased feature selection and determining the relevance of single features. The problem of dealing with multidimensional, multicollinear data is well-known in the statistical literature under the term of "the curse of dimensionality" (Bellman, 1961). Penalized regression has been developed since the 1940s to deal with this problem, encompassing techniques like ridge regression (Hoerl, 1970), the Lasso (Tibshirani, 1996), and more recently elastic net regression (Zou and Hastie, 2005), which combines both regularization techniques within the same model. Different to ridge regression, and similar to the Lasso, elastic net regression not only shrinks the feature coefficients but sets some of the coefficients to zero, thus reducing the dimensionality of the feature space. Different to the Lasso, elastic net regression is designed to select highly correlated features as a group rather than selecting only a single feature out of such a set of highly correlated variables, thus preserving a potentially meaningful correlation structure of the original feature space (Zou and Hastie, 2005).

Previous neuroimaging studies have successfully applied elastic net regression to multimodal neuroimaging data for feature selection for dementia prediction in subjects with mild cognitive impairment (MCI), and AD cases (Trzepacz et al., 2014; Teipel S. J. et al., 2015; de Vos et al., 2016). A previous study has applied this approach to rs-fMRI data of people with mild to moderate AD dementia from one scanner (Schouten et al., 2016), reaching 77% accuracy in the mild AD subgroup. In our multicenter study, cross-validated accuracy of 80% discrimination between AD cases and controls from elastic net regression was higher than the accuracy in this previous study

TABLE 3 | Selected features.

Frequency [%]	Anatomical regions	Functional networks (Shirer et al., 2012)
FEATURES FROM ELASTIC NET		
94.5	Left/right gyrus temporalis superior	Auditory network
87.1	Right gyrus frontalis superior <-> left gyrus occipitalis medialis	Basal ganglia network <-> visuospatial network
79.8	Left gyrus frontalis medialis <-> bilateral precuneus	Anterior salience network <-> precuneus network
69.9	Left/right precentral gyrus	Sensorimotor network
68.7	Right gyrus frontalis inferior <-> cingulate gyrus body	Anterior salience network <-> dorsal DMN
60.9	Right gyrus angularis <-> right gyrus frontalis medialis	Dorsal DMN <-> right executive control network
59.9	Bilateral anterior cingulate gyrus/ left gyrus frontalis superior/left gyrus frontalis medialis <-> left lobulus parietalis inferior/superior	Dorsal DMN <-> left executive control network
FEATURES FROM STEPWISE LOGISTIC REGRESSION		
56.9	Left/right precentral gyrus	Sensorimotor network
53.8	Right gyrus frontalis superior <-> left gyrus occipitalis medialis	Basal ganglia network <-> visuospatial network

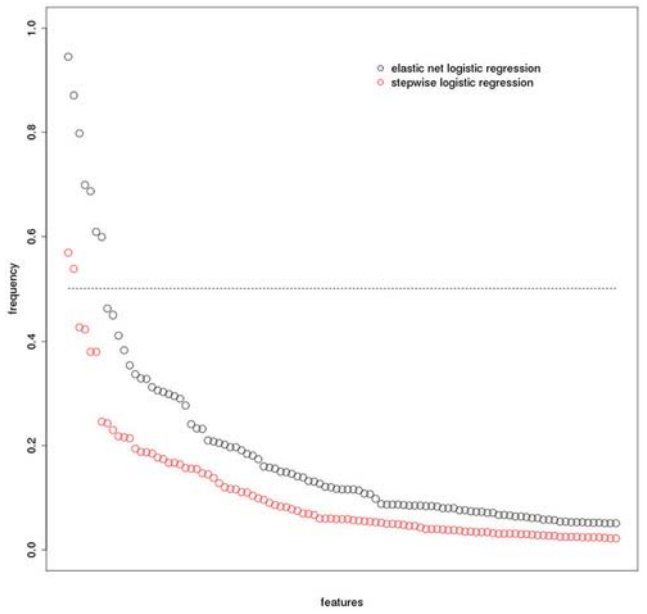


FIGURE 3 | Feature selection frequency plot. Frequency of selected features (based on 1000 bootstrap iterations) for elastic net and stepwise logistc regression. Please note that the x-axis represents the features that were sorted according to their frequency independently within each model. Therfore, the same position on the x-axis does not indicate the same feature for the elastic net and the stepwise logistic regression models, respectively.

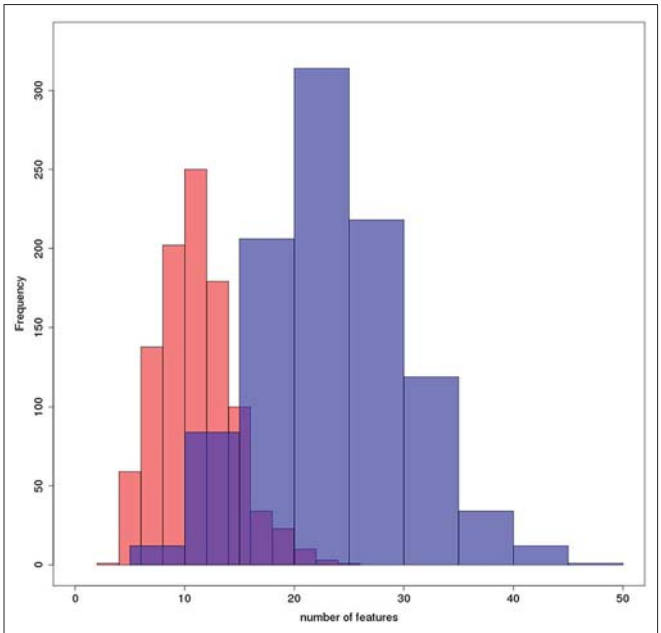


FIGURE 4 | Number of features selected per model. Histograms plotting the frequency with which a number of features was selected across all bootstrapping iterations for elastic net (blue) and stepwise logistc regression (red).

(Schouten et al., 2016), but still lower than results from previous monocenter studies lacking cross-validation (Koch et al., 2012; Balthazar et al., 2014). Our findings level of accurcy agrees with estimates from previous cross-validated monocenter studies using non-linear machine learning techniques for classification (Challis et al., 2015; Dyrba et al., 2015). One recent study yielded 100% group discrimination between 20 AD patients and 20 controls using support vector machine classification (Khazaee et al., 2015). From the method description, however, the feature selection prior to cross-validated machine learning was based on the entire data set and was not part of the cross-validation.

Features selected by the elastic net regression were more consistent across repeated cross-validations than features selected by stepwise regression. Previous research on rs-fMRI in AD dementia has often focused on the DMN regions (Greicius et al., 2004; Koch et al., 2012; Balthazar et al., 2014). This approach reduces potential problems from collinearity through a priori feature selection. At the same time, it restricts the analysis to a single preselected functional network. Using elastic net regression, we retrieved the dorsal part of the DMN as key part of altered functional connectivity in AD. This agrees with previous analyses based on preselected DMN regions (Greicius et al., 2004; Koch et al., 2012; Balthazar et al., 2014) and underscores the overall validity of our approach. In addition, we found decreased

functional connectivity in AD in the superior temporal gyrus, a region that is involved in language processing (Zhuang et al., 2014), and prefrontal parts of the salience network, prefrontal and parietal components of executive control networks, as well as the medial occipital gyrus as part of the ventral visual stream involved in object recognition (Teipel et al., 2007) and recognition of limb movements (Astafiev et al., 2004). These findings support the extended nature of AD pathology affecting several higher order cognitive networks, as previously found in topographic lesion driven studies (Grothe et al., 2016) and rs-fMRI analysis in small samples of 12 to 16 AD cases and 12–22 controls (Zhou et al., 2010; Agosta et al., 2012; Dai et al., 2012), and one large scale study (Brier et al., 2012). Different to two of these previous studies (Zhou et al., 2010; Agosta et al., 2012), we found only reductions, but no increases of functional connectivity in AD. This difference may have two possible causes. The first possible cause would be different severity of disease within the dementia stage of AD. However, the MMSE scores were similar between the AD cases of our and the previous studies. Another cause may be the different metric used as prediction features: we used correlation between regions irrespective of preselected networks, whereas the previous studies used regional loadings on independent components associated with specific functional networks (Agosta et al., 2012; Zhou et al., 2010).

Compared with elastic net regression, stepwise regression yielded only 70% accuracy. In addition, selection of the most relevant features was much less consistent across the 1000 iterations, compromising only two functional connections between sensorimotor and visuospatial regions, and no connection involving the DMN. These findings suggest that feature selection in step-wise regression was more sensitive to multicollinearity, where small differences in explained variance drive almost arbitrarily selection of a single feature among a set of highly collinear variables (Farrar and Glauber, 1967).

Stepwise logistic regression was sensitive to scanner effects, with a slight drop in prediction accuracy and a further loss of consistency in feature selection when scanner was forced into the model. In contrast, elastic net regression was insensitive to scanner effects; both accuracy of group discrimination and frequency of feature selection were unaffected when we repeated the analyses with scanner forced into the cross-validated models. This finding is of particular relevance given the sensitivity of rs-fMRI data to multiscanner effects, as has been reported in test-retest studies of rs-fMRI even in healthy people repeatedly scanned at the same scanner (Meindl et al., 2010; Chen et al., 2015; Lin et al., 2015; Orban et al., 2015; Shirer et al., 2015; Jovicich et al., 2016), including long-term evaluation after more than 12 months (Chou et al., 2012; Guo et al., 2012; Blautzik et al., 2013). Moreover, the use of multiple scanners typically results in high variability of signal-to-noise and contrast-to-noise ratios, particularly when using field strengths of 3T and higher (Magnotta et al., 2006; Lin et al., 2015; Jovicich et al., 2016).

We need to consider two main limitations of our study. First, the scan protocols were different between scanners. Multiscanner acquisition helps to increase sample size, a problem of many previous monocenter studies. In addition, estimates of accuracy derived from multicenter data may more easily generalize to future use of an imaging technology in routine care than estimates derived from single center data acquisition. We employed preprocessing steps that had been shown in a previous study to reduce multiscanner effects (Shirer et al., 2015), and used cross-correlation of regional signal time series which in a previous study had yielded more stable results across scanners than other connectivity metrics, such as cross-coherence or partial cross-correlation (Fiecas et al., 2013). Secondly, the reference standard in our sample was a clinical diagnosis of AD dementia, but independent PET or CSF based biomarker validation was not available in the majority of cases. Data came from expert centers experienced in the early diagnosis of AD. Still, a final judgment of the added value of rs-fMRI for AD diagnosis must await systematic evaluation of diagnostic accuracy in multicenter data from biomarker stratified cases.

In summary our findings point to an extended network of functional disconnection, including the dorsal DMN, but also involving functional networks employed in attention, object recognition and language processing. In a multicenter sample of AD and control cases, elastic net regression yielded cross-validated diagnostic accuracy that approached, but did not reach, the benchmark for a useful biomarker of AD (Consensus-Group, 1998); diagnostic approaches based on stepwise regression came not even close to this benchmark. These findings question the future wide-spread use of rs-fMRI as a stand-alone diagnostic marker of AD (Teipel S. et al., 2015). This does not exclude an important role of rs-fMRI as add-on diagnostic marker (Dai et al., 2012) and to identify mechanisms of functional disconnection and resilience in future prospective studies. Our data suggest that regularized regression should be preferred over still more widely used but less robust stepwise feature selection to retrieve homogeneous and stable estimates of altered functional networks in AD.

ETHICS STATEMENT

Institutional Review Board of the University Medicine Rostock. Written informed consent was provided by all subjects, or their representatives. The study was approved by local ethics committees at each of the participating centers, and has been conducted in accord with the Helsinki Declaration of 1975. For people with dementia, informed consent involves oral and written presentation of study procedures, and information of caregivers.

AUTHOR CONTRIBUTIONS

ST: Conception of the work; acquisition, analysis, and interpretation of data for the work; drafting the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. MG, MD: Acquisition, analysis, and interpretation of data for the work; revising the work critically for important intellectual content; final approval

of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. CM, TG, CS, ME, NF, EM, SK, VB, MW: Interpretation of data for the work; revising the work critically for important intellectual content; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

ST received support by a grant of the Federal Ministry of Research (BMBF) (AgeGain, 1GQ1425B).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnagi.2016.00318/full#supplementary-material

REFERENCES

Agosta, F., Pievani, M., Geroldi, C., Copetti, M., Frisoni, G. B., and Filippi, M. (2012). Resting state fMRI in Alzheimer's disease: beyond the default mode network. *Neurobiol. Aging* 33, 1564–1578. doi: 10.1016/j.neurobiolaging.2011.06.007

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Astafiev, S. V., Stanley, C. M., Shulman, G. L., and Corbetta, M. (2004). Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nat. Neurosci.* 7, 542–548. doi: 10.1038/nn1241

Balthazar, M. L., de Campos, B. M., Franco, A. R., Damasceno, B. P., and Cendes, F. (2014). Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease. *Psychiatry Res.* 221, 37–42. doi: 10.1016/j.pscychresns.2013.10.010

Bellman, R. (1961). *Adaptive Control Processes. A Guided Tour*. Princeton, NJ: P*rinceton University Press.*

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Chichester: John Wiley & Sons.

Blautzik, J., Keeser, D., Berman, A., Paolini, M., Kirsch, V., Mueller, S., et al. (2013). Long-term test-retest reliability of resting-state networks in healthy elderly subjects and with amnestic mild cognitive impairment patients. *J. Alzheimers. Dis.* 34, 741–754. doi: 10.3233/JAD-111970

Brier, M. R., Thomas, J. B., Snyder, A. Z., Benzinger, T. L., Zhang, D., Raichle, M. E., et al. (2012). Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression. *J. Neurosci.* 32, 8890–8899. doi: 10.1523/JNEUROSCI.5698-11.2012

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., and Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *Neuroimage* 112, 232–243. doi: 10.1016/j.neuroimage.2015.02.037

Chao-Gan, Y., and Yu-Feng, Z. (2010). DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Front. Syst. Neurosci.* 4:13. doi: 10.3389/fnsys.2010.00013

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., et al. (2015). Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS ONE* 10:e0144963. doi: 10.1371/journal.pone.0144963

Chou, Y. H., Panych, L. P., Dickey, C. C., Petrella, J. R., and Chen, N. K. (2012). Investigation of long-term reproducibility of intrinsic connectivity network mapping: a resting-state fMRI study. *AJNR Am. J. Neuroradiol.* 33, 833–838. doi: 10.3174/ajnr.A2894

Consensus-Group (1998). Consensus report of the working group on: "Molecular and Biochemical Markers of Alzheimer's Disease." The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group. *Neurobiol. Aging* 19, 109–116.

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., et al. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *Neuroimage* 59, 2187–2195. doi: 10.1016/j.neuroimage.2011.10.003

de Vos, F., Schouten, T. M., Hafkemeijer, A., Dopper, E. G., van Swieten, J. C., de Rooij, M., et al. (2016). Combining multiple anatomical MRI measures improves Alzheimer's disease classification. *Hum. Brain Mapp.* 37, 1920–1929. doi: 10.1002/hbm.23147

Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* 36, 2118–2131. doi: 10.1002/hbm.22759

Farrar, D. E., and Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* 49, 92–107. doi: 10.2307/1937887

Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A., and Feng, D. (2013). Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fMRI. *Neuroimage* 65, 231–241. doi: 10.1016/j.neuroimage.2012.09.052

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507. doi: 10.2307/2331838

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental-state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W. D. (2007). *Statistical Parametric Mapping: The Analysis of Funtional Brain Images, 1st Edn* (Amsterdam; Boston, MA: Elsevier/Academic Press).

Gaser, C., Volz, H.-P., Kiebel, S., Riehemann, S., and Sauer, H. (1999). Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research. *Neuroimage* 10, 107–113. doi: 10.1006/nimg.1999.0458

Greicius, M. D., Srivastava, G., Reiss, A. L., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4637–4642. doi: 10.1073/pnas.0308627101

Grothe, M. J., Teipel, S. J., and Alzheimer's Disease Neuroimaging Initiative (2016). Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks. *Hum. Brain Mapp.* 37, 35–53. doi: 10.1002/hbm.23018

Guo, C. C., Kurth, F., Zhou, J., Mayer, E. A., Eickhoff, S. B., Kramer, J. H., et al. (2012). One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. *Neuroimage* 61, 1471–1483. doi: 10.1016/j.neuroimage.2012.03.027

Hoerl, A. E. (1970). Ridge regression. *Biometrics* 26, 603.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression - biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

James, G. A., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York, NY: S*pringer*.

Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartres-Faz, D., et al. (2016). Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fMRI study. *Neuroimage* 124(Pt A), 442–454. doi: 10.1016/j.neuroimage.2015.07.010

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2015). Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin. Neurophysiol.* 126, 2132–2141. doi: 10.1016/j.clinph.2015.02.060

Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A. L., et al. (2012). Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiol. Aging* 33, 466–478. doi: 10.1016/j.neurobiolaging.2010.04.013

Lin, Q., Dai, Z., Xia, M., Han, Z., Huang, R., Gong, G., et al. (2015). A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. *Sci. Data* 2:150056. doi: 10.1038/sdata.2015.56

Magnotta, V. A., Friedman, L., and First, B. (2006). Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study. *J. Digit. Imaging* 19, 140–147. doi: 10.1007/s10278-006-0264-x

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* 34, 939–944. doi: 10.1212/WNL.34.7.939

Meindl, T., Teipel, S., Elmouden, R., Mueller, S., Koch, W., Dietrich, O., et al. (2010). Test-retest reproducibility of the default-mode network in healthy individuals. *Hum. Brain Mapp.* 31, 237–246. doi: 10.1002/hbm.20860

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., et al. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39, 1159–1165. doi: 10.1212/WNL.39.9.1159

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage* 44, 893–905. doi: 10.1016/j.neuroimage.2008.09.036

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Orban, P., Madjar, C., Savard, M., Dansereau, C., Tam, A., Das, S., et al. (2015). Test-retest resting-state fMRI in healthy elderly persons with a family history of Alzheimer's disease. *Sci. Data* 2:150043. doi: 10.1038/sdata.2015.43

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320–341. doi: 10.1016/j.neuroimage.2013.08.048

Schouten, T. M., Koini, M., de Vos, F., Seiler, S., van der Grond, J., Lechner, A., et al. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *Neuroimage Clin.* 11, 46–51. doi: 10.1016/j.nicl.2016.01.002

Sheline, Y. I., and Raichle, M. E. (2013). Resting state functional connectivity in preclinical Alzheimer's disease. *Biol. Psychiatry* 74, 340–347. doi: 10.1016/j.biopsych.2012.11.028

Shirer, W. R., Jiang, H., Price, C. M., Ng, B., and Greicius, M. D. (2015). Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. *Neuroimage* 117, 67–79. doi: 10.1016/j.neuroimage.2015.05.015

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb. Cortex* 22, 158–165. doi: 10.1093/cercor/bhr099

Teipel, S., Drzezga, A., Grothe, M. J., Barthel, H., Chetelat, G., Schuff, N., et al. (2015). Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. *Lancet Neurol.* 14, 1037–1053. doi: 10.1016/S1474-4422(15)00093-9

Teipel, S., Grothe, M. J., Zhou, J., Sepulcre, J., Dyrba, M., Sorg, C., et al. (2016). Measuring cortical connectivity in Alzheimer's Disease as a brain neural network pathology: toward clinical applications. *J. Int. Neuropsychol. Soc.* 22, 138–163. doi: 10.1017/S1355617715000995

Teipel, S. J., Bokde, A. L., Born, C., Meindl, T., Reiser, M., Moller, H. J., et al. (2007). Morphological substrate of face matching in healthy ageing and mild cognitive impairment: a combined MRI-fMRI study. *Brain* 130(Pt 7), 1745–1758. doi: 10.1093/brain/awm117

Teipel, S. J., Kurth, J., Krause, B., Grothe, M. J., and Alzheimer's Disease Neuroimaging Initiative (2015). The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - beyond classical regression. *Neuroimage Clin.* 8, 583–593. doi: 10.1016/j.nicl.2015.05.006

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* 58, 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. B* 73, 273–282. doi: 10.1111/j.1467-9868.2011.00771.x

Trzepacz, P. T., Yu, P., Sun, J., Schuh, K., Case, M., Witte, M. M., et al. (2014). Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. *Neurobiol. Aging* 35, 143–151. doi: 10.1016/j.neurobiolaging.2013.06.018

Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., et al. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* 76, 183–201. doi: 10.1016/j.neuroimage.2013.03.004

Zhou, J., Greicius, M. D., Gennatas, E. D., Growdon, M. E., Jang, J. Y., Rabinovici, G. D., et al. (2010). Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. *Brain* 133(Pt 5), 1352–1367. doi: 10.1093/brain/awq075

Zhuang, J., Tyler, L. K., Randall, B., Stamatakis, E. A., and Marslen-Wilson, W. D. (2014). Optimally efficient neural systems for processing spoken language. *Cereb. Cortex* 24, 908–918. doi: 10.1093/cercor/bhs366

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zou, H., and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* 37, 1733–1751. doi: 10.1214/08-AOS625

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix: Penalized regression

The following description except the elastic net penalty follows the description of penalized regression in James et al. (1, Chapter 6.2).

To address the curse of dimensionality (2), the parameter estimate β from the classical linear regression model

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2,$$

with the number of cases n and number of features p, has been constrained by various penalty terms with different ensuing characteristics of the estimates of the β coefficients.

The penalty terms decrease fit variance at the cost of increasing bias. Through the *variance-bias trade-off,* overall fit of a penalized parameter estimate can be improved over the simple least square estimation of the parameter without penalization.

In 1970, Hoerl and Kennard (3) introduced the **ridge penalty**, with an L2 quadratic penalty

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

with the tuning parameter $\lambda \geq 0$, and the shrinkage penalty $\lambda\sum_j \beta_j^2$, and the number of features p. This penalty shrinks the magnitude of all coefficients, but sets no coefficient exactly to zero. As a result, it increases prediction accuracy, but has limited model interpretability in the presence of a high number of features.

In 1996, Tibshirani (4) described the **Least Absolute Shrinkage and Selection Operator (Lasso)**, which is characterized by an L1 (absolute value) penalty,

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|,$$

with the tuning parameter $\lambda \geq 0$, the shrinkage penalty $\lambda\sum_j|\beta_j|$, and number of features p.

The Lasso forces some of the coefficients to become zero, so that it provides both shrinkage and subset selection. The fit for the Ridge regression is superior to the Lasso if many variables are truly related to the outcome, whereas the fit of the Lasso is superior to the Ridge if only few variables are truly related to the outcome.

For collinear data, which are characterized by subgroups of features with high intercorrelation, such as has to be expected in brain imaging data with hubs that are interconnected in partly overlapping networks, the Lasso has the unfavorable characteristics that it selects only one among a group of highly correlated features. In addition, if $p > n$, i.e. the number of features is larger than the number of cases, the Lasso selects at most n variables.

To overcome these limitations, in 2005, Zou and Hastie (5) introduced the **elastic net penalty** that was extended to non-linear regression, such as logistic regression, in 2010 (6). Elastic net penalty regression features both, the L2 norm (quadratic) Ridge and the L1 norm Lasso penalty that are governed by an additional parameter $\alpha\epsilon[0,1]$,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left[(1-\alpha)\frac{1}{2}\beta_j^2+\alpha\left|\beta_j\right|\right]$$
$$=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\left[(1-\alpha)\frac{1}{2}\beta_j^2+\alpha\left|\beta_j\right|\right].$$

If $\alpha=0$, the elastic net penalty becomes the Ridge penalty, while if $\alpha=1$, the elastic net penalty becomes the Lasso penalty. In addition, in simulated data, the elastic net penalty has been shown to select or discard highly correlated features as a group (5).

References

1. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York, NY: Springer New York (2013).
2. Bellman R, Bellman RE. *Adaptive Control Processes: A Guided Tour*: Princeton University Press (1961). 255 p.
3. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* (1970) **12**(1):55-67. doi: 10.1080/00401706.1970.10488634.
4. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* (1996) **58**(1):267–88.
5. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2005) **67**(2):301-20. doi: 10.1111/j.1467-9868.2005.00503.x.
6. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* (2010) **33**(1):1-22. PubMed PMID: 20808728; PubMed Central PMCID: PMC2929880.

Journal of Alzheimer's Disease 64 (2018) 801–813
DOI 10.3233/JAD-180106
IOS Press

Multicenter Resting State Functional Connectivity in Prodromal and Dementia Stages of Alzheimer's Disease

Stefan J. Teipel[a,b,*], Coraline D. Metzger[c,d,e], Frederic Brosseron[f,g], Katharina Buerger[h,i], Katharina Brueggen[b], Cihan Catak[i], Dominik Diesing[j], Laura Dobisch[e], Klaus Flieβbach[f,g], Christiana Franke[l,m], Michael T. Heneka[f,g], Ingo Kilimann[a,b], Barbara Kofler[f,k], Felix Menne[j,m], Oliver Peters[j,m], Alexandra Polcher[f], Josef Priller[l,m], Anja Schneider[f,g], Annika Spottke[f,n], Eike J. Spruth[l,m], Manuela Thelen[f,o], René J. Thyrian[p], Michael Wagner[f,g], Emrah Düzel[c,e], Frank Jessen[f,o] and Martin Dyrba[b] the DELCODE study group[1]
[a]*Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany*
[b]*German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*
[c]*Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University, Magdeburg, Germany*
[d]*Department of Psychiatry and Psychotherapy, Otto-von-Guericke University, Magdeburg, Germany*
[e]*German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany*
[f]*German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*
[g]*Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany*
[h]*German Center for Neurodegenerative Diseases (DZNE), Munich, Germany*
[i]*Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany*
[j]*Department of Psychiatry and Psychotherapy, Campus Benjamin Franklin, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[k]*Department of Psychiatry and Psychotherapy, University of Bonn, Bonn, Germany*
[l]*Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[m]*German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany*
[n]*Department of Neurology, University of Bonn, Bonn, Germany*
[o]*Department of Psychiatry, University of Cologne, Cologne, Germany*
[p]*German Center for Neurodegenerative Diseases (DZNE), Greifswald, Germany*

Accepted 9 May 2018

Abstract.
Background: Alterations of intrinsic networks from resting state fMRI (rs-fMRI) have been suggested as functional biomarkers of Alzheimer's disease (AD).
Objective: To determine the diagnostic accuracy of multicenter rs-fMRI for prodromal and preclinical stages of AD.

[1]The members of the DELCODE study group are listed in the acknowledgments.

*Correspondence to: Stefan J. Teipel, MD, Department of Psychosomatic Medicine, University Medicine Rostock, and DZNE, Gehlsheimer Str. 20, 18147 Rostock, Germany. Tel.: +49 01149 381 494 9470; Fax: +49 01149 381 494 9682; E-mail: stefan.teipel@med.uni-rostock.de.

ISSN 1387-2877/18/$35.00

Methods: We determined rs-fMRI functional connectivity based on Pearson's correlation coefficients and amplitude of low-frequency fluctuation in people with subjective cognitive decline, people with mild cognitive impairment, and people with AD dementia compared with healthy controls. We used data of 247 participants of the prospective DELCODE study, a longitudinal multicenter observational study, imposing a unified fMRI acquisition protocol across sites. We determined cross-validated discrimination accuracy based on penalized logistic regression to account for multicollinearity of predictors.
Results: Resting state functional connectivity reached significant cross-validated group discrimination only for the comparison of AD dementia cases with healthy controls, but not for the other diagnostic groups. AD dementia cases showed alterations in a large range of intrinsic resting state networks, including the default mode and salience networks, but also executive and language networks. When groups were stratified according to their CSF amyloid status that was available in a subset of cases, diagnostic accuracy was increased for amyloid positive mild cognitive impairment cases compared with amyloid negative controls, but still inferior to the accuracy of hippocampus volume.
Conclusion: Even when following a strictly harmonized data acquisition protocol and rigorous scan quality control, widely used connectivity measures of multicenter rs-fMRI do not reach levels of diagnostic accuracy sufficient for a useful biomarker in prodromal stages of AD.

Keywords: Aging, diagnostic accuracy, functional MRI, multicenter, subjective cognitive decline

INTRODUCTION

Today, early diagnosis of Alzheimer's disease (AD) in prodromal and dementia stages is supported by the detection of disease characteristic biomarkers, such as amyloid accumulation or tau pathology [1–3]. In addition, people in prodromal or clinical at risk stages of AD such as people with mild cognitive impairment (MCI) [4] or subjective cognitive decline (SCD) [5] show characteristic pattern of metabolic decline in cortical areas belonging to cortical functional networks, particularly the default mode network [6], in ^{18}FDG-PET studies [7–10]. In comparison to ^{18}FDG-PET, functional MRI during resting state (rs-fMRI) is more widely available, less costly and has no radiation exposure. A large range of studies has shown that seed based as well as data driven analyses can reveal distinct intrinsic connectivity networks, including the default network, based on low frequency resting state fluctuations of the blood oxygenation level dependent signal (BOLD) (for review, see [11]). The BOLD signal is considered a measure of neuronal activity, similar to FDG-PET metabolism, and intrinsic connectivity networks from rs-fMRI have been suggested to carry diagnostically useful information in MCI and AD dementia [12]. However, the coupling between neuronal metabolism and activity and the oxygen extraction fraction, the physiological driver of the BIOLD signal, is complex, varying with blood flow, cerebral pathology, and activity [13]. Thus, findings on the accuracy of group discrimination based on rs-fMRI between prodromal and dementia stages of AD and healthy controls vary even between monocenter studies [14–17].

A step further in the evaluation of rs-fMRI as a diagnostic tool is the assessment of the multicenter stability of this measure across different scanners. In healthy people the pattern of functional connectivity was found to be affected by different scanners [18–21]. Studies on the effect of multi scanner variability on the diagnostic value of rs-fMRI in AD are still scarce. In a previous study using retrospectively collected multicenter rs-fMRI data [22] we found a medium sized effect of scanner on between group differences. The discriminatory accuracy between AD dementia and MCI patients and healthy controls after cross validation was about 80 to 70%, respectively, rendering multicenter rs-fMRI inferior in diagnostic accuracy compared with hippocampus volume [22], a well-established structural MRI marker of AD [23].

We expected that the retrospective collection of fMRI data without controlling for the acquisition parameters increased multicenter variability and decreased group discrimination compared with harmonized rs-fMRI acquisitions across scanners. Therefore, in the present study we determined the diagnostic accuracy of two frequently employed rs-fMRI indices, functional connectivity [16] and amplitude of low-frequency fluctuation (ALFF) [24], from a prospective cohort study with a harmonized acquisition protocol and prospectively employed scan quality control. Extending previous studies, we included not only people with AD dementia, MCI or healthy controls, but also a group of people with SCD which represent a preMCI at risk stage of AD [5]. We hypothesized that multicenter variability of functional connectivity and ALFF would be

lower compared with previous multicenter studies using harmonized retrospective data collection. In addition, we hypothesized that the cross-validated diagnostic accuracy would be higher in the prospective cohort than in previous multicenter studies which were based on retrospective data. To address the issue of multicollinearity of functional imaging features, we employed the technique of elastic net regression [25] which was found superior for feature selection in several previous studies, including fMRI, FDG-PET, structural MRI and amyloid PET data of people with AD and MCI, compared with classical stepwise selection [26–29].

MATERIALS AND METHODS

Subjects

For the current study we used data from the DELCODE study, an ongoing observational longitudinal memory clinic-based multicenter study in Germany [30]. The sample included 27 patients with a clinical diagnosis of probable AD dementia according to the National Institute on Aging-Alzheimer's Association (NIA-AA) workgroups guidelines [2], 50 patients meeting the core clinical criteria for MCI according to NIA-AA workgroups guidelines [1], 90 people with a diagnosis of SCD, and 80 cognitively healthy older controls. Participants with SCD were cognitively unimpaired and stated to have decline in cognitive functioning unrelated to an event or condition explaining the cognitive deficits according to research criteria [5]. Healthy controls never reported SCD and had no history of neurological or psychiatric disease or any sign of cognitive decline. All participants were tested with an extensive cognitive battery [30] including the Consortium to Establish a Registry of Alzheimer's Disease (CERAD) cognitive battery [31], the Mini-Mental Status Examination (MMSE) [32], the clinical dementia rating scale (CDR) [33], the immediate and delayed story recall logical memory subtest of the revised Wechsler Memory Scale [34], and the Geriatric Depression Scale (GDS) [35]. Participants were excluded from the study if they fulfilled one of the following exclusion criteria: current major depressive episode; major psychiatric disorders (e.g., psychotic disorder, bipolar disorder, or substance abuse); neurodegenerative disorder other than AD; vascular dementia; or history of stroke with residual clinical symptoms. Further information is given in [30]. A detailed overview about relevant comorbidities and medication is provided in the Supplementary Material.

Groups were matched in respect to age and sex distribution. These data originated from an interim data set of the first 400 cases of the DELCODE study at baseline. From these 400 cases only 320 had rs-fMRI data available. From these 320 scans we excluded 9 cases due to neurologic conditions, 20 cases due to image quality issues, and 44 cases due to unbalanced age and sex distribution, leaving 247 scans originating from six sites of the DELCODE study. The demographic characteristic of the participants of the current analysis can be found in Table 1.

Written informed consent was provided by all participants, or their representatives. The study was approved by local ethics committees at each of the participating centers and has been conducted in accord with the Helsinki Declaration of 1975.

Of the 80 controls, 28 had $A\beta_{42}$ and $A\beta_{42}/A\beta_{40}$ ratio measures available from CSF, with a normal $A\beta_{42}/A\beta_{40}$ ratio in 25 cases; of the 90 people with SCD 36 had CSF markers available with a normal $A\beta_{42}/A\beta_{40}$ ratio in 29 cases; of the 50 people with MCI 27 had CSF markers available, with a normal $A\beta_{42}/A\beta_{40}$ ratio in 14 cases; in AD dementia CSF markers were available in 14 cases with a normal $A\beta_{42}/A\beta_{40}$ ratio in one case. Lumbar puncture followed the MRI scanning within on average 19 days (ranging from 0 to 78 days).

Imaging data acquisition

Data were obtained from six Siemens 3.0 Tesla MRI scanners (4 Verio, one Skyra, one TimTrio) with unified scanning protocols and instructions. In all centers, the participants were instructed to keep their eyes closed, relax, but not to fall asleep. Initially, the field-of-view (FOV) was orientated to be in plane with the anterior–posterior commissure line covering the whole brain. Functional MRI was based on a T_2*-weighted echo-planar imaging sequence using a 64×64 image matrix with 47 axial slices (thickness 3.5 mm, no gap) and interleaved acquisition. The FOV was $224 \times 224 \times 165$ mm, isotropic voxel size of 3.5 mm, echo time 30 ms, repetition time 2,580 ms, flip angle 80°, and parallel imaging acceleration factor 2. The sequence took 7 min 54 s. High-resolution T_1-weighted anatomical scans were obtained from all participants using the magnetization-prepared rapid gradient echo (MPRAGE) sequence during the same session. Image matrix was 256×256 with 192 sagittal slices, FOV $250 \times 250 \times 192$ mm, isotropic voxel

Table 1
Demographic characteristics

	AD	MCI	SCD	Controls
No. cases (women)[1]	27 (17)	50 (19)	90 (49)	80 (42)
Age (SD) [y][2]	72.7 (7.0)	72.5 (4.8)	72.0 (5.2)	71.0 (4.2)
MMSE (SD)[3]	23.7 (3.3)	27.9 (1.6)	29.1 (1.0)	29.3 (0.9)
GDS[4]	2.0 (1.6)	1.8 (1.8)	1.8 (1.8)	0.8 (1.6)
Education (SD)[5]	13.7 (3.1)	14.0 (3.1)	14.7 (3.2)	14.7 (2.8)

MMSE, Mini Mental State Examination [32]; GDS, Geriatric Depression Scale [35]. [1]not significantly different between groups, $\chi^2 = 5.4$, 3 df, $p = 0.14$. [2]not significantly different between groups, $F(3, 243) = 7.5$, $p = 0.25$. [3]significantly different between groups, Kruskal Wallis Test, $p < 0.001$. [4]significantly different between groups, $F(3, 243) = 6.0$, $p < 0.001$. [5]not significantly different between groups, $F(3, 243) = 1.3$, $p = 0.26$.

size of 1 mm, echo time 4.37 ms, repetition time 2,500 ms, flip angle 7°, and parallel imaging acceleration factor 2. The duration of the sequence was 5 min 8 s.

Biomaterial sampling

Biomaterial sampling included CSF in those participants, who consented. Trained study assistants performed the collection, processing and storage of the samples up to the shipment to the central biorepository of the DZNE according to SOP. After the centrifugation CSF was aliquoted and stored at –80°C.

MR processing

The **anatomical T_1-weighted image** for each participant was segmented into gray matter, white matter, and cerebrospinal fluid (CSF) partitions partitions using the New Segment routines included in Statistical Parametric Mapping (SPM12) [36]. The Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm [37] was applied to normalize the T_1-weighted gray matter and white matter partitions to the Montreal Neurological Institute (MNI) reference coordinate system using the default brain template included in CAT12 [38]. Individual flow-fields resulting from the DARTEL registration to the reference template were used to warp the gray matter segments and to apply modulation to preserve the total amount of grey matter in the scans.

Functional MRI data processing was carried out using Data Processing Assistant for Resting-State fMRI (DPARSF 4.3) [39], considering the recommendations from a recent systematic evaluation of processing alternatives [40]. After the removal of the first ten images to account for gradient field stabilization, the rs-fMRI data was slice time corrected and realigned to the temporal mean image. The anatomical T_1-weighted image for each participant was coregistered to the mean functional image. The deformation fields generated by DARTEL from the anatomical T_1-weighted images were used to project the functional scans from each subjects' native image space into the MNI reference space. We combined this step with the reslicing of all functional data to an isotropic resolution of 3 mm. The subsequent nuisance regression included covariates of head movement (rotation, translation, and derivatives) and the mean time courses for the global brain signal, the white matter segment signal, and the CSF segment signal. Although global signal regression was found to introduce negative correlations [40, 41], studies consistently reported that it effectively improves the signal-to-noise ratio [40, 42, 43]. Recently, Shirer et al. evaluated the influence of global signal regression on group separation but only found a minor, non-significant effect [40]. Subsequently, the images were band-pass filtered using the frequency band 0.1–0.01 Hz. For each individual the time series of signal was extracted for each of the 90 functionally defined regions of the Greicius atlas [44]. Functional connectivity was defined as the Pearson's correlation coefficients between all 4005 possible pairs of correlations between these 90 regions [44]; this means that we included both connectivity *within* as well as *between* intrinsic connectivity networks. Additionally, to test the sensitivity of the results to this specific atlas, we repeated functional connectivity analyses for two widely used alternative functional atlases, the Craddrock atlas [45] containing 200 regions and the Schaefer-Yeo atlas [46] containing 100 regions. Finally, Pearson correlation coefficients were adjusted to be normally distributed using Fisher's Z-transform [47]: $z = 0.5 \ln [(1+r)/(1-r)]$. ALFF maps were calculated for the frequency band 0.1–0.01 Hz based on the slice time corrected and realigned data in native space, subsequently pro-

jected to the MNI reference space, and resliced to an isotropic resolution of 3 mm. Mean ALFF values were then obtained for each region of the Greicius atlas [44].

Extraction of hippocampus volumes

A mask for the hippocampus was obtained by manual delineation of the hippocampus in the reference template following the harmonized protocol for hippocampus segmentation [48, 49]. Individual gray matter volumes of the hippocampus were extracted automatically from the warped gray matter segments by summing up the modulated gray matter voxel values within hippocampus ROI in the reference space and proportional scaling to total intracranial volume (TIV) to adjust for head size.

CSF AD biomarker assessment

CSF $A\beta_{42}$ and $A\beta_{40}$ levels were determined using commercially available kits according to vendor specifications (V-PLEX Aβ Peptide Panel 1 (6E10) Kit). Cut-offs for normal and abnormal concentrations of $A\beta_{42}$ (<496 pg/ml), and of the ratio $A\beta_{42}/A\beta_{40}$ (<0.09) were derived from the literature, which applied the respective assays [50].

Statistical analysis

Demographics

We compared demographic characteristic between groups using parametric and non-parametric tests as required: Age and years of education were compared between groups using Student's *t*-test, sex distribution using Chi-square test, MMSE and GDS scores using the Mann-Whitney-U- test.

Data quality

We determined the variability of the temporal signal to noise ratio (tSNR) across centers to assess the degree of between center variability. Additionally, we assessed the standardized DVARS, a measure of how much the intensity of a brain image changes in comparison to the previous time point [42]. Standardized DVARS is scaled by the temporal standard deviation and temporal autocorrelation so that it is approximately 1 if there are no artifacts in the functional data. This measure detects scan artifacts that are not necessarily related to head motion [42]. We had found both measures, tSNR and DVARS, highly sensitive to indicate insufficient image quality in a previous multicenter rs-fMRI study [22].

Diagnostic accuracy of resting state fMRI features

We determined group discrimination based on a penalized logistic regression model with an elastic net penalty, using the R package *glmnet* (available at http://cran.r-project.org/web/packages/glmnet/index.html). Elastic net regression extends the traditional linear regression model, which minimizes the residual sum of squares, such that two penalty terms are added [25]: $\hat{\beta} = \text{argmin}_{\beta}\,[\|y - X\beta\| + \lambda((1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1)]$. Elastic net regression is governed by two parameters. The parameter $\alpha \in [0, 1]$, which determines the type of regularization, which lies between the extreme of Ridge regression, (regularization by squared L_2 norm $\|\beta\|^2$ if $\alpha = 0$) and the Lasso (Least Absolute Shrinkage and Selection Operator, regularization by L_1 norm $\|\beta\|_1$ if $\alpha = 1$), and (ii) $\lambda > 0$, defining the strength of regularization [51]. While the L_1 norm $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ forces some of the β_j coefficients to become zero, it was reported to keep only one of a group of inter-correlated predictors in the model and to discards the others, thus leading to unstable models [25]. In contrast, the L_2 norm $\|\beta\|^2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ keeps groups of inter-correlated predictors in the model, but will not discard useless predictors [25]. We preselected α to be 0.5 which corresponds to a full elastic net penalty with an equal weighting of the L_1 and L_2 norm regularization as we previously observed a tendency for overfitting with low α (not dropping predictors) when determining the optimal value for α based on cross-validation [27]. The parameter λ was determined using grid search with 100 times repeated cross-validation. The optimal λ was determined as the mean across 100 iteratively determined λ values minimizing the deviance of the model. We followed a strict cross-validation procedure. We draw random samples of 2/3 of the data 1,000 times to train the prediction models (training data) including the λ parameter. The prediction accuracy than was determined using the remaining 1/3 of the data as the test set. Prior to model building, the feature space was restricted through determining the set of variables which correlated with diagnosis with a Pearson's correlation coefficient of an absolute value at 97.5th percentile in the training data, resulting in an average number of about 100 predictor variables across the 1,000 cross-validation repetitions. In a sensitivity analysis, we

forced a dummy coded center variable in to the model to determine to what extent center would account for group differences. To compare diagnostic accuracy of functional connectivity and ALFF with the accuracy of an established measure, we used hippocampus volume as a reference test.

We used a support vector machine (SVM) as alternative classifier to assess if the outcome was sensitive to the employed classifier. Details of the SVM classifier are described in the Supplementary Material.

RESULTS

Demographics

Groups were matched in respect to age and sex, so that we did not find significant group differences in these parameters; in addition, groups were not significantly different in education (Table 1). As expected, healthy controls and people with SCD scored higher in the MMSE than the MCI and AD dementia cases, with the lowest performance in the AD dementia cases. GDS scores differed significantly between groups, i.e., healthy controls had significantly lower values than SCD, MCI and AD dementia cases; however, all participants scored within a clinically normal range.

Between-center variability

The temporal signal to noise ratio and the standardized DVARS across the six sites included in the analysis are shown in Fig. 1. Both indices showed wide overlap between centers and less variation in comparison to the corresponding values from the retrospectively collected psymri data (shown as boxplot on the left for both indices), suggesting a low between center variability in these key measures of scan quality.

Group discrimination

As shown in Fig. 2, the cross-validated discrimination accuracy for ALFF and functional connectivity ranged between 56% and 81% in the people with AD dementia compared to healthy controls, was 57% in the MCI cases and ranged between 56% and 57% in the people with SCD. Using functional connectivity measures based on other atlases than the Greicius atlas [44] yielded essentially identical results with on average slightly lower AUC for the Craddock [45] (–3%) and Schaefer-Yeo [46] (–4%) atlases (Supplementary Figure 1). Notably, the Schaefer-Yeo atlas [46] performed substantially worse (–15% AUC) for amyloid-positive AD dementia cases versus amyloid-negative controls, which might result from the missing subcortical regions in that specific atlas. With exception of the AD dementia cases, the bootstrapped 95% confidence intervals (CI) of the cross validated accuracy levels included the value of 50%, suggesting that the degree of diagnostic accuracy did not significantly exceed random guessing accuracy in all groups except the AD dementia cases. For comparison, the diagnostic accuracy based on the hippocampus volume was only 54% in people with SCD, but reached 75% in people with MCI and 88% in people with AD dementia. When we repeated the analysis with a dummy coded center covariate forced into the model, the diagnostic accuracy remained essentially unchanged.

When we assessed the diagnostic accuracy in the MCI subgroup of 13 cases with an abnormal $A\beta_{42}/A\beta_{40}$ ratio in CSF and the 25 healthy controls with a normal $A\beta_{42}/A\beta_{40}$ ratio, we found a cross-validated AUC for ALFF and functional connectivity between 69% and 73%, with the lower level of the 95% CI including 50%. For comparison in this subsample, hippocampus volume reached a cross-validated AUIC of 78%, where the 95% CI included 50% as well.

When we determined the 50% most frequently selected functional connectivity networks using elastic net regression for the AD dementia cases versus controls comparison, these networks involved functional connectivity between regions from the dorsal DMN, the anterior and posterior salience network, the language network, the left executive network, the visuo-spatial and the sensori-motor network (Table 2 and Fig. 3). Since the other comparisons did not significantly exceed random guessing accuracy, we did not assess the regional distribution of discriminatory features for these comparisons.

When we used a support vector machine classifier, results were similar to the elastic net classification (see Supplementary Figure 2).

DISCUSSION

Compared with previous multicenter studies using retrospective data acquisition, in this prospective multicenter rs-fMRI analysis we found less pronounced variations of global scan quality parameters across sites, including measures of temporal

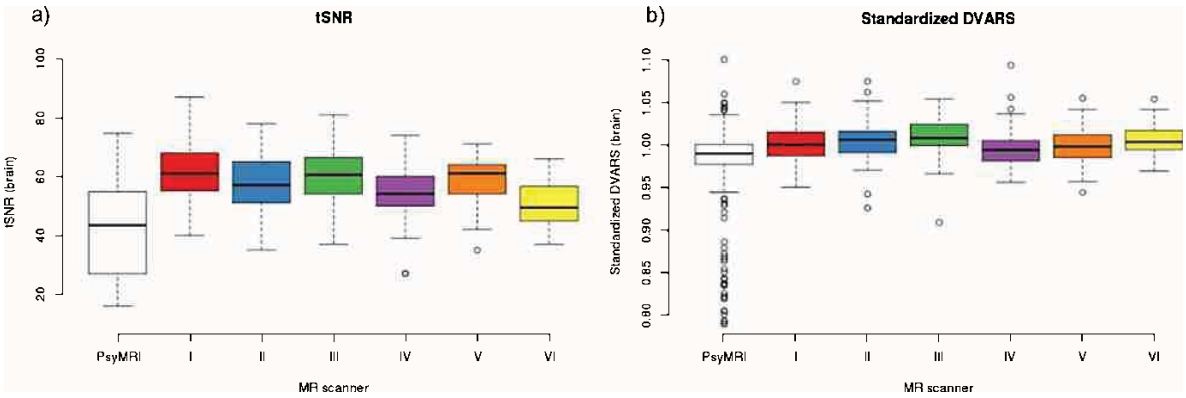


Fig. 1. Multicenter data quality indices. a) Mean whole brain temporal Signal to Noise Ratio (tSNR). b) Standardized DVARS. Box plots for the quality indices with open circles indicating outlying values for the six centers (denoted center I to center VI). For comparison, the boxplots to the left indicate the corresponding values from the retrospectively collected fMRI data of the psymri cohort.

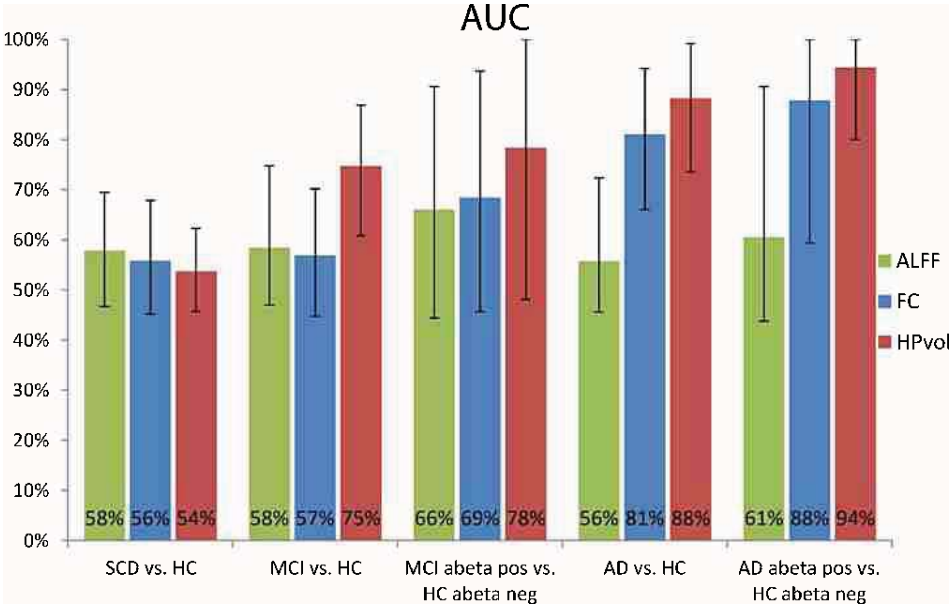


Fig. 2. Cross-validated areas under the ROC curves. Bar diagrams for the group discrimination accuracy as measured by cross-validated AUC with 95% confidence intervals. SCD, subjective cognitive decline; HC, healthy controls; MCI, mild cognitive impairment; AD, Alzheimer's disease; ALFF, amplitude of low-frequency fluctuation; FC, functional connectivity; HPvol, hippocampal volume.

signal to noise ratio that we had found to be particularly sensitive to between scanner variability in a previous study [22]. Despite the smaller inter-scanner variability, between group effects of resting state functional connectivity and ALFF reached only small effect size in prodromal and at risk stages of AD, such as SCD and MCI. Only in manifest stages of AD dementia, we found 80% accuracy of group discrimination, where the cross-validated level of accuracy was significantly above chance level. This level of accuracy is similar to the level of between group discrimination in previous mono-center studies comparing AD dementia and healthy controls using cross validated diagnostic accuracy. In contrast, the discrimination between MCI patients and controls and even between MCI patients with pronounced episodic memory impairment and controls did not significantly differ from chance level. The effect was even smaller for the people with SCD. These results were not sensitive to the atlas selection and were replicated using support vector machine as alternative classifier.

The level of accuracy found in our amnestic MCI cases is lower than the levels of accuracy found in

Table 2
Functional connectivity on regional and network level discriminating AD patients from controls

Beta	network 1	network 2	Corresponding AAL region 1	Corresponding AAL region 2	mean FC controls	mean FC AD
–1.90	dDMN	sensorimotor	l. hippocampus + parahippocampal gyrus	l. + r. cerebellum	0.01	–0.19
–1.25	dDMN	post. salience	r. angular gyrus	l. insula	0.23	0.04
–1.01	language	precuneus	l. middle temporal gyrus	l. inferior parietal lobule + angular gyrus	0.35	0.21
–0.89	dDMN	dDMN	l. + r. thalamus	l. hippocampus + parahippocampal gyrus	0.50	0.32
–0.85	language	post. salience	r. middle and sup. temporal gyrus	r. mid. cingulum	0.30	0.11
–0.78	dDMN	dDMN	l. + r. superior medial frontal gyrus + ant. cingulum	l. angular gyrus	0.22	0.10
–0.47	dDMN	precuneus	l. + r. thalamus	l. post. cingulum	0.21	0.03
0.15	ant. salience	dDMN	l. insula	Angular R	–0.25	–0.06
0.38	dDMN	visuospatial	l. + r. thalamus	r. angular, supramarginal, and postcentral gyrus, and parietal inferior lobule	0.08	0.23
0.52	language	LECN	l. inf. orbitofrontal	L superior and inferior frontal gyrus	–0.17	–0.03
0.52	dDMN	sensorimotor	l. + r. superior medial frontal gyrus + ant. cingulum	l. thalamus	–0.14	0.01
0.55	dDMN	post. salience	r. angular gyrus	l. precuneus	–0.18	–0.03
0.83	auditory	sensorimotor	r. thalamus	r. thalamus	–0.17	–0.04
0.92	dDMN	post. salience	r. angular gyrus	r. mid. cingulum	–0.28	–0.14
1.00	dDMN	visuospatial	l.+ r. precuneus + post. cingulum	r. cerebellum	–0.17	–0.04
1.01	ant. salience	visuospatial	r. middle frontal gyrus	r. angular, supramarginal, and postcentral gyrus, and parietal inferior lobule	–0.05	0.08
1.25	dDMN	RECN	l. + r. thalamus	r. middle frontal gyrus	0.33	0.45
1.90	dDMN	post. salience	l. + r. thalamus	r. middle frontal gyrus	–0.09	0.04
2.40	dDMN	post. salience	l. + r. superior medial frontal gyrus + ant. cingulum	l. thalamus	–0.10	0.02

r., right; l., left; ant., anterior; post., posterior; FC, functional connectivity; dDMN, dorsal default mode network; RECN, right executive network; LECN, left executive network.

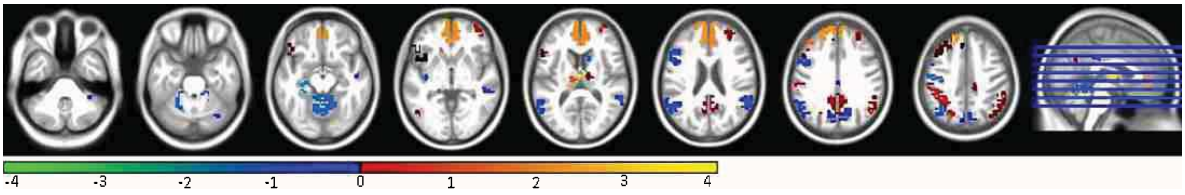


Fig. 3. Resting state network components contributing to the discrimination between AD patients and controls. Resting state network components whose functional connectivity was contributing to the discrimination between AD patients and controls projected on a structural MRI scan in MNI space. Axial sections go through MNI z coordinates –38, –26, –14, –2, 10, 22, 34, and 46 mm. Colors code the sum of the beta weights for each region.

previous mono-center studies when these did not consider the cross-validated accuracy [14, 15], and two studies using cross-validation [16, 17]. One recent study based on multicenter fMRI data from ADNI, found a level of 77% cross-validated accuracy for the discrimination between 54 MCI cases and 54 controls based on functional connectivity, without explicitly addressing the multicenter characteristics of the data [52]. Our findings agree with a previous multicenter study which was based on retrospective collected fMRI data sets, where we found cross validated accuracy of group discrimination between MCI patients and controls between 66% and 72% [22]. Together these data suggest that despite a reduction of multicenter variability through a prospectively harmonized acquisition protocol the usefulness of

resting state functional connectivity for the discrimination between amnestic MCI cases and healthy controls is very limited. The level of accuracy was higher when comparing the subset of MCI cases who were amyloid positive with the controls who were amyloid negative. The level of accuracy here ranged between 66% and 69%; with the large confidence interval likely reflecting the small number of cases. One previous fMRI study showed steeper age related decline in DMN, posterior cingulate and frontoparietal resting state network connectivity in CSF amyloid positive versus amyloid negative cognitively healthy older people [53]. Consistently, previous rs-fMRI studies showed that resting state connectivity measures were related to CSF levels of amyloid across diagnostic groups, including cognitively healthy older people, people with MCI and AD dementia patients [54–58]. Taking these previous findings into account, our finding would indicate that rs-fMRI functional connectivity more closely reflects the underlying molecular pathology than the clinical phenotype of amnestic MCI; this is different to hippocampus volume that reached similar levels of accuracy both in amnestic MCI cases and the amyloid positive subgroup, suggesting that hippocampus volume reflects the amnestic phenotype of MCI, relatively independent from the underlying molecular pathology.

The number of previous monocenter studies in people with SCD is limited. One study used independent component analysis to decompose the rs-fMRI data of 13 healthy controls, 8 people with Alzheimer's disease dementia, 21 with MCI and 16 with SCD into distinct functional connectivity networks and assessed the association of these networks with neuropsychological performance in different cognitive tests. Group differences in functional connectivity were not reported [59]. To our knowledge, the current study is the first multicenter study to report diagnostic accuracy of functional connectivity from rs-fMRI to discriminate people with SCD from healthy controls. Interestingly, two previous monocenter studies reported not only decreases, but also significant increases of functional connectivity measures in SCD people compared to controls, one study comprising 25 SCD cases and 60 controls [60] and the other study comprising 25 people with SCD and 29 controls [61].

The lack of a significant discrimination between SCD cases and controls in our study may be related to two factors: First, in the subset of SCD cases where CSF was available, the proportion of amyloid positive cases was low with only 19%, albeit comparable with previous studies in subjective memory complainers with a rate of 21% amyloid positive cases [62]. If we interpolate this proportion to the entire group of SCD cases this would indicate that the majority of the SCD cases was not in a preclinical state of AD so that the lack of discrimination may simply reflect the lack of underlying neuropathology. The few previous studies on resting state fMRI in SCD [59–61] did not control for amyloid status. Unfortunately, with only 7 amyloid positive SCD cases we were not able to determine a meaningful level of accuracy. Secondly, following the previous observation that SCD cases showed not only decreased, but also increased connectivity [60, 61], people with SCD may present with a mixture of increases and decreases of functional connectivity that together may lead to a poor diagnostic accuracy when searching for between group differences across the entire brain.

In addition, relevant for all comparisons multicenter acquisition may additionally reduce between group effects through sources of inter-scanner variance that remain even after prospective harmonization of acquisition protocols and strict image quality control. The negative finding in the SCD group is based on a relatively large sample of 90 cases compared to 80 controls with a strict matching in respect to age, sex distribution and MMSE score performance. These are the first data on the discriminatory power of rs-fMRI in SCD, however, due to the relatively high number of cases, the risk of a false negative finding is small.

When we assessed the networks and regions contributing to significant group discrimination in AD dementia, we found not only a strong involvement of default mode network (DMN) regions, including not only hippocampal and neocortical but also thalamic components of this network, but also language, executive, salience and visuo-spatial networks. This agrees with previous studies showing reduced within network connectivity in AD patients compared with controls in the DMN, ventral and dorsal salience network, executive network and a frontal-parietal network involving the precuneus [63] as well as altered between network connectivity involving the dDMN, executive, visuospatial, fronto-parietal and anterior and posterior salience networks [54, 64, 65], including the thalamus. Indeed, we found a reduced strength of positive associations within the DMN or the DMN and the precuneus resting state network in AD patients compared with controls, and a reduced strength of negative associations between

the DMN and the anterior salience network. Different to previous studies [54, 64, 65], however, we did not *a priori* exclude resting state networks from the analyses. Using such a purely data driven approach, we found alterations also including networks that would be considered to be relatively spared by AD, such as the sensorimotor or auditory network. This does not imply decline of connectivity within these latter networks, but rather a change in coupling between relatively spared and more affected networks. So, for example, the negative association between sensorimotor and auditory network components became less strong in AD patients compared with controls or sensorimotor network and DMN components became more negatively associated in AD, suggesting a potential dysfunctional compensatory effect in people with AD.

A strength of our study is the large number of people with subjective cognitive decline. To our knowledge only the INSIGHT-preAD cohort [62] with 318 participants has a larger number of people with subjective cognitive decline and rs-fMRI, however, the INSIGHT-preAD cohort does not have a control group so that the analysis of the discriminatory accuracy of rs-fMRI has to await the longitudinal follow-up. A limitation of our study is that the number of available CSF samples was too small in the SCD group to allow for a meaningful comparison of the amyloid positive SCD cases with the amyloid negative controls. The DELCODE study is ongoing so that we will have access to a larger number of amyloid positive SCD cases to study the association of rs-fMRI with the presence or absence of biomarker-confirmed preclinical at risk stages of AD in the future.

A strength of the study in comparison to previous retrospective multicenter rs-fMRI studies is the strict harmonization of acquisition protocols as well as the strict implementation of scan quality controls throughout the study. Indeed, compared to our previous study based on retrospective data [22], key parameters of scan quality showed much less variability in the prospectively acquired data. Still, although the degree of variability between scanners was lower compared to the previous multicenter study, the achievable level of diagnostic accuracy was comparable to the previous study, suggesting that between scanner variability is not the main source of limited diagnostic accuracy from multicenter rs-fMRI data. Future follow-up data of this cohort will allow us to determine predictive accuracy of resting state functional connectivity for a clinically relevant functional outcome such as decline of cognitive function or conversion to dementia.

The prospective design of our study with a strict harmonization of acquisition protocols is a strength of the study when we interpret the findings in respect to future application of the technique in controlled multicenter trials. Since the outcome of our prospective study suggests a low diagnostic accuracy of rs-fMRI functional connectivity in a highly controlled setting, we expect the level of accuracy would even be lower in a less controlled routine care setting.

In summary, in a prospective multicenter resting state fMRI acquisition we found significant group discrimination between AD dementia patients and controls that was, however, inferior to the widely established measure of hippocampus volume. Functional connectivity or ALFF did not reach significant group discrimination above chance level neither in SCD nor in amnestic MCI cases. These findings suggest that measures of functional connectivity based on Pearson's correlation or ALFF, independently of the selected atlas and the classifier, are not useful markers in prodromal stages of AD. Future multicenter studies should explore the diagnostic usefulness of alternative functional connectivity measures, such as partial correlation [66], regional homogeneity [67] or functional dynamics based on shorter time windows of the resting state scan [66] that are currently less widely used than ALFF and Pearson's correlation metrics.

ACKNOWLEDGMENTS

SJT received support by a grant of the Federal Ministry of Research (BMBF) (AgeGain, 1GQ1425B).

DELCODE study group: J. Acosta-Cabronero, S. Altenstein, H. Amthauer, I. Apostolova, M. Barkhoff, D. Berron, M. Betts, M. Beuth, D. Bittner, F. Brosseron, K. Brüggen, K. Bürger, A. Cardenas-Blanco, C. Catak, Y. Cheng, L. Coloma Andrews, M. Dichgans, D. Diesing, L. Dobisch, A. Dörr, E. Düzel, M. Dyrba, M. Ehrlich, B. Ertl-Wagner, J. Faber, K. Fließbach, D. Frimmer, I. Frommann, M. Fuentes, W. Glanz, D. Grieger-Klose, D. Hartmann, D. Hauser, Ch. Heine, G. Hennes, G. Herrmann, B. Huber, A. Hufen, H. Janecek-Meyer, D. Janowitz, F. Jessen, Ch. Kainz, P. Kalbhen, J. Kalzendorf, E. Kasper, I. Kilimann, X. Kobeleva, B. Kofler, Ch. Korp, M. Kreißl, M. Kreuzer, A. Langenfurth, E. Lau, C. Lindlar, K. Lindner, A. Lohse, E. Markov, H. Megges, F. Menne, C. Metzger, E. Meyer, L. Miebach, K. Möhring, A. Müller,

C. Müller, P. Nestor, K. Neumann, O. Peters, H. Pfaff, A. Polcher, J. Priller, H. Raum, A. Rominger, S. Röske, Ch. Ruß, P. Sabik, P. Sänger, J. Schmid, M. Schmidt, A. Schneider, Ch. Schneider, H. Schulz, F. Schulze, P. Schulze, H. Schütze, S. Schwarzenboeck, A. Seegerer, O. Speck, A. Spottke, E. Spruth, J. Stephan, A. Szagarus, S. Teipel, C. Tempelmann, F. van der Ven, I. Villar Munoz, I. Vogt, M. Wagner, M.-A. Weber, S. Weschke, Ch. Westerteicher, C. Widmann, I. Wienhöft, S. Wolfsgruber, R. Yakupov, S. Yilmaz, G. Ziegler, A. Zollver.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/18-0106r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-180106.

REFERENCES

[1] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) 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* **7**, 270-279.

[2] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack Jr CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia 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* **7**, 263-269.

[3] Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R, Cappa S, Crutch S, Engelborghs S, Frisoni GB, Fox NC, Galasko D, Habert MO, Jicha GA, Nordberg A, Pasquier F, Rabinovici G, Robert P, Rowe C, Salloway S, Sarazin M, Epelbaum S, de Souza LC, Vellas B, Visser PJ, Schneider L, Stern Y, Scheltens P, Cummings JL (2014) Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurol* **13**, 614-629.

[4] Petersen RC (2004) Mild cognitive impairment as a diagnostic entity. *J Intern Med* **256**, 183-194.

[5] Jessen F, Amariglio RE, van Boxtel M, Breteler M, Ceccaldi M, Chetelat G, Dubois B, Dufouil C, Ellis KA, van der Flier WM, Glodzik L, van Harten AC, de Leon MJ, McHugh P, Mielke MM, Molinuevo JL, Mosconi L, Osorio RS, Perrotin A, Petersen RC, Rabin LA, Rami L, Reisberg B, Rentz DM, Sachdev PS, de la Sayette V, Saykin AJ, Scheltens P, Shulman MB, Slavin MJ, Sperling RA, Stewart R, Uspenskaya O, Vellas B, Visser PJ, Wagner M, Subjective Cognitive Decline Initiative Working Group (2014) A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement* **10**, 844-852.

[6] Liu K, Chen K, Yao L, Guo X (2017) Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the Cox model. *Front Hum Neurosci* **11**, 33.

[7] Vannini P, Hanseeuw B, Munro CE, Amariglio RE, Marshall GA, Rentz DM, Pascual-Leone A, Johnson KA, Sperling RA (2017) Hippocampal hypometabolism in older adults with memory complaints and increased amyloid burden. *Neurology* **88**, 1759-1767.

[8] Teipel S, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative (2016) Does posterior cingulate hypometabolism result from disconnection or local pathology across preclinical and clinical stages of Alzheimer's disease? *Eur J Nucl Med Mol Imaging* **43**, 526-536.

[9] Mosconi L, Mistur R, Switalski R, Tsui WH, Glodzik L, Li Y, Pirraglia E, De Santi S, Reisberg B, Wisniewski T, de Leon MJ (2009) FDG-PET changes in brain glucose metabolism from normal cognition to pathologically verified Alzheimer's disease. *Eur J Nucl Med Mol Imaging* **36**, 811-822.

[10] Mosconi L (2005) Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. *Eur J Nucl Med Mol Imaging* **32**, 486-510.

[11] Teipel S, Grothe MJ, Zhou J, Sepulcre J, Dyrba M, Sorg C, Babiloni C (2016) Measuring cortical connectivity in Alzheimer's disease as a brain neural network pathology: Toward clinical applications. *J Int Neuropsychol Soc* **22**, 138-163.

[12] Vemuri P, Jones DT, Jack Jr CR (2012) Resting state functional MRI in Alzheimer's disease. *Alzheimers Res Ther* **4**, 2.

[13] Freeman RD, Li B (2016) Neural-metabolic coupling in the central visual pathway. *Philos Trans R Soc Lond B Biol Sci* **371**, 20150357.

[14] Koch W, Teipel S, Mueller S, Benninghoff J, Wagner M, Bokde AL, Hampel H, Coates U, Reiser M, Meindl T (2012) Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiol Aging* **33**, 466-478.

[15] Balthazar ML, de Campos BM, Franco AR, Damasceno BP, Cendes F (2014) Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease. *Psychiatry Res* **221**, 37-42.

[16] Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirste T, Teipel SJ, EDSD study group (2015) Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. *J Neuroimaging* **25**, 738-747.

[17] De Marco M, Beltrachini L, Biancardi A, Frangi AF, Venneri A (2017) Machine-learning support to individual diagnosis of mild cognitive impairment using multimodal MRI and cognitive assessments. *Alzheimer Dis Assoc Disord* **31**, 278-286.

[18] Magnotta VA, Friedman L, First B (2006) Measurement of signal-to-noise and contrast-to-noise in the fBIRN Multicenter Imaging Study. *J Digit Imaging* **19**, 140-147.

[19] Lin Q, Dai Z, Xia M, Han Z, Huang R, Gong G, Liu C, Bi Y, He Y (2015) A connectivity-based test-retest dataset of

multi-modal magnetic resonance imaging in young healthy adults. *Sci Data* **2**, 150056.

[20] Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartres-Faz D, Arnold J, Benninghoff J, Fiedler U, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Sein J, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargallo N, Ferretti A, Caulo M, Aiello M, Cavaliere C, Soricelli A, Parnetti L, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Schonknecht P, Hensch T, Hoffmann KT, Kuijer JP, Visser PJ, Barkhof F, Frisoni GB, PharmaCog Consortium (2016) Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: A multicentric resting-state fMRI study. *Neuroimage* **124**, 442-454.

[21] Suckling J, Barnes A, Job D, Brennan D, Lymer K, Dazzan P, Marques TR, MacKay C, McKie S, Williams SR, Williams SC, Deakin B, Lawrie S (2012) The Neuro/PsyGRID calibration experiment: Identifying sources of variance and bias in multicenter MRI studies. *Hum Brain Mapp* **33**, 373-386.

[22] Teipel SJ, Wohlert A, Metzger C, Grimmer T, Sorg C, Ewers M, Meisenzahl E, Kloppel S, Borchardt V, Grothe MJ, Walter M, Dyrba M (2017) Multicenter stability of resting state fMRI in the detection of Alzheimer's disease and amnestic MCI. *Neuroimage Clin* **14**, 183-194.

[23] Moodley KK, Chan D (2014) The hippocampus in neurodegenerative disease. *Front Neurol Neurosci* **34**, 95-108.

[24] Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* **29**, 162-173.

[25] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc B* **67**, 301-320.

[26] de Vos F, Schouten TM, Hafkemeijer A, Dopper EG, van Swieten JC, de Rooij M, van der Grond J, Rombouts SA (2016) Combining multiple anatomical MRI measures improves Alzheimer's disease classification. *Hum Brain Mapp* **37**, 1920-1929.

[27] Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Kloppel S, Borchardt V, Walter M, Dyrba M (2017) Robust detection of impaired resting state functional connectivity networks in Alzheimer's disease using elastic net regularized regression. *Front Aging Neurosci* **8**, 318.

[28] Teipel SJ, Kurth J, Krause B, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative (2015) The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - Beyond classical regression. *Neuroimage Clin* **8**, 583-593.

[29] Trzepacz PT, Yu P, Sun J, Schuh K, Case M, Witte MM, Hochstetler H, Hake A, Alzheimer's Disease Neuroimaging Initiative (2014) Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. *Neurobiol Aging* **35**, 143-151.

[30] Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, Fliessbach K, Franke C, Fuentes M, Heneka MT, Janowitz D, Kilimann I, Laske C, Menne F, Nestor P, Peters O, Priller J, Pross V, Ramirez A, Schneider A, Speck O, Spruth EJ, Teipel S, Vukovich R, Westerteicher C, Wiltfang J, Wolfsgruber S, Wagner M, Düzel E (2018) Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimers Res Ther* **10**, 15.

[31] Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C (1989) The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **39**, 1159-1165.

[32] Folstein MF, Folstein SE, McHugh PR (1975) Mini-mental-state: A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189-198.

[33] Morris JC (1993) The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412-2414.

[34] Wechsler D (1987) *Wechsler Memory Scale-Revised*, The Psychological Corporation, New York.

[35] Yesavage JA, Sheikh JI (1986) Geriatric Depression Scale (GDS). *Clin Gerontol* **5**, 165-173.

[36] Friston KJ, Ashburner J, Kiebel S, Nichols T, Penny WD (2007) *Statistical parametric mapping: The analysis of funtional brain images*, Elsevier/Academic Press, Amsterdam, Boston.

[37] Ashburner J (2007) A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95-113.

[38] Kurth F, Gaser C, Luders E (2015) A 12-step user guide for analyzing voxel-wise gray matter asymmetries in statistical parametric mapping (SPM). *Nat Protoc* **10**, 293-304.

[39] Chao-Gan Y, Yu-Feng Z (2010) DPARSF: A MATLAB Toolbox for "pipeline" data analysis of resting-state fMRI. *Front Syst Neurosci* **4**, 13.

[40] Shirer WR, Jiang H, Price CM, Ng B, Greicius MD (2015) Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. *Neuroimage* **117**, 67-79.

[41] Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA (2009) The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *Neuroimage* **44**, 893-905.

[42] Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE (2014) Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320-341.

[43] Yan CG, Cheung B, Kelly C, Colcombe S, Craddock RC, Di Martino A, Li Q, Zuo XN, Castellanos FX, Milham MP (2013) A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* **76**, 183-201.

[44] Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD (2012) Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex* **22**, 158-165.

[45] Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS (2012) A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapp* **33**, 1914-1928.

[46] Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT (2017) Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb Cortex*. doi: 10.1093/cercor/bhx179

[47] Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507.

[48] Frisoni GB, Jack Jr CR, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, Preboske G, Swihart T, Blair M, Cavedo E, Grothe MJ, Lanfredi M, Martinez O, Nishikawa M, Portegies M, Stoub T, Ward C, Apostolova LG, Ganzola R, Wolf D, Barkhof F, Bartzokis G, DeCarli C, Csernansky JG,

deToledo-Morrell L, Geerlings MI, Kaye J, Killiany RJ, Lehericy S, Matsuda H, O'Brien J, Silbert LC, Scheltens P, Soininen H, Teipel S, Waldemar G, Fellgiebel A, Barnes J, Firbank M, Gerritsen L, Henneman W, Malykhin N, Pruessner JC, Wang L, Watson C, Wolf H, deLeon M, Pantel J, Ferrari C, Bosco P, Pasqualetti P, Duchesne S, Duvernoy H, Boccardi M, EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative (2015) The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimers Dement* **11**, 111-125.

[49] Wolf D, Bocchetta M, Preboske GM, Boccardi M, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative (2017) Reference standard space hippocampus labels according to the EADC-ADNI harmonized protocol: Utility in automated volumetry. *Alzheimers Dement* **13**, 893-902.

[50] Janelidze S, Zetterberg H, Mattsson N, Palmqvist S, Vanderstichele H, Lindberg O, van Westen D, Stomrud E, Minthon L, Blennow K, Swedish Bio Fsg, Hansson O (2016) CSF Abeta42/Abeta40 and Abeta42/Abeta38 ratios: Better diagnostic markers of Alzheimer disease. *Ann Clin Transl Neurol* **3**, 154-165.

[51] Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1-22.

[52] Chen X, Zhang H, Zhang L, Shen C, Lee SW, Shen D (2017) Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification. *Hum Brain Mapp* **38**, 5019-5034.

[53] Brier MR, Thomas JB, Snyder AZ, Wang L, Fagan AM, Benzinger T, Morris JC, Ances BM (2014) Unrecognized preclinical Alzheimer disease confounds rs-fcMRI studies of normal aging. *Neurology* **83**, 1613-1619.

[54] Weiler M, de Campos BM, Teixeira CVL, Casseb RF, Carletti-Cassani A, Vicentini JE, Magalhaes TNC, Talib LL, Forlenza OV, Balthazar MLF (2017) Intranetwork and internetwork connectivity in patients with Alzheimer disease and the association with cerebrospinal fluid biomarker levels. *J Psychiatry Neurosci* **42**, 366-377.

[55] Malpas CB, Saling MM, Velakoulis D, Desmond P, O'Brien TJ (2016) Differential functional connectivity correlates of cerebrospinal fluid biomarkers in dementia of the Alzheimer's type. *Neurodegener Dis* **16**, 147-151.

[56] Luo X, Qiu T, Jia Y, Huang P, Xu X, Yu X, Shen Z, Jiaerken Y, Guan X, Zhou J, Zhang M, Alzheimer's Disease Neuroimaging Initiative (2017) Intrinsic functional connectivity alterations in cognitively intact elderly APOE epsilon4 carriers measured by eigenvector centrality mapping are related to cognition and CSF biomarkers: A preliminary study. *Brain Imaging Behav* **11**, 1290-1301.

[57] Demirtas M, Falcon C, Tucholka A, Gispert JD, Molinuevo JL, Deco G (2017) A whole-brain computational modeling approach to explain the alterations in resting-state functional connectivity during progression of Alzheimer's disease. *Neuroimage Clin* **16**, 343-354.

[58] Binnewijzend MA, Adriaanse SM, Van der Flier WM, Teunissen CE, de Munck JC, Stam CJ, Scheltens P, van Berckel BN, Barkhof F, Wink AM (2014) Brain network alterations in Alzheimer's disease measured by eigenvector centrality in fMRI are related to cognition and CSF biomarkers. *Hum Brain Mapp* **35**, 2383-2393.

[59] Contreras JA, Goni J, Risacher SL, Amico E, Yoder K, Dzemidzic M, West JD, McDonald BC, Farlow MR, Sporns O, Saykin AJ (2017) Cognitive complaints in older adults at risk for Alzheimer's disease are associated with altered resting-state networks. *Alzheimers Dement (Amst)* **6**, 40-49.

[60] Sun Y, Dai Z, Li Y, Sheng C, Li H, Wang X, Chen X, He Y, Han Y (2016) Subjective Cognitive decline: Mapping functional and structural brain changes-a combined resting-state functional and structural MR imaging study. *Radiology* **281**, 185-192.

[61] Hafkemeijer A, Altmann-Schneider I, Oleksik AM, van de Wiel L, Middelkoop HA, van Buchem MA, van der Grond J, Rombouts SA (2013) Increased functional connectivity and brain atrophy in elderly with subjective memory complaints. *Brain Connect* **3**, 353-362.

[62] Teipel SJ, Cavedo E, Weschke S, Grothe MJ, Rojkova K, Fontaine G, Dauphinot L, Gonzalez-Escamilla G, Potier MC, Bertin H, Habert MO, Dubois B, Hampel H, INSIGHT-preAD study group (2017) Cortical amyloid accumulation is associated with alterations of structural integrity in older people with subjective memory complaints. *Neurobiol Aging* **57**, 143-152.

[63] Agosta F, Pievani M, Geroldi C, Copetti M, Frisoni GB, Filippi M (2012) Resting state fMRI in Alzheimer's disease: Beyond the default mode network. *Neurobiol Aging* **33**, 1564-1578.

[64] Dai Z, Yan C, Li K, Wang Z, Wang J, Cao M, Lin Q, Shu N, Xia M, Bi Y, He Y (2015) Identifying and mapping connectivity patterns of brain network hubs in Alzheimer's disease. *Cereb Cortex* **25**, 3723-3742.

[65] Song J, Qin W, Liu Y, Duan Y, Liu J, He X, Li K, Zhang X, Jiang T, Yu C (2013) Aberrant functional organization within and between resting-state networks in AD. *PLoS One* **8**, e63727.

[66] de Vos F, Koini M, Schouten TM, Seiler S, van der Grond J, Lechner A, Schmidt R, de Rooij M, Rombouts SARB (2018) A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *Neuroimage* **167**, 62-72.

[67] Zhang Z, Liu Y, Jiang T, Zhou B, An N, Dai H, Wang P, Niu Y, Wang L, Zhang X (2012) Altered spontaneous activity in Alzheimer's disease and mild cognitive impairment revealed by regional homogeneity. *Neuroimage* **59**, 1429-1440.

Supplementary Material

Comorbidities and Medication

To give an overview of the participants' comorbidities and medication, we listed the diagnoses and drugs in relevant categories (Supplementary Tables 1 and 2). The data refer to the participants' conditions at the first examination date (DELCODE baseline) [1].

Overall four participants had a clinical history of stroke, but no current symptoms and no visible lesions in the MRI scans (T_1, T_2 and fMRI). Two participants had a polyneuropathy diagnosis, one suffered from restless legs syndrome and one had an unclassified disorder of the peripheral nervous system (without medical treatment), listed as "Others" in Supplementary Table 1 (line 4). Interestingly, eight participants had a history of depression, while 18 subjects were treated with antidepressants (17 with regular antidepressants, one with Quetiapine). This is not surprising when considering the frequent subscription of antidepressants through non-psychiatrists [2]. No participant fulfilled criteria of current major depression (see below). One person (healthy control) was diagnosed with generalized anxiety. A total of 36 participants reported substituted hypothyroidism as preexisting condition, but 43 subjects took Levothyroxine, hence we conclude the difference is caused by incomplete declaration from subjects.

Supplementary Table 2 gives an overview of the medication of all subjects. 17 subjects were treated with antidementia drugs (6 with Rivastigmine, 5 with Donepezil, 4 with Galatamine, 1 with Memantine; 1 was treated with Rivastigmine and Memantine). Two participants with MCI diagnosis were also treated with anti-dementia drugs. With regard to current references and clinical practice [3], the medication of those two subjects was ceased

during the ongoing trial. Other preparations with effect to the central nervous system included Gingko biloba, Levodopa, Pregabalin, and Gabapentin as well as Beta blockers. The listing of other potentially relevant drugs, such as proton-pump inhibitors or Statins is given in the lower part of Supplementary Table 2. The group "anticholinergic drugs" includes both drugs with a primary anticholinergic effect (e.g., the antimuscarinergic drug Solifenacin) and drugs with anticholinergic side-effects (e.g., Cetirizine). Please note that participants can appear multiple times in the table. On average, participants took 2.6 drugs with the most drugs in the dementia group (3.7). Only 37 participants did not take any drugs regularly.

Supplementary Methods

We repeated the functional connectivity analysis using the support vector machine (SVM) as classifier. We applied a radial basis kernel function to allow for nonlinear projections to the decision space. Optimal values for the two parameters cost factor C and radial basis width γ were determined using grid search in the range $C = 10^{-1}, 10^{0}, \ldots, 10^{4}$ and $\gamma = 2^{-10}, 2^{-9}, \ldots, 2^{2}$ based on tenfold cross-validation. As for the elastic net models, within the cross-validation procedure, the feature space was restricted to include approximately 100 of most informative features, determined as the set of variables which correlated with the diagnosis with a Pearson's correlation coefficient of an absolute value higher than the 97.5^{th} percentile. Cross-validation was repeated 1,000 times to assess the stability of group separation.

REFERENCES

[1] Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, Fliessbach K, Franke C, Fuentes M, Heneka MT, Janowitz D, Kilimann I, Laske C, Menne F, Nestor P, Peters O, Priller J, Pross V, Ramirez A, Schneider A, Speck O, Spruth EJ, Teipel S, Vukovich R, Westerteicher C, Wiltfang J, Wolfsgruber S, Wagner M, Düzel E (2018) Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimers Res Ther* **10**, 15.

[2] Mojtabai R, Olfson M (2011) Proportion of antidepressants prescribed without a psychiatric diagnosis is growing. *Health Aff (Millwood)* **30**, 1434-1442.

[3] Tricco AC, Soobiah C, Berliner S, Ho JM, Ng CH, Ashoor HM, Chen MH, Hemmelgarn B, Straus SE (2013) Efficacy and safety of cognitive enhancers for patients with mild cognitive impairment: a systematic review and meta-analysis. *Can Med Assoc J* **185**, 1393-1401.

Supplementary Table 1. Summary of participants' comorbidities

	HC	SCD	MCI	AD	Total of each class
Neurological diseases		**3**		**5**	**8**
• History of stroke		1		3	4
• Others[1]		2		2	4
Psychiatric diseases	**1**	**4**	**2**	**2**	**9**
• History of depression (currently remitted)		4	2	2	8
• History of anxiety disorder (currently remitted)	1				1
Hypothyroidism (substituted)	**14**	**15**	**3**	**4**	**36**
Total of each group	**15**	**22**	**5**	**11**	**53**

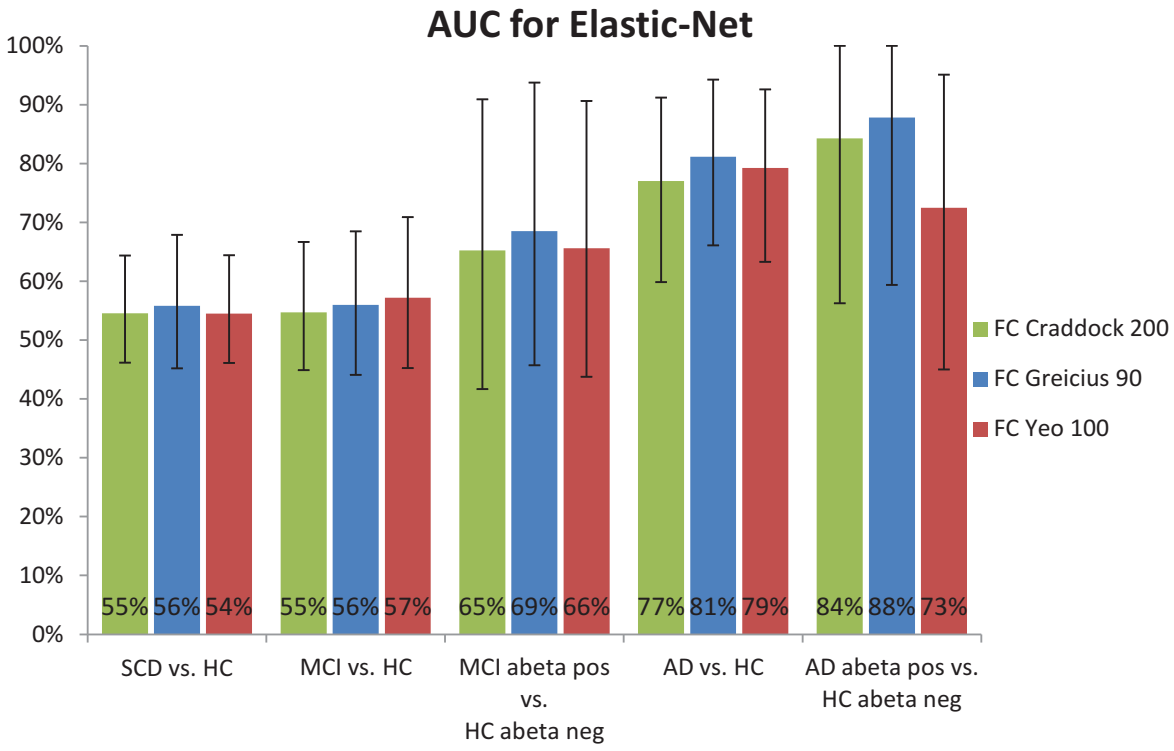
Number of participants diagnosed with relevant diseases sorted by clinical classifications.
[1] includes: peripheral neuropathy, restless legs, unclassified disorder of the peripheral nervous system
HC, healthy controls; SCD, subjective cognitive decline; MCI, mild cognitive impairment; AD, Alzheimer's diseasea

Supplementary Table 2. Overview of participants' medication

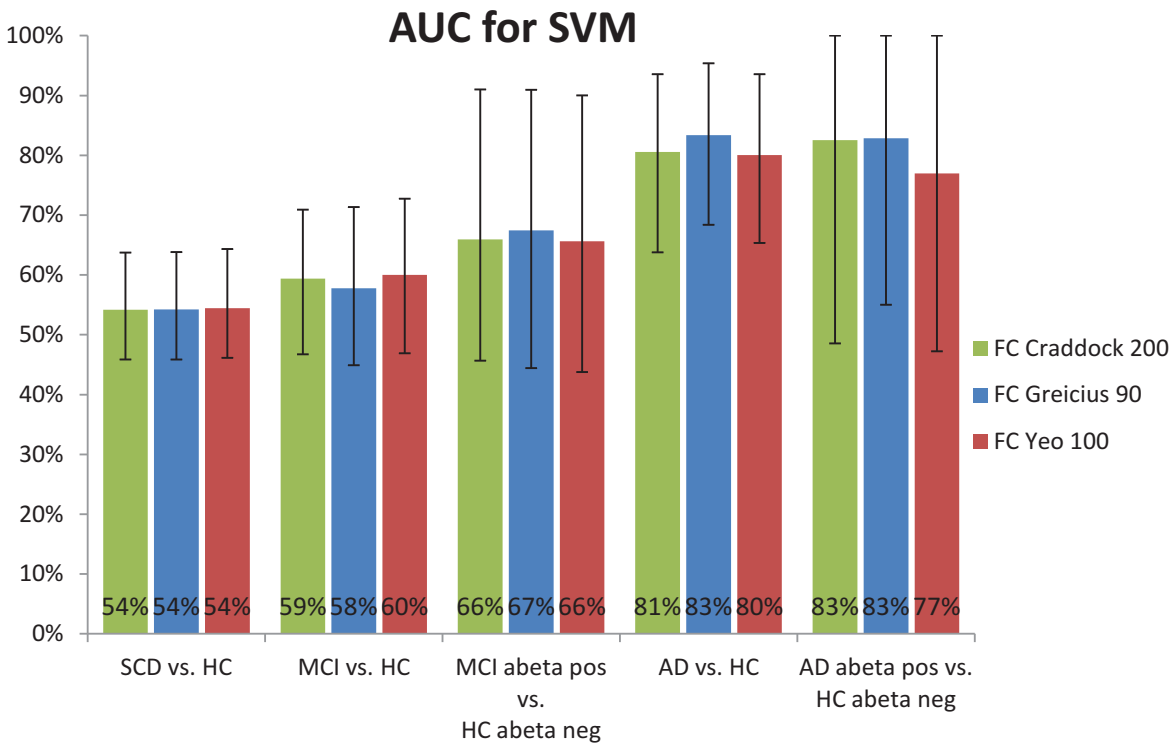
	HC	SCD	MCI	AD	Total of each class
Dementia treatment			**2**	**18**	**20**
• Acetylcholinesterase inhibitors			1	16	17
• NMDA-antagonists			1	2	3
Psychiatric medication	**2**	**8**	**5**	**4**	**19**
• Antidepressants	2	8	5	2	17
• Low potential antipsychotics				2	2
Others with known effect to CNS[1]		**7**	**1**	**3**	**11**
Beta blockers	**14**	**19**	**11**	**8**	**52**
Others					
• Anticholinergic drugs	2	4		1	7
• Levothyroxine	14	21	3	5	43
• NSAIDs	1	2	1		4
• Proton-pump inhibitors	5	10	3	3	21
• Statins	9	21	14	6	50

Number of participants treated with relevant drugs, sorted by drug category. Participants can appear multiple times (for different drug categories).
HC, healthy controls; SCD, subjective cognitive decline; MCI, mild cognitive impairment; AD, Alzheimer's disease
[1]Gingko biloba, Levodopa (restless legs), pregabalin, gabapentin

Supplementary Fig. 1. Comparison of the AUC for functional connectivity based on different atlases using elastic-net penalized logistic regression



Supplementary Fig. 2. Comparison of the AUC for functional connectivity based on different atlases using support vector machine classification



Journal of Alzheimer's Disease 72 (2019) 455–465
DOI 10.3233/JAD-190446
IOS Press

Multicenter Tract-Based Analysis of Microstructural Lesions within the Alzheimer's Disease Spectrum: Association with Amyloid Pathology and Diagnostic Usefulness

Stefan J. Teipel[a,b,*], Jan O. Kuper-Smith[a], Claudia Bartels[c,d], Frederic Brosseron[e,f], Martina Buchmann[g,h], Katharina Buerger[i,j], Cihan Catak[j], Daniel Janowitz[j], Peter Dechent[k], Laura Dobisch[l], Birgit Ertl-Wagner[m,x], Klaus Fließbach[e,f], John-Dylan Haynes[n], Michael T. Heneka[e,f], Ingo Kilimann[a,b], Christoph Laske[g,h], Siyao Li[o], Felix Menne[o], Coraline D. Metzger[l,p,q], Josef Priller[r,s], Verena Pross[t], Alfredo Ramirez[u], Klaus Scheffler[v], Anja Schneider[e,f], Annika Spottke[e,w], Eike J. Spruth[r,s], Michael Wagner[e,f], Jens Wiltfang[c,d], Steffen Wolfsgruber[e], Emrah Düzel[l,p], Frank Jessen[e,u], Martin Dyrba[b]
and the DELCODE study group
[a]*Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany*
[b]*German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*
[c]*German Center for Neurodegenerative Diseases (DZNE), Goettingen, Germany*
[d]*Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, University of Goettingen, Goettingen, Germany*
[e]*German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*
[f]*Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany*
[g]*German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany*
[h]*Section for Dementia Research, Hertie Institute for Clinical Brain Research and Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany*
[i]*German Center for Neurodegenerative Diseases (DZNE, Munich), Munich, Germany*
[j]*Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany*
[k]*MR-Research in Neurology and Psychiatry, Georg-August-University Göttingen, Göttingen, Germany*
[l]*German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany*
[m]*Institute for Clinical Radiology, Ludwig-Maximilians-University, Munich, Germany*
[n]*Bernstein Center for Computational Neuroscience, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[o]*Institute of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[p]*Institute of Cognitive Neurology and Dementia Research (IKND), Otto-von-Guericke University, Magdeburg, Germany*
[q]*Department of Psychiatry and Psychotherapy, Otto-von-Guericke University, Magdeburg, Germany*

*Correspondence to: Stefan J. Teipel, MD, Department of Psychosomatic Medicine, University Medicine Rostock, and DZNE, Gehlsheimer Str. 20, 18147 Rostock, Germany. Tel.: +49 381 494 9470; Fax: +49 381 494 9682; E-mail: stefan.teipel@med.uni-rostock.de.

ISSN 1387-2877/19/$35.00

[r]*Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[s]*German Center for Neurodegenerative Diseases (DZNE), Berlin, Germany*
[t]*Study Center Bonn, Medical Faculty, Bonn, Germany*
[u]*Department of Psychiatry, University of Cologne, Cologne, Germany*
[v]*Department for Biomedical Magnetic Resonance, University of Tübingen, Tübingen, Germany*
[w]*Department of Neurology, University of Bonn, Bonn, Germany*
[x]*Division of Neuroradiology, Department of Medical Imaging, The Hospital for Sick Children, Toronto, Canada*

Handling Associate Editor: Juan Helen Zhou

Accepted 30 August 2019

Abstract. Diffusion changes as determined by diffusion tensor imaging are potential indicators of microstructural lesions in people with mild cognitive impairment (MCI), prodromal Alzheimer's disease (AD), and AD dementia. Here we extended the scope of analysis toward subjective cognitive complaints as a pre-MCI at risk stage of AD. In a cohort of 271 participants of the prospective DELCODE study, including 93 healthy controls and 98 subjective cognitive decline (SCD), 45 MCI, and 35 AD dementia cases, we found reductions of fiber tract integrity in limbic and association fiber tracts in MCI and AD dementia compared with controls in a tract-based analysis ($p < 0.05$, family wise error corrected). In contrast, people with SCD showed spatially restricted white matter alterations only for the mode of anisotropy and only at an uncorrected level of significance. DTI parameters yielded a high cross-validated diagnostic accuracy of almost 80% for the clinical diagnosis of MCI and the discrimination of Aβ positive MCI cases from Aβ negative controls. In contrast, DTI parameters reached only random level accuracy for the discrimination between Aβ positive SCD and control cases from Aβ negative controls. These findings suggest that in prodromal stages of AD, such as in Aβ positive MCI, multicenter DTI with prospectively harmonized acquisition parameters yields diagnostic accuracy meeting the criteria for a useful biomarker. In contrast, automated tract-based analysis of DTI parameters is not useful for the identification of preclinical AD, including Aβ positive SCD and control cases.

Keywords: amyloid, anisotropy, cerebral white matter, cognition, diagnosis, diffusion tensor imaging, mild cognitive impairment, subjective cognitive decline

INTRODUCTION

Microstructural lesions of associative fiber tracts in Alzheimer's disease (AD) likely result from primary cell loss in grey matter regions but also reflect primary white matter pathology such as myelin break down (for review, see [1]). Diffusion tensor imaging (DTI) can usefully be employed for the *in vivo* detection of such lesions [2, 3] showing moderate to high diagnostic accuracy in mild cognitive impairment (MCI) as a prodromal stage of AD compared with healthy controls in monocenter studies [4, 5]. In addition, a multicenter study from retrospectively pooled DTI data [6] suggested a high diagnostic utility (about 77% cross-validated accuracy) for the discrimination between amyloid positive people with mild cognitive impairment (MCI) and healthy controls [7]. In this study, DTI was superior to volumetric measures despite high vulnerability of the DTI parameters to multicenter variability [8]. This finding was very interesting because a promising biomarker for AD should also prove itself in a multicenter setting [9], which is much closer to the future application in routine care than a monocenter study.

Subjective cognitive decline (SCD) is a clinical at-risk stage for MCI and dementia [10]. SCD cases have a two-fold increased risk to develop dementia and a six-fold increased risk to develop MCI over on average four years of clinical follow-up compared with cognitively normal people without subjective cognitive complaints [11]. Several studies found significant differences in DTI parameters, such as fractional anisotropy or mean diffusivity, in SCD cases compared with controls [12–15]. Studies on the diagnostic utility of DTI markers, however, for the discrimination between SCD cases and controls are still scarce. One recent monocenter study reported an area under the ROC curve of 78% for the discrimination between

20 SCD cases and 22 controls, but only in the training sample, i.e., without cross-validation [16].

Since reliable acquisition across different sites is an important prerequisite for a potentially useful biomarker, here we tested group differences and diagnostic usefulness of microstructural lesion markers across the entire AD spectrum from a prospective multicenter DTI acquisition with harmonized acquisition parameters. The cohort spans from cognitively healthy controls through SCD and MCI to AD dementia. The current analysis focused on two endpoints, discrimination of the clinical diagnoses SCD and MCI from healthy controls and the discrimination of amyloid positive SCD and MCI cases, representing preclinical or prodromal stages of AD [17], from amyloid negative controls based on cerebrospinal fluid (CSF) amyloid-β (Aβ). We used tract-based statistics that was found in a previous multicenter reliability study to be less prone to multicenter effects than voxel-based analysis [8]. The motivation of our study was that if found useful in a prospective cohort tract-based statistics of microstructural lesion markers may be employed for risk stratification of study participants in future clinical trials.

MATERIALS AND METHODS

Participants

Data used in this study came from baseline data of the DELCODE (*DZNE Longitudinal Cognitive Impairment and Dementia*) study, an ongoing observational longitudinal memory clinic-based multicenter study in Germany [18]. A total of 282 participants from nine centers were included in this study. Two cases were excluded due to neuroimaging issues. One scan deviated in the number of slices (47 instead of 72) and slice spacing (3.5 mm instead of 2 mm). Another participant had a small falx meningioma and was excluded to avoid problems with the image processing algorithms. This left us with a final number of 280 participants from nine sites. DELCODE exclusion criteria dictate that no participant should have past or present unstable medical conditions, major psychiatric disorders, including a current major depressive episode, or neurologic diseases that are not AD [18]. The basis of these exclusion criteria was provided by a clinical assessment of cognitive status, which included the Geriatric Depression Scale (GDS) [19], and an extensive neuropsychological testing battery [18].

The sample included people with SCD who were cognitively unimpaired and stated to have decline in cognitive functioning unrelated to an event or condition explaining the cognitive deficits according to research criteria [10], MCI who met National Institute on Aging – Alzheimer's Association (NIA-AA) workgroup core criteria for MCI [20], AD dementia who met the NIA-AA probable AD dementia criteria [21], and cognitively normal controls who never reported SCD and had no history of neurological or psychiatric disease or any sign of cognitive decline.

Written informed consent was provided by all participants, or their representatives. The study was approved by local ethics committees at each of the participating centers, and has been conducted in accord with the Helsinki Declaration of 1975.

Imaging data acquisition

Imaging data at the nine different DZNE sites were obtained using Siemens 3.0 Tesla MRI scanners (three Verio, three TimTrio, one Prisma, two Skyra) using the same acquisition parameters and instructions. An axial diffusion sequence was measured based on a single shot echo planar imaging sequence (acquisition time: 14 min 45 s, field of view: 240x240 mm, isotropic voxel size: 2 mm, repetition time: 12100 ms, echo time: 88 ms, flip angle: 90°, number of gradients: 60, b-values: 700 s/mm^2 and 1000 s/mm^2, number of slices: 72, parallel imaging acceleration factor: 2). High-resolution T1-weighted anatomical images were obtained using a sagittal magnetization-prepared rapid gradient echo (MPRAGE) sequence (acquisition time: 5 min 8 s, field of view: 256×256 mm, isotropic voxel size: 1 mm, echo time: 4.37 ms, flip angle: 7°, repetition time: 2500 ms, number of slices: 192, parallel imaging acceleration factor: 2). To ensure high image quality, all scans had to pass a semi-automated check for SOP conformity and scan quality during data collection so that protocol deviations were reported to the study sites promptly, in order to allow the sites to adjust their acquisition. Additionally, all scans were visually inspected and controlled for 1) proper alignment of the field-of-view to cover the whole brain, 2) screened for severe imaging artifacts (e.g., aliasing/ghosting, strong noise/motion or susceptibility artifacts from metallic dental fillings), and 3) checked for incidental findings such as old strokes or meningiomas.

Biomaterial sampling

Biomaterial sampling included CSF in those participants, who consented. Trained study assistants performed the collection, processing and storage of the samples up to the shipment to the central biorepository of the DZNE according to SOP. After the centrifugation CSF was aliquoted and stored at –80°C.

Image processing

Due to varying degrees of atrophy between participants, accurately registering white matter (WM) into a standard space is problematic for whole-brain deformation approaches [22], especially when considering smaller anatomical structures such as the fornix [23, 24]. In addition, in a previous clinical and physical phantom study, tract-based statistics (TBSS) was found less prone to scanner effects than voxel-based analysis [8]. Therefore, we used TBSS [25] in FSL 5.0.9 for DTI data analysis. First, diffusion scans were corrected for distortions using a gradient-echo field map and the T1-weighted scans by applying fsl_prepare_fieldmap and epi_reg commands. After head motion and eddy current correction [26] using eddy_correct with spline interpolation, surrounding skull matter was removed from the non-diffusion-weighted images using FSL's brain extraction tool (bet2) [27] and diffusion tensor models were fitted using FSL's dtifit command to derive voxel-wise FA, MD, and mode of anisotropy values. The next steps involved aligning all subjects' FA images in a voxelwise nonlinear registration to MNI152 reference space, and creating a mean FA average from these transformed FA images. We then created a custom mean FA skeleton, which was thresholded at 0.2 in order to exclude more peripheral tracts with lower inter-subject correspondence. The individual participants' FA maps were then projected onto the skeleton by assigning the maximum FA value in perpendicular tract direction to the skeleton voxel at each point of the skeleton. This projection information was subsequently applied to mode of anisotropy and MD maps as well.

For comparison, we used classification accuracy based on hippocampus volume, an established biomarker of AD [28]. For hippocampus volumetry, we used the harmonized hippocampus segmentation protocol, an internationally driven effort under the auspices of the Alzheimer's Association [29]. Further details can be found on the project's website (http://www.hippocampal-protocol.net/SOPs/index.php). More recently, the manual hippocampus labels were integrated into an automated volumetry pipeline to ease processing of larger numbers of cases [30]. A high correspondence between manual and automated segmented hippocampi based on the harmonized protocol was shown in 135 MRI scans that were measured using both manual segmentation and automated segmentation. Following this automated processing pipeline, the T1-weighed MPRAGE images were normalized to the MNI reference template from CAT12 using SPM12 new segment and the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm [31]. Subsequently, hippocampus volume was automatically computed from all voxels within the harmonized hippocampus mask regions of the normalized and modulated grey matter maps [30]. The raw volume estimates were proportionally scaled to total intracranial volume to adjust for head size.

CSF AD biomarker assessment

CSF $A\beta_{42}$ and $A\beta_{40}$ levels were determined using commercially available kits according to vendor specifications (V-PLEX Aβ Peptide Panel 1 (6E10) Kit). Cut-offs for normal and abnormal concentrations of $A\beta_{42}$ (<496 pg/ml), and of the ratio $A\beta_{42}/A\beta_{40}$ (<0.09) were derived from the literature, which applied the respective assays [32]. Correspondingly, cases below the cut-off of 0.09 for the ratio $A\beta_{42}/A\beta_{40}$ were designated amyloid positive, cases above this cutoff as amyloid negative.

Data analysis

Demographic data

Participants' demographic information was analyzed using appropriate tests as needed: gender distribution was assessed using χ^2-test, while differences in age, years of education, and MMSE scores were assessed with ANOVA models.

Voxelwise TBSS analysis

Voxelwise cross-subject comparisons were performed with the control group as reference; i.e., controls versus SCD, controls versus MCI, and controls versus AD. The models included age, sex, and scanner as covariates. For the main results, we applied a significance threshold of $p < 0.05$, family-wise error (FWE) corrected, and for statistical trends

$p < 0.01$ uncorrected for multiple comparisons. Statistical differences were estimated by a permutation test approach with 5000 random permutations defining a null distribution of regression parameters.

We used variance component analysis to assess the vulnerability of tract-based DTI parameters to multicenter effects. For this, we extracted the subject level TBSS clusters using the FSL function "fslmeants" averaging the values of all significant voxels for the group comparisons of SCD cases versus controls and of MCI cases versus controls, respectively. This resulted in a scalar value for each individual for each comparison and each DTI parameter. We determined a random effects model in R using library "lmer" with the averaged cluster values as dependent variable and scanner as random effects variable. We then extracted the amount of variance attributable to the random effect of scanner divided by the total amount of variance, providing an estimate of the variance component for scanner for a given comparison and DTI parameter.

Elastic net regression

We calculated group discrimination using a penalized logistic regression model with an elastic net penalty, using the R package glmnet (available at http://cran.r-project.org/web/packages/glmnet/index.html). In an elastic net regression, two penalty terms are added as an extension of the residual sum of squares minimization of traditional linear regression models to account for high collinearity of regression features [33]. A detailed description of this method as applied to multicenter imaging data can be found in [34]. Mean values were extracted from the normalized and skeletonized FA, mode of anisotropy and MD maps using the JHU-ICBM DTI atlas containing labels for 48 major white matter tracts [35]. The vectors of all three DTI indices were concatenated and entered as predictors in the logistic regression models. To assess diagnostic accuracy, we followed a stringent cross-validation procedure based on 100 times repeated 2/3 by 1/3 cross-validation. Estimations of the area under the receiver-operating-characteristics curve (AUC) were made for each of the iterations from the test sample. Due to currently missing extended inference techniques for elastic-net models, beta coefficients estimates based on the whole sample as well as selection rates for each beta based on cross validation iterations, are reported.

Models were calculated with clinical diagnosis as dependent variable as well as with the discrimination of $A\beta_{42}/A\beta_{40}$-ratio positive controls/SCD/MCI cases versus $A\beta_{42}/A\beta_{40}$-ratio negative controls where CSF data were available.

Calculating AUCs for classifying patient groups using hippocampal volumetry used cross-validated unpenalized logistic regression due to lack of collinear covariates.

All model calculations were repeated after adding age, sex, and scanner to assess sensitivity of outcomes to these parameters.

RESULTS

Demographic data

Of the 280 included participants, 93 were classified as healthy controls, 9 as cognitively normal first degree relatives of people with AD dementia, 98 as SCD, 45 as MCI, and 35 as AD dementia (see Table 1 for additional participant information). Due to the small number of cases the group of people with first degree relatives with AD dementia was left out from the subsequent analyses.

The remaining groups differed in respect to gender distribution, age, and years of education (see Table 1 for details on demographic characteristics). As expected and required by diagnostic criteria, SCD and controls did not differ in MMSE scores ($t = 1.57$, 189 df, $p = 0.12$), while MCI and AD cases showed significantly lower MMSE scores compared with controls ($t = 7.5$, 136 df, $p < 0.001$, and $t = 18.4$, 126 df, $p < 0.001$, respectively).

CSF was available in 36 controls with 25 having a normal $A\beta_{42}/A\beta_{40}$-ratio, in 43 SCD cases with 17 having an abnormal $A\beta_{42}/A\beta_{40}$-ratio, and in 31 MCI cases with 20 having an abnormal $A\beta_{42}/A\beta_{40}$-ratio.

Table 1
Participants' demographic characteristics

N=271	AD	MCI	SCD	Controls
No. cases (women)[a]	35 (19)	45 (14)	98 (47)	93 (55)
Age (SD) [y][b]	73.5 (6.8)	72.3 (5.7)	71.3 (5.9)	68.5 (5.1)
Education (SD)[c] [y]	13.4 (3.1)	14.4 (3.1)	14.6 (3.1)	15.1 (2.6)
MMSE (SD)[d]	23.1 (3.1)	28.0 (1.6)	29.3 (0.9)	29.5 (0.8)

[a]significantly different between groups, $\chi^2 = 9.95$, 3 df, $p = 0.02$; [b]significantly different between groups, F(3, 267)=8.7, $p < 0.001$; [c]significantly different between groups, F(3, 267)=2.7, $p < 0.05$; [d]significantly different between groups, F(3, 267)=184.3, $p < 0.001$.

Voxelwise TBSS analysis

All models were controlled for age, sex, and scanner as covariates. For SCD versus controls, we did not find significant group differences in FA, MD, nor mode of anisotropy at $p < 0.05$ FWE-corrected. Only at an uncorrected $p < 0.01$ we found reduced mode of anisotropy in SCD compared with controls in left predominant fornix, fusiform gyrus and superior temporal gyrus white matter, and anterior thalamic radiation (Fig. 1).

For MCI versus controls, we found widespread reductions of FA and increases of MD across limbic and association white matter fiber tracts at $p < 0.05$ FWE-corrected (Fig. 2). Reductions in the mode of anisotropy were most focused on the corpus callosum and cingulate gyrus, but also involving external and internal capsule and uncinate fasciculus.

For AD versus controls, reductions of FA and increases of MD were similarly widespread as for MCI cases at $p < 0.05$, FWE-corrected. Similar to MCI cases, reductions in the mode of anisotropy were focused on the corpus callosum and cingulate gyrus, and also involved external and internal capsule and uncinate fasciculus.

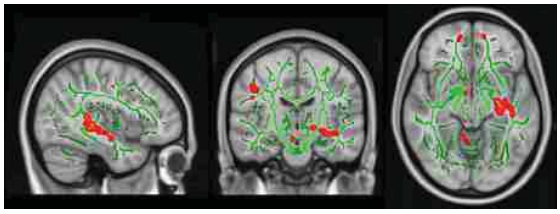


Fig. 1. Differences of mode of anisotropy in SCD cases compared with controls. Projection of the differences of mode of anisotropy values between SCD cases and controls (mode of anisotropy smaller in SCD than in controls) in red to yellow color on the group specific averaged TBSS fiber tract skeleton (green color) in MNI standard space. Effects are thresholded at $p < 0.01$, uncorrected for multiple comparisons

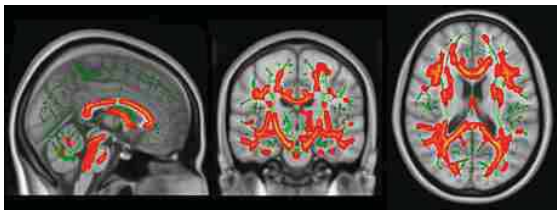


Fig. 2. Differences of FA in in MCI cases compared with controls. Projection of the differences of FA values between MCI cases and controls (FA smaller in MCI than in controls) in red to yellow color on the group specific averaged TBSS fiber tract skeleton (green color) in MNI standard space. Effects are thresholded at $p < 0.05$, FWE corrected for multiple comparisons

In a complementary analysis, we studied tract-based changes of axial and radial diffusivity. We found widespread increase of radial and axial diffusivity in AD and MCI, but not in SCD, compared with controls at $p < 0.05$, FWE-corrected, with a large spatial overlap with the FA decreases (data not shown).

The variance component for scanner was 9.6% for the mode of anisotropy in the combined clusters discriminating SCD cases and controls. For comparison, the variance component for hippocampus volume in the subsample of SCD and control participants was 1.3%. For the MCI versus controls comparison, the variance component for scanner for the pooled FA clusters was 20.7%, for the MD cluster 29.2%, and for the mode of anisotropy cluster 11.5%; for the hippocampus it was below 1%.

Assessment of diagnostic accuracy

The variance inflation factor (VIF) for each predictor was calculated as the corresponding diagonal element of the inverse of the cross-correlation matrix [36]. The mean VIF was 44.0 across all DTI parameters and diagnoses. This suggested a high level of collinearity and motivated the use of elastic net logistic regression to account for it.

Mean area under the curve (AUC) with the 95% confidence intervals for the cross-validated elastic net regression of each group compared to classification via hippocampal volumetry are shown in Figure 3. For the comparison of SCD cases with controls, DTI parameters were numerically superior to hippocampus volume with 69% versus 62% AUC. However, for the comparison of $A\beta_{42}/A\beta_{40}$-ratio positive SCD cases (i.e., preclinical AD) versus $A\beta_{42}/A\beta_{40}$-ratio negative controls, diagnostic accuracy reached only 55% AUC for the DTI parameters and 60% AUC for hippocampus volume, respectively. For MCI cases versus controls and for $A\beta_{42}/A\beta_{40}$-ratio positive MCI cases (i.e., prodromal AD) versus $A\beta_{42}/A\beta_{40}$-ratio negative controls, diagnostic accuracy for DTI parameters was 78% for both comparisons, and 77% and 83% for hippocampus volume, respectively. These numbers were almost unchanged when adding age, sex, and scanner to the classification models.

Table 2 lists important diffusivity measures of specific tract locations that were selected at least in 90% of the bootstrapped models for classifying SCD, SCD-$A\beta_{42/40}$-ratio-positive, MCI, and MCI $A\beta_{42/40}$-ratio-positive participants.

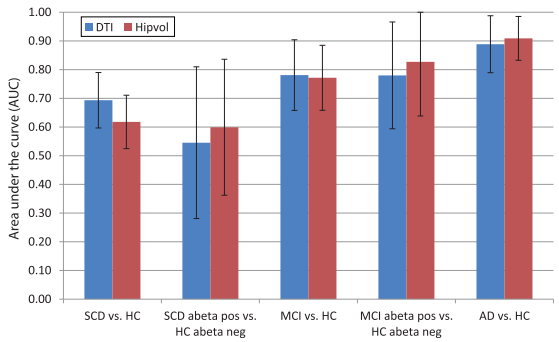


Fig. 3. Group discrimination based on multimodal DTI parameters and hippocampus volume. Cross-validated (100 iterations) areas under the ROC curves (AUC) with 95% confidence intervals for the group classification of participants of SCD, MCI, and AD versus controls in addition to MCI/SCD amyloid-β positive versus amyloid-β negative controls, based on multimodal tract-based DTI parameters (DTI) and hippocampal volume (Hipvol), respectively.

In an additional analysis, the diagnostic accuracy of DTI parameters comparing 11 $A\beta_{42/40}$-ratio-positive controls plus 17 $A\beta_{42/40}$-ratio-positive SCD cases, representing a preclinical AD group, versus 25 $A\beta_{42/40}$-ratio-negative controls plus 26 $A\beta_{42/40}$-ratio-negative SCD cases, reached a cross-validated AUC of 50%, i.e., random level accuracy.

DISCUSSION

Here, we studied diffusion changes as surrogate markers of microstructural lesions in cerebral white matter in clinically defined at risk stages of AD, including SCD and MCI cases, as well as in preclinical and prodromal AD cases, represented by CSF amyloid positive cognitively normal people and SCD cases. To assess the potential usefulness of these markers for future diagnostic applications we used a prospective multicenter data set within a cross-validation framework to assess the diagnostic accuracy of diffusion markers in the presence of multicenter variability.

In agreement with previous studies, we found reductions of FA and mode of anisotropy and increases of MD, both in MCI cases in general as well as in Aβ-positive MCI cases compared with controls in widespread white matter fiber tracts including projections from hippocampus, as well as association fiber tracts but also parts of the internal capsule, projecting into brain stem white matter. This regional pattern was very similar to previous reports from monocenter studies [4, 5, 37–40] and the pattern of diffusion changes found in the AD dementia group of the current cohort. In conclusion, these findings suggest widespread white matter microstructural degeneration already in prodromal stages of AD. The reductions of FA in MCI and AD cases spatially widely overlapped with increases of radial and axial diffusivity, in agreement with an earlier report in AD and MCI cases from the ADNI cohort [41]. Consistent with a recent review [42], this would indicate widespread damage including impaired axonal integrity, edema and myelin damage as cause of reduced anisotropy.

Cross-validated diagnostic accuracy of DTI parameters for the distinction of MCI and MCI Aβ-positive cases from controls reached almost 80%, but was not superior to hippocampus volume, the best-established structural imaging marker of AD to date.

Table 2
Most frequently selected features for diagnostic group discrimination

SCD versus controls				
β coefficient	Frequency (%)	Diffusivity measure and region	Mean tract diffusivity value	
			Patient	Controls
–0.372	100	MO retrolenticular part of internal capsule	0.225	0.233
–0.302	96	MO posterior thalamic radiation L	0.217	0.227
–0.260	93	MO posterior limb of internal capsule R	0.172	0.176
–0.237	91	MO fornix/stria terminalis R	0.113	0.124
–0.239	91	MO uncinate fasciculus L	0.114	0.120
$A\beta_{42/40}$-ratio positive SCD versus $A\beta_{42}$ negative controls				
		_[1]		
MCI versus controls				
–0.023	100	MO medial lemniscus R	0.207	0.219
$A\beta_{42/40}$-ratio positive MCI versus $A\beta_{42}$ negative controls				
		_[2]		

[1]Highest selection frequency was 71%; [2]highest selection frequency was 88%.

This agrees with the level of accuracy found in previous monocenter studies [4, 5], including studies using tract-based statistics [43]. Interestingly, the elastic net algorithm selected only few fiber tracts with very high frequency (>90%), another set of association fiber tracts had between 50% to 80% selection frequency. This is consistent with the widespread alterations of white matter fiber tracts, suggesting comparable diagnostic value of a wide range of tracts. The mode of anisotropy is a scalar diffusion marker that describes an important aspect of the shape of the diffusion tensor. It ranges between –1 and 1 as the shape of the diffusion tensor ranges from planar anisotropy (in areas with crossing fiber populations or adjacent orthogonal fiber orientations) through orthotropy to linear anisotropy (in areas with one predominant fiber direction) [44]. In a study using joint independent component analysis, the mode of anisotropy was decreased in MCI subjects compared to controls mainly in anterior and posterior corpus callosum and in superior and inferior longitudinal fasciculus [45]. Consistently, in the current analysis we found reductions of the mode of anisotropy in MCI in corpus callosum and cingulate gyrus, but also involving external and internal capsule and uncinate fasciculus. Thus, the mode of anisotropy reductions involved mainly regions with directed fiber tracts such as corpus callosum or cingulate bundle, indicating a loss of these directed fibers, resulting in a more orthotropic or planar shape of the diffusion tensor.

SCD cases exhibited no significant fiber tract alterations at family wise error corrected p-values. Only at a liberal uncorrected level of significance of $p < 0.01$, we found reductions in the mode of anisotropy in restricted white matter regions, including left predominant medial temporal lobe projections. This agrees with previous monocenters studies, where diffusion parameter changes were found in some [13, 16, 46–48], but not all studies [49, 50], and one of the positive studies did not apply a correction for multiple comparisons [13]. In addition, none of these previous studies included amyloid markers to assess preclinical AD status of the SCD cases. The inconsistency of results across studies may reflect an only low to moderate effect size of diffusion parameter changes in SCD cases. Mode of anisotropy changes in SCD cases compared with controls have not been reported before, rendering a comparison of our findings with previous results unfeasible. The only previous study on mode of anisotropy changes in SCD used the SCD group as reference group for comparison with AD and MCI [15].

The likelihood of false positive findings was high in the analysis that was uncorrected for multiple comparisons. Still, it is interesting to discuss why the mode of anisotropy, indicating a change in the diffusion tensor toward a more sphere like shape, may be the earliest diffusion marker affected in SCD in our analysis. One possible reason may be that among the three diffusion markers studied, the mode of anisotropy was least affected by multicenter effects, with a variance component of about 10% attributable to scanner, as compared with 20% for FA and almost 30% for MD. In addition, one could speculate that the reduction of the mode of anisotropy indicates a selective loss of highly directional fiber tracts within particularly vulnerable regions, such as fornix and temporal lobe white matter, but this needs to be confirmed in subsequent studies.

Diagnostic accuracy for the discrimination of SCD cases from healthy controls was only moderate to low with 69%, and mainly driven by reductions of mode of anisotropy in the uncinate fascicle, fornix and retrolenticular part of internal capsule, but substantially higher than for hippocampus volume with 62%. This compares with an AUC of 78% for the discrimination between 20 SCD cases and 22 controls in a previous study that did, however, not use cross-validation and therefore strongly overestimated the level of accuracy [16].

When comparing Aβ-positive SCD cases, representing a preclinical stage of AD, with Aβ-negative controls, DTI parameters only reached random level guessing accuracy. The same was true for the comparison of the Aβ-positive and Aβ-negative controls and the combined analysis of Aβ-positive SCD and controls versus the Aβ-negative SCD cases and controls. The latter analysis included 28 amyloid positive and 51 amyloid negative cases; this substantial number of cases suggests that the lack of an effect is not just a false negative outcome. In conclusion, the white matter alterations in the SCD cases (found at an uncorrected level of significance) may be related to the clinical phenotype rather than the underlying amyloid pathology. SCD is an unspecific syndrome related to "numerous conditions such as normal aging, personality traits, psychiatric conditions, neurologic and medical disorders, substance use, and medication. It may also be affected by the individual cultural background." ([10], pages 845/646). Thus, the white matter alterations in the SCD phenotype cases may reflect a trait from a broad range of conditions which are independent from the state of Aβ pathology. Consistently, not all cases with SCD

progress to MCI [11] and not all MCI cases have transited through a state of SCD [51]. This would suggest that similar to white matter changes preceding the first episode of major depression [52], the white matter alterations in SCD cases may reflect a functional (and in some cases reversible) clinical syndrome, rather than the effect from neuropathological lesions. The DELCODE protocol excludes current or past episodes of major depression as well as significant cerebrovascular disease. However, subsyndromal depressive symptoms or subclinical personality traits, such as increased anxiety, that may be related to subtle white matter alterations were not excluded.

There are several limitations associated with this study. First, the number of cases with available CSF was smaller than one would have wished for. Still, to our knowledge, this is the first DTI study featuring a substantial number of SCD cases stratified according to their amyloid status. Secondly, multicenter variability affects the accuracy of diffusion markers. This was even true for the prospectively harmonized DTI data acquisition and the use of tract bases spatial statistic (that was found less sensitive to multicenter effects than voxel-based analysis in a clinical phantom study [8]). Multicenter acquisition, however, is required if one wants to test the useful of DTI measures for the application in future (multicenter) clinical trials or routine care. The variance component analysis revealed that between 9% and 29% of variance was attributable to scanner for the DTI parameters. This compares favorably with previous analysis on multicenter DTI parameters from retrospectively pooled data with more than 40% of variance attributable to scanner for FA and MD parameters in a voxel-based analysis [53]. But even mode of anisotropy that was the least affected by scanner effects among the diffusion parameters was still much more affected than hippocampus volume. Other shortcomings are the different age and sex distribution across the diagnostic groups. However, including these variables together with scanner did not affect the outcome of the diagnostic models in sensitivity analyses.

In conclusion, we found significant differences in widespread white matter tracts in MCI and AD cases compared with controls, including Aβ positive MCI cases, representing prodromal AD. In contrast, white matter alterations were detectable only at an uncorrected level of significance and spatially restricted in the SCD cases and were entirely absent in Aβ positive compared with Aβ negative SCD and control cases, suggesting an effect of clinical phenotype of SCD rather than of preclinical Aβ pathology on white matter tract integrity in this cohort. In the near future, we will have access to the longitudinal data of the DELCODE cohort allowing assessment of the predictive accuracy of DTI parameters for cognitive decline within the AD spectrum.

DISCLOSURE STATEMENT

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/19-0446r1).

REFERENCES

[1] Caso F, Agosta F, Filippi M (2016) Insights into white matter damage in Alzheimer's disease: From postmortem to *in vivo* diffusion tensor MRI studies. *Neurodegener Dis* **16**, 26-33.

[2] Le Bihan D, Mangin JF, Poupon C, Clark CA, Pappata S, Molko N, Chabriat H (2001) Diffusion tensor imaging: Concepts and applications. *J Magn Reson Imaging* **13**, 534-546.

[3] Moore EE, Hohman TJ, Badami FS, Pechman KR, Osborn KE, Acosta LMY, Bell SP, Babicz MA, Gifford KA, Anderson AW, Goldstein LE, Blennow K, Zetterberg H, Jefferson AL (2018) Neurofilament relates to white matter microstructure in older adults. *Neurobiol Aging* **70**, 233-241.

[4] Cui Y, Wen W, Lipnicki DM, Beg MF, Jin JS, Luo S, Zhu W, Kochan NA, Reppermund S, Zhuang L, Raamana PR, Liu T, Trollor JN, Wang L, Brodaty H, Sachdev PS (2012) Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: A combined spatial atrophy and white matter alteration approach. *Neuroimage* **59**, 1209-1217.

[5] Henf J, Grothe MJ, Brueggen K, Teipel S, Dyrba M (2018) Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction. *Neuroimage Clin* **17**, 579-586.

[6] Brueggen K, Grothe MJ, Dyrba M, Fellgiebel A, Fischer F, Filippi M, Agosta F, Nestor P, Meisenzahl E, Blautzik J, Frolich L, Hausner L, Bokde ALW, Frisoni G, Pievani M, Kloppel S, Prvulovic D, Barkhof F, Pouwels PJW, Schroder J, Hampel H, Hauenstein K, Teipel S (2017) The European DTI Study on Dementia – A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. *Neuroimage* **144**, 305-308.

[7] Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirste T, Teipel SJ; EDSD study group (2015) Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. *J Neuroimaging* **25**, 738-747.

[8] Teipel SJ, Reuter S, Stieltjes B, Acosta-Cabronero J, Ernemann U, Fellgiebel A, Filippi M, Frisoni G, Hentschel F, Jessen F, Kloppel S, Meindl T, Pouwels PJW, Hauenstein KH, Hampel H (2011) Multicenter stability of diffusion tensor imaging measures: A European clinical and physical phantom study. *Psychiatry Res* **194**, 363-371.

[9] Frisoni GB, Boccardi M, Barkhof F, Blennow K, Cappa S, Chiotis K, Demonet JF, Garibotto V, Giannakopoulos

P, Gietl A, Hansson O, Herholz K, Jack CR, Jr., Nobili F, Nordberg A, Snyder HM, Ten Kate M, Varrone A, Albanese E, Becker S, Bossuyt P, Carrillo MC, Cerami C, Dubois B, Gallo V, Giacobini E, Gold G, Hurst S, Lonneborg A, Lovblad KO, Mattsson N, Molinuevo JL, Monsch AU, Mosimann U, Padovani A, Picco A, Porteri C, Ratib O, Saint-Aubert L, Scerri C, Scheltens P, Schott JM, Sonni I, Teipel S, Vineis P, Visser PJ, Yasui Y, Winblad B (2017) Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *Lancet Neurol* **16**, 661-676.

[10] Jessen F, Amariglio RE, van Boxtel M, Breteler M, Ceccaldi M, Chetelat G, Dubois B, Dufouil C, Ellis KA, van der Flier WM, Glodzik L, van Harten AC, de Leon MJ, McHugh P, Mielke MM, Molinuevo JL, Mosconi L, Osorio RS, Perrotin A, Petersen RC, Rabin LA, Rami L, Reisberg B, Rentz DM, Sachdev PS, de la Sayette V, Saykin AJ, Scheltens P, Shulman MB, Slavin MJ, Sperling RA, Stewart R, Uspenskaya O, Vellas B, Visser PJ, Wagner M, Subjective Cognitive Decline Initiative Working Group (2014) A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimers Dement* **10**, 844-852.

[11] Mitchell AJ, Beaumont H, Ferguson D, Yadegarfar M, Stubbs B (2014) Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: Meta-analysis. *Acta Psychiatr Scand* **130**, 439-451.

[12] Hong YJ, Kim CM, Jang EH, Hwang J, Roh JH, Lee JH (2016) White matter changes may precede gray matter loss in elderly with subjective memory impairment. *Dement Geriatr Cogn Disord* **42**, 227-235.

[13] Hong YJ, Yoon B, Shim YS, Ahn KJ, Yang DW, Lee JH (2015) Gray and white matter degenerations in subjective memory impairment: Comparisons with normal controls and mild cognitive impairment. *J Korean Med Sci* **30**, 1652-1658.

[14] Selnes P, Aarsland D, Bjornerud A, Gjerstad L, Wallin A, Hessen E, Reinvang I, Grambaite R, Auning E, Kjaervik VK, Due-Tonnessen P, Stenset V, Fladby T (2013) Diffusion tensor imaging surpasses cerebrospinal fluid as predictor of cognitive decline and medial temporal lobe atrophy in subjective cognitive impairment and mild cognitive impairment. *J Alzheimers Dis* **33**, 723-736.

[15] Doan NT, Engvig A, Persson K, Alnaes D, Kaufmann T, Rokicki J, Cordova-Palomera A, Moberget T, Braekhus A, Barca ML, Engedal K, Andreassen OA, Selbaek G, Westlye LT (2017) Dissociable diffusion MRI patterns of white matter microstructure and connectivity in Alzheimer's disease spectrum. *Sci Rep* **7**, 45131.

[16] Shao W, Li X, Zhang J, Yang C, Tao W, Zhang S, Zhang Z, Peng D (2019) White matter integrity disruption in the predementia stages of Alzheimer's disease: From subjective memory impairment to amnestic mild cognitive impairment. *Eur J Neurol* **26**, 800-807.

[17] Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Contributors (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535-562.

[18] Jessen F, Spottke A, Boecker H, Brosseron F, Buerger K, Catak C, Fliessbach K, Franke C, Fuentes M, Heneka MT, Janowitz D, Kilimann I, Laske C, Menne F, Nestor P, Peters O, Priller J, Pross V, Ramirez A, Schneider A, Speck O, Spruth EJ, Teipel S, Vukovich R, Westerteicher C, Wiltfang J, Wolfsgruber S, Wagner M, Duzel E (2018) Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). *Alzheimers Res Ther* **10**, 15.

[19] Gauggel S, Birkner B (1999) Validity and reliability of a German version of the Geriatric Depression Scale (GDS). *Z Klin Psychol Forsch Praxis* **28**, 18-27.

[20] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) 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* **7**, 270-279.

[21] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr., Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia 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* **7**, 263-269.

[22] Acosta-Cabronero J, Williams GB, Pengas G, Nestor PJ (2010) Absolute diffusivities define the landscape of white matter degeneration in Alzheimer's disease. *Brain* **133**, 529-539.

[23] Oishi K, Mielke MM, Albert M, Lyketsos CG, Mori S (2011) DTI analyses and clinical applications in Alzheimer's disease. *J Alzheimers Dis* **26**(**Suppl 3**), 287-296.

[24] Stricker NH, Schweinsburg BC, Delano-Wood L, Wierenga CE, Bangen KJ, Haaland KY, Frank LR, Salmon DP, Bondi MW (2009) Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis. *Neuroimage* **45**, 10-16.

[25] Smith SM, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols TE, Mackay CE, Watkins KE, Ciccarelli O, Cader MZ, Matthews PM, Behrens TE (2006) Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *Neuroimage* **31**, 1487-1505.

[26] Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825-841.

[27] Smith S (2002) Fast robust automated brain extraction. *Hum Brain Mapp* **17**, 143-155.

[28] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[29] Frisoni GB, Jack CR, Jr., Bocchetta M, Bauer C, Frederiksen KS, Liu Y, Preboske G, Swihart T, Blair M, Cavedo E, Grothe MJ, Lanfredi M, Martinez O, Nishikawa M, Portegies M, Stoub T, Ward C, Apostolova LG, Ganzola R, Wolf D, Barkhof F, Bartzokis G, DeCarli C, Csernansky JG, deToledo-Morrell L, Geerlings MI, Kaye J, Killiany RJ, Lehericy S, Matsuda H, O'Brien J, Silbert LC, Scheltens P, Soininen H, Teipel S, Waldemar G, Fellgiebel A, Barnes J, Firbank M, Gerritsen L, Henneman W, Malykhin N, Pruessner JC, Wang L, Watson C, Wolf H, deLeon M, Pantel J, Ferrari C, Bosco P, Pasqualetti P, Duchesne S, Duvernoy

H, Boccardi M, EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative (2015) The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimers Dement* **11**, 111-125.

[30] Wolf D, Bocchetta M, Preboske GM, Boccardi M, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative (2017) Reference standard space hippocampus labels according to the European Alzheimer's Disease Consortium-Alzheimer's Disease Neuroimaging Initiative harmonized protocol: Utility in automated volumetry. *Alzheimers Dement* **13**, 893-902.

[31] Ashburner J (2007) A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95-113.

[32] Janelidze S, Zetterberg H, Mattsson N, Palmqvist S, Vanderstichele H, Lindberg O, van Westen D, Stomrud E, Minthon L, Blennow K, Swedish BioFINDER study group, Hansson O (2016) CSF Abeta42/Abeta40 and Abeta42/Abeta38 ratios: Better diagnostic markers of Alzheimer disease. *Ann Clin Transl Neurol* **3**, 154-165.

[33] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* **67**, 301-320.

[34] Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Kloppel S, Borchardt V, Walter M, Dyrba M (2016) Robust detection of impaired resting state functional connectivity networks in Alzheimer's disease using elastic net regularized regression. *Front Aging Neurosci* **8**, 318.

[35] Mori S, Oishi K, Jiang H, Jiang L, Li X, Akhter K, Hua K, Faria AV, Mahmood A, Woods R, Toga AW, Pike GB, Neto PR, Evans A, Zhang J, Huang H, Miller MI, van Zijl P, Mazziotta J (2008) Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *Neuroimage* **40**, 570–582.

[36] Belsley DA (1991) *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley & Sons, Chichester.

[37] Cho H, Yang DW, Shon YM, Kim BS, Kim YI, Choi YB, Lee KS, Shim YS, Yoon B, Kim W, Ahn KJ (2008) Abnormal integrity of corticocortical tracts in mild cognitive impairment: A diffusion tensor imaging study. *J Korean Med Sci* **23**, 477-483.

[38] Alves GS, O'Dwyer L, Jurcoane A, Oertel-Knochel V, Knochel C, Prvulovic D, Sudo F, Alves CE, Valente L, Moreira D, Fusser F, Karakaya T, Pantel J, Engelhardt E, Laks J (2012) Different patterns of white matter degeneration using multiple diffusion indices and volumetric data in mild cognitive impairment and Alzheimer patients. *PLoS One* **7**, e52859.

[39] Clerx L, Visser PJ, Verhey F, Aalten P (2012) New MRI markers for Alzheimer's disease: A meta-analysis of diffusion tensor imaging and a comparison with medial temporal lobe measurements. *J Alzheimers Dis* **29**, 405-429.

[40] Liu Y, Spulber G, Lehtimaki KK, Kononen M, Hallikainen I, Grohn H, Kivipelto M, Hallikainen M, Vanninen R, Soininen H (2011) Diffusion tensor imaging and tract-based spatial statistics in Alzheimer's disease and mild cognitive impairment. *Neurobiol Aging* **32**, 1558-1571.

[41] Lee SH, Coutu JP, Wilkens P, Yendiki A, Rosas HD, Salat DH, Alzheimer's disease Neuroimaging Initiative (2015) Tract-based analysis of white matter degeneration in Alzheimer's disease. *Neuroscience* **301**, 79-89.

[42] Winklewski PJ, Sabisz A, Naumczyk P, Jodzio K, Szurowska E, Szarmach A (2018) Understanding the physiopathology behind axial and radial diffusivity changes-what do we know? *Front Neurol* **9**, 92.

[43] Zhuang L, Wen W, Zhu W, Trollor J, Kochan N, Crawford J, Reppermund S, Brodaty H, Sachdev P (2010) White matter integrity in mild cognitive impairment: A tract-based spatial statistics study. *Neuroimage* **53**, 16-25.

[44] Ennis DB, Kindlmann G (2006) Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magn Reson Med* **55**, 136-146.

[45] Teipel SJ, Grothe MJ, Filippi M, Fellgiebel A, Dyrba M, Frisoni GB, Meindl T, Bokde AL, Hampel H, Kloppel S, Hauenstein K; EDSD study group (2014) Fractional anisotropy changes in Alzheimer's disease depend on the underlying fiber tract architecture: A multiparametric DTI study using joint independent component analysis. *J Alzheimers Dis* **41**, 69-83.

[46] Li XY, Tang ZC, Sun Y, Tian J, Liu ZY, Han Y (2016) White matter degeneration in subjective cognitive decline: A diffusion tensor imaging study. *Oncotarget* **7**, 54405-54414.

[47] Ryu SY, Lim EY, Na S, Shim YS, Cho JH, Yoon B, Hong YJ, Yang DW (2017) Hippocampal and entorhinal structures in subjective memory impairment: A combined MRI volumetric and DTI study. *Int Psychogeriatr* **29**, 785-792.

[48] Selnes P, Fjell AM, Gjerstad L, Bjornerud A, Wallin A, Due-Tonnessen P, Grambaite R, Stenset V, Fladby T (2012) White matter imaging changes in subjective and mild cognitive impairment. *Alzheimers Dement* **8**, S112-121.

[49] Kiuchi K, Kitamura S, Taoka T, Yasuno F, Tanimura M, Matsuoka K, Ikawa D, Toritsuka M, Hashimoto K, Makinodan M, Kosaka J, Morikawa M, Kichikawa K, Kishimoto T (2014) Gray and white matter changes in subjective cognitive impairment, amnestic mild cognitive impairment and Alzheimer's disease: A voxel-based analysis study. *PLoS One* **9**, e104007.

[50] Wang Y, West JD, Flashman LA, Wishart HA, Santulli RB, Rabin LA, Pare N, Arfanakis K, Saykin AJ (2012) Selective changes in white matter integrity in MCI and older adults with cognitive complaints. *Biochim Biophys Acta* **1822**, 423-430.

[51] Sargent-Cox K, Cherbuin N, Sachdev P, Anstey KJ (2011) Subjective health and memory predictors of mild cognitive disorders and cognitive decline in ageing: The Personality and Total Health (PATH) through Life Study. *Dement Geriatr Cogn Disord* **31**, 45-52.

[52] Chen G, Guo Y, Zhu H, Kuang W, Bi F, Ai H, Gu Z, Huang X, Lui S, Gong Q (2017) Intrinsic disruption of white matter microarchitecture in first-episode, drug-naive major depressive disorder: A voxel-based meta-analysis of diffusion tensor imaging. *Prog Neuropsychopharmacol Biol Psychiatry* **76**, 179-187.

[53] Teipel SJ, Wegrzyn M, Meindl T, Frisoni G, Bokde AL, Fellgiebel A, Filippi M, Hampel H, Kloppel S, Hauenstein K, Ewers M (2012) Anatomical MRI and DTI in the diagnosis of Alzheimer's disease: A European multicenter study. *J Alzheimers Dis* **31**(**Suppl 3**), S33-47.

https://doi.org/10.1093/braincomms/fcac239 BRAIN COMMUNICATIONS 2022: Page 1 of 11 | 1

BRAIN COMMUNICATIONS

Cognitive and behavioural but not motor impairment increases brain age in amyotrophic lateral sclerosis

Andreas Hermann,[1,2,3] Gaël Nils Tarakdjian,[1,3] Anna Gesine Marie Temp,[1,3] Elisabeth Kasper,[4] Judith Machts,[5,6,7] Jörn Kaufmann,[8] Stefan Vielhaber,[7,8] Johannes Prudlo,[4] James H. Cole,[9,10] Stefan Teipel[3,11] and Martin Dyrba[3]

Age is the most important single risk factor of sporadic amyotrophic lateral sclerosis. Neuroimaging together with machine-learning algorithms allows estimating individuals' brain age. Deviations from normal brain-ageing trajectories (so called predicted brain age difference) were reported for a number of neuropsychiatric disorders. While all of them showed increased predicted brain-age difference, there is surprisingly few data yet on it in motor neurodegenerative diseases. In this observational study, we made use of previously trained algorithms of 3377 healthy individuals and derived predicted brain age differences from volumetric MRI scans of 112 amyotrophic lateral sclerosis patients and 70 healthy controls. We correlated predicted brain age difference scores with voxel-based morphometry data and multiple different motoric disease characteristics as well as cognitive/behavioural changes categorized according to Strong and Rascovsky. Against our primary hypothesis, there was no higher predicted brain-age difference in the amyotrophic lateral sclerosis patients as a group. None of the motoric phenotypes/characteristics influenced predicted brain-age difference. However, cognitive/behavioural impairment led to significantly increased predicted brain-age difference, while slowly progressive as well as cognitive/behavioural normal amyotrophic lateral sclerosis patients had even younger brain ages than healthy controls. Of note, the cognitive/behavioural normal amyotrophic lateral sclerosis patients were identified to have increased cerebellar brain volume as potential resilience factor. Younger brain age was associated with longer survival. Our results raise the question whether younger brain age in amyotrophic lateral sclerosis with only motor impairment provides a cerebral reserve against cognitive and/or behavioural impairment and faster disease progression. This new conclusion needs to be tested in subsequent samples. In addition, it will be interesting to test whether a potential effect of cerebral reserve is specific for amyotrophic lateral sclerosis or can also be found in other neurodegenerative diseases with primary motor impairment.

1 Translational Neurodegeneration Section "Albrecht Kossel", Department of Neurology, University Medical Center Rostock, University of Rostock, 18147 Rostock, Germany
2 Center for Transdisciplinary Neurosciences Rostock (CTNR), University Medical Center Rostock, University of Rostock, 18147 Rostock, Germany
3 Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) Rostock/Greifswald, 18147 Rostock, Germany
4 Department of Neurology, University Medical Center Rostock, University of Rostock, 18147 Rostock, Germany
5 Institute for Cognitive Neurology and Dementia Research, Otto-von-Guericke University Magdeburg, 39120 Magdeburg, Germany
6 Center for Behavioral Brain Sciences CBBS, 39104 Magdeburg, Germany
7 Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) Magdeburg, 39120 Magdeburg, Germany
8 Department of Neurology, Otto-von-Guericke University Magdeburg, 39120 Magdeburg, Germany
9 Centre for Medical Image Computing, Department of Computer Science, UCL, London, UK
10 Dementia Research Centre, Queen Square Institute of Neurology, UCL, London, UK
11 Department of Psychosomatic Medicine, University Medical Center Rostock, University of Rostock, 18147 Rostock, Germany

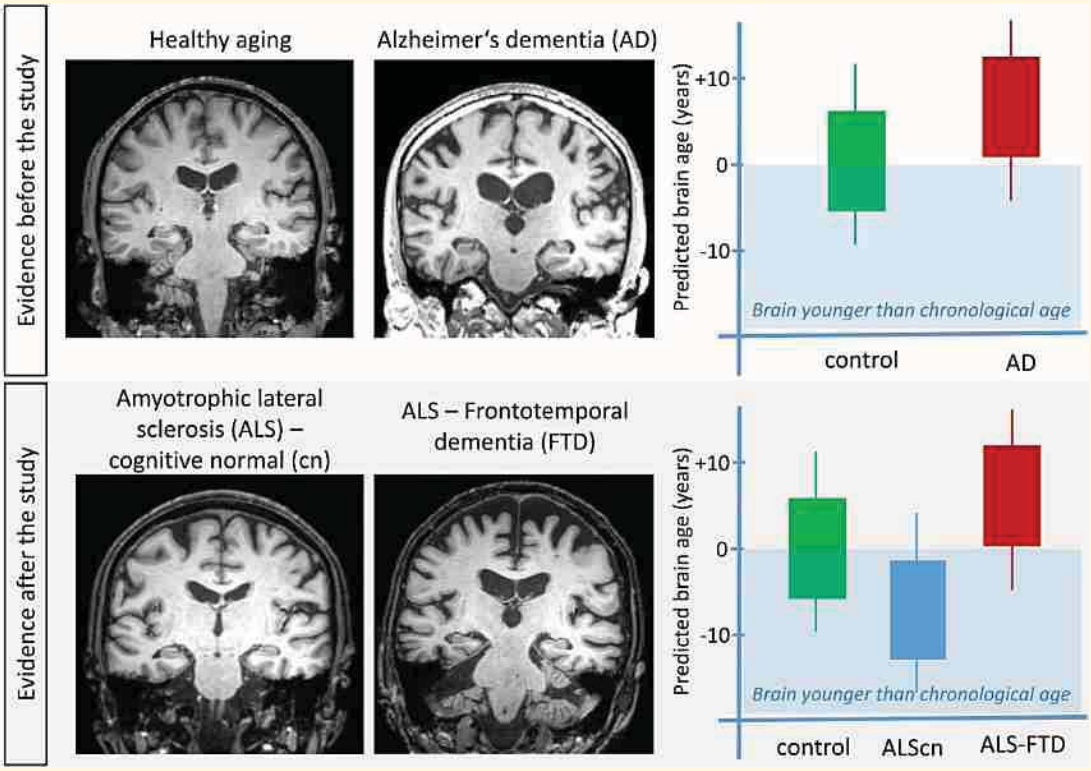
Received March 22, 2022. Revised July 01, 2022. Accepted September 21, 2022. Advance access publication September 22, 2022

Correspondence to: Andreas Hermann
Translational Neurodegeneration Section 'Albrecht Kossel'
Department of Neurology, University Medical Center Rostock
Gehlsheimer Straße 20, 18147 Rostock, Germany
E-mail: Andreas.Hermann@med.uni-rostock.de

Keywords: ageing; frontotemporal dementia; frontotemporal lobar degeneration; motor neurodegenerative diseases; cognitive reserve

Abbreviations: ALS = amyotrophic lateral sclerosis; ALSFRS-R = ALS functional rating scale revised; ALScn = ALS without cognitive/behavioural impairments ('cognitive normal'); ALSci = ALS with cognitive impairment; ALSbi = ALS with behavioural impairment; ALScbi = ALS with cognitive and behavioural impairments; ALS-FTD = ALS with frontotemporal dementia; HCs = healthy controls; LMN ALS = lower motor neuron predominant ALS; MoCA = Montreal cognitive assessment; PAD = predicted brain age difference; PMA = progressive muscular atrophy; UMN ALS = upper motor neuron predominant ALS; VBM = voxel-based morphometry

Graphical Abstract



Introduction

Amyotrophic lateral sclerosis (ALS) is the most common motor neuron disease. It is characterized by upper and lower motor neuron demise, leading to progressive paralysis and death within 1–5 years after symptom onset. On a group level, patients with ALS exhibit central nervous system involvement beyond the upper motor neuron system, including for example the frontotemporal lobes,[1-3] hypothalamus[4] and corpus callosum.[3] Several factors may modify risk and speed of disease progression, including the initial disease manifestation (bulbar versus spinal)[5] and the extend of frontotemporal impairment.[6] Cognitive and behavioural impairment accompany motor decline in over half of the patients with ALS over the course of the disease. The most common cognitive deficits in ALS concern executive functions, especially verbal fluency.[7] The revised consensus criteria of frontotemporal dysfunction in ALS by Strong *et al.*[8] and by Rascovsky *et al.*[9] in the case of frontotemporal dementia (FTD) patients with ALS characterize patients according to their cognitive deficits, extent of behavioural impairment and presence or absence of FTD. Notably, cognitive and behavioural impairments predict shorter survival time.[6]

The single most relevant risk factor for sporadic ALS is age, with the highest prevalence of disease in patients over 60 years of age.[10] This points towards an important role of

the ageing process itself. Normal ageing is a process of gradual accumulation of pathologies associated with cognitive and physical decline, which also affects brain volume and nerve cell loss.[11,12] Thus, ALS might be considered as an increased ageing process of specific brain systems.

Neuroimaging data combined with machine-learning techniques can be used to predict the age of a healthy individual's brain and allow measuring a potential deviance of an individual's predicted brain age from chronologic age, termed 'the predicted brain-age difference' (PAD).[13] This approach has been successfully applied not only in both early brain development and ageing in the healthy elderly (for a review see Franke and Gaser[13]), but also in a number of disease conditions, including Alzheimer's disease, schizophrenia, major depression, multiple sclerosis[14] and epilepsy (for systematic review see Cole *et al.*[15] and Wrigglesworth *et al.*[16]). All disease conditions exhibited a remarkably increased PAD score, indicating that brain atrophy exceeded normal brain ageing. This was also true when investigating effects of known cardiovascular risk factors on brain ageing.[17] In addition, increased PAD scores correlated well with increased mortality or decreased survival in a range of different conditions.[15,18]

To date, there are only few reports on classical 'motor' neurodegenerative disorders, such as Parkinson's disease and none on ALS. Hence, we chose ALS as a paradigmatic motor neurodegenerative disorder, since it has the advantage of clear definitions of motoric and cognitive/behavioural impairment. We hypothesized that patients with ALS would show increased brain age compared with healthy controls (HCs), and that this ageing process would be more pronounced in the presence of additional cognitive and/or behaviour impairment.

Materials and methods

Design

This two-centre prospective, observational cross-sectional study was conducted between April 2011 and August 2013. Local ethics committees of both universities approved the study (Rostock: A 2011 56; Magdeburg: 75/11) and all subjects gave written informed consent prior to their inclusion.

General methods used in this study have been already published[19] and respective details are reported in the supplementary material. Specific methods relevant to the estimation of brain-age algorithm are described here.

Participants

We recruited 182 German participants in Rostock and Magdeburg, Germany. Persons with a history of brain injury, epilepsy or psychiatric illness were excluded. Control participants were screened for cognitive impairment using the Montreal cognitive assessment (MoCA) with a cut off of ≤26/30. Seventy HCs and 112 patients diagnosed with ALS according to Swinnen and Robberecht[20] were included (Supplementary Fig. 1). These cases were characterized into ALS without cognitive/behavioural impairments (ALScn), ALS with cognitive impairment (ALSci), ALS with behaviour impairment (ALSbi), ALS with cognitive and behavioural impairments (ALScbi) and ALS with FTD (ALS-FTD) following the Strong and Rascovsky criteria.[8,9] Different motoric phenotypes of ALS were classified as classical ALS, upper/lower motor neuron predominant (UMN/LMN) ALS, flail arm, flail leg and progressive muscular atrophy (PMA). None of the patient presented with pure primary lateral sclerosis (PLS). Demographic details are summarized in Table 1.

Clinical and neuropsychologic measures

Clinical and neuropsychologic measures were reported previously[19] and thus reported in the Supplementary material. Examinations were done at the respective recruitment side.

MRI acquisition and processing

MRI scanning was performed with two 3 T Siemens Magnetom VERIO scanners (Erlangen, Germany) using a 32-channel head coil; one single scanner at each site (Rostock and Magdeburg, Germany). The anatomical T_1-weighted images were segmented into grey matter, white matter and cerebrospinal fluid partitions using the SPM12 toolbox in Matlab 2019a. Then, the *Diffeomorphic Anatomical Registration Through Exponentiated Lie* (DARTEL) algebra algorithm[21] was used in combination with a custom brainAgeR brain template to normalize the T_1-weighted images to the *Montreal Neurological Institute* (MNI) reference coordinate system. The estimated deformation fields were subsequently applied to the grey-matter segments to bring them in MNI space as well, followed by modulation to preserve the total amount of grey matter and smoothing with an 8 mm Gaussian kernel for the voxel-based morphometry (VBM) analysis.

Brain-age model and PAD

We estimated brain age in R, using the package 'brainAgeR', available at https://github.com/james-cole/brainageR. This algorithm was trained on $n = 3377$ healthy individuals and validated on 857 people. To predict brain age, we followed an automated pipeline starting with T_1-weighted image segmentation and normalization using SPM12 with smoothing with a 4 mm Gaussian kernel to match with the training sample. Then, the spatially normalized grey- and white-matter segments as well as cerebrospinal fluid segments were loaded into R. They were masked to exclude voxels with <30% probability for cerebrospinal fluid, white matter or grey matter, respectively. Subsequently, these segments were vectorized to apply a principal component transformation. The transformed data were then entered in the pretrained Gaussian progress regression model to obtain the predicted

Table 1 Demographic Background of the Participants

	HC ($N = 70$)	ALScn ($N = 58$)	ALSci ($N = 29$)	ALSbi ($N = 12$)	ALScbi ($N = 5$)	ALS-FTD ($N = 8$)	BF_{01} between all ALS & HC
Sex (f/m; %)	40/60	38/62	34/66	33/67	20/80	38/62	4.57
Age at examination	61.00 (10.67)	59.94 (9.74)	62.13 (11.29)	57.40 (12.01)	65.67 (14.62)	61.21 (10.38)	$5.88 \pm 1.37e-5\%$
Education (years)	13.36 (1.62)	13.48 (2.64)	11.86 (1.57)	13.83 (1.99)	11.60 (1.52)	13.00 (2.20)	$3.24 \pm 8.90e-6\%$
MoCA	27.50 (1.29)[a]	25.90 (2.45)	23.6 (3.57)[b]	26.33 (3.42)	21.00 (3.94)[c]	19.13 (5.49)[a]	$5.52 \pm 8.76e-6\%$
Disease duration until TPI (months)		29.19 (38.76)	30.69 (50.68)[b]	32.00 (30.21)[b]	14.60 (11.55)	12.38 (6.59)	
Total disease duration (months)		63.59 (53.64)	42.81 (33.80)	43.89 (29.02)	30.50 (32.03)	31.75 (18.74)	
Age at onset (years)		57.25 (10.07)	58.35 (12.36)	54.25 (13.23)	63.78 (14.74)	59.44 (10.36)	
EL Escorial criteria at test time (possible/probable/definitive/unknown; %)		38/20/14/28	41/31/14/14	17/42/33/8	20/60/0/20	38/38/25/0	
Onset site (bulbar/spinal/no data; %)	-	31/51.7/17.2	37.9/44.8/17.2	50/50/0	40/40/20	75/25/0	
ALS-FRSR (as close as possible to test time)		39.00 (5.93)	38.07 (6.96)	34.82(4.98)	36.80(5.54)	41.57(3.87)	
Δ ALS-FRSR (as close as possible to test time)		0.65(0.73)	1.40(1.82)[b]	1.12(1.20)	0.91(1.17)	0 .99(0.95)	
Delta ALS-FRSR (at diagnosis)		0.49 (0.40)	0.57 (0.40)	0.89 (0.92)	1.72 (1.67)[a]	0.45 (0.40)	

Matching took place between HC and patients with ALS as a whole. Sex, age and education were matched successfully: independent Student's *t*-tests supported the absence of differences in age and education, and a χ^2 test supported the absence of differences in sex distribution. Depicted are mean and SD if not mentioned differentially. [a]$BF_{10} > 100$ in favour of differences to ALSni group. [b]$BF_{10} > 3$ in favour of differences to ALSni group > 3. [c]$BF_{10} > 10$ in favour of differences to ALSni group.

brain age. Finally, the predicted age was subtracted from the chronologic age to calculate the PAD. While a positive PAD indicates an older appearing brain, a negative score suggests a younger appearing brain.

Voxel-based analysis of group differences

Complementarily, we performed a whole-brain VBM analysis for which the normalized and smoothed grey-matter maps were analysed using Statistical Parametric Mapping (SPM12; http://www.fil.ion.ucl.ac.uk/spm). All voxel-based analyses were controlled for total intracranial volume, chronologic age, sex and site, as these were potential nuisance variables. The statistical threshold for the analyses was set to an uncorrected $P < 0.001$ and only clusters with at least 50 voxels extent were retained in the results.

Statistical analysis

As classical null hypothesis significance testing only enables us to reject the null hypothesis that there are no effects of clinical presentation on PAD, we opted for *Bayes factor hypothesis testing* (BFHT) using an analysis of covariance (ANCOVA). This Bayesian approach allows for the estimation of the likelihood of such effects given the observed data and, hence, more directly infer and compare the actual effects. Specifically, we compared the effects of Strong profile, progressor type, phenotype, onset type, disease duration until MRI scanning and age at disease onset, while controlling for age at MRI, sex and recruitment location by adding them to the null model. We conducted one multi-factorial ANCOVA which compared all these effects against one another, and against the corrected null hypothesis model. A priori, we assumed all models to be equally likely. We applied default Jeffreys-Zellner-Siow priors, with the seed set to 84 293. Please see Table 2 for a summary of the statistical measures we will be reporting. All Bayesian analyses were conducted in *Jeffreys's Amazing Statistics Program* (JASP, 0.14.3). JASP was set to report the corrected null model on top, and to compare all other models against it using BF_{10}. Bayes factors do not require thresholding akin to $P < 0.05$ to determine statistical significance: instead they fall on a continuum ranging from support for the null hypothesis via no support for either hypothesis to support for the alternative hypothesis.[22] Additionally, we can add qualitative descriptors by stating that $BF_{10} > 100$ constitutes 'extreme evidence' for H_1, $BF_{10} > 30$ constitutes 'very strong' evidence for H_1, $BF_{10} > 10$ constitutes 'strong' evidence for H_1 and $BF_{10} > 3$ constitutes 'moderate' support for H_1.

Data availability

The original, individual MRI files are not available due to participant confidentiality and privacy concerns. The brainAgeR toolbox is freely available at https://github.com/james-cole/brainageR. The PAD score information was

Table 2 Statistical Measures in Bayesian Probability

Notation/ Abbreviation	Full Name	Interpretation
Prior	Prior distribution	Distribution of the effect size, as assumed prior to data collection/analysis
Posterior	Posterior distribution	Actual distribution of the effect size after the data at hand have been analysed
P(M)	Prior model probability	Probability of this particular statistical model being supported by the data at hand, as assumed prior to data collection/analysis
P(M|data)	Posterior model probability	Posterior probability of this particular model being supported by the data at hand, after they have been analysed
BF	Bayes factor	The strength of evidence in favour of a given statistical model, relative to another statistical model (see below)
BF_{01}	Bayes factor 0/1	The strength of evidence in favour of Model 0, relative to Model 1
BF_{10}	Bayes factor 1/0	The strength of evidence in favour of Model 1, relative to Model 0
$BF_{10} > 100$		'Extreme evidence' favouring Model 1, relative to Model 0
$BF_{10} > 30$		'Very strong evidence' favouring Model 1, relative to Model 0
$BF_{10} > 10$		'Strong evidence' favouring Model 1, relative to Model 0
$BF_{10} > 3$		'Moderate evidence' favouring Model 1, relative to Model 0
$BF_{10} = 1$		Model 1 and Model 0 are equally supported by the evidence
$BF_{10} < 0.33$		'Moderate evidence' against Model 1, relative to Model 0 (equivalent to $BF_{01} > 3$)
$BF_{10} < 0.10$		'Strong evidence' against Model 1, relative to Model 0 (equivalent to $BF_{01} > 10$)
$BF_{10} < 0.03$		'Very strong evidence' against Model 1, relative to Model 0 (equivalent to $BF_{01} > 30$)
$BF_{10} < 0.01$		'Extreme evidence' against Model 1, relative to Model 0 (equivalent to $BF_{01} > 100$)
Error%	Stability of the BF	The range of the BF over the chosen Markov chain Monte Carlo iterations, e.g. $BF_{10} = 10$ with error% = 20 means that the BF_{10} ranged from 8 to 12
95% CI	Credible interval	With 95% certainty, the true effect size lies within these bounds

extracted and included in a.csv file, alongside necessary clinical information. These data supporting the BFHT and Figs 1–3 are publicly available from: https://osf.io/fyt7d/, alongside a JASP analysis file, an HTML results file and the R code supporting the figure generation. The MRI data supporting Fig. 4 are not publicly available.

Results

Patient cohort

The patients' onset types included bulbar ($n = 41$), spinal ($n = 53$) or uncertain ($n = 18$). Phenotypically, the patients presented with classical ($n = 68$), upper motor neuron dominant ($n = 12$), flail arm ($n = 6$), flail leg ($n = 6$) or other ($n = 18$) ALS. According to the El Escorial criteria, 40/29/21 patients had a possible/probable/definitive ALS, but 22 patients exhibited pure upper or lower motor neuron involvement and thus did not meet the El Escorial criteria. Patient classification according to the Strong and Rascovsky criteria indicated that most patients were profiled as ALScn (ALS with no cognitive or behavioural impairments, $n = 58$), alongside 29 ALSci, 12 ALSbi, 5 ALScbi and $n = 8$ ALS-FTD patients. All patients underwent genetic testing, with four cases with mutations in *C9ORF72*, three cases of *superoxide dismutase 1* (SOD1), one case of *vesicle-associated membrane protein-associated protein B/C* (VAPB), one case of a juvenile ALS with *senataxin* (SETX) mutation and an uncertain familial link emerging. The remaining patients had sporadic ALS. Demographic and clinical characteristics of the study populations are shown in Table 1. The recruitment flow is shown in Supplementary Fig. 1.

Predicted brain age was only increased in cognitively/behaviourally impaired patients with ALS

Prior to analysing possible disease effects, we checked that the pretrained brain-age model included in the brainAgeR software was appropriate for our dataset, as it was established on 3377 independent healthy people. For this, we evaluated the PAD (i.e. calculated brain age—chronologic age at time point of MRI) in our HC cohort. As shown in Fig. 1A, the control cohort revealed a (perfectly) matching PAD score of -1.30 ± 6.00 years (mean ± SD) with homogeneous variability across the age range.

We first investigated whether PAD differed between HCs and patients with ALS in general. Surprisingly, patients with ALS in general did not show increased brain age (Fig. 1A). The hypothesis that PAD score of -1.06 ± 7.14 years (mean ± SD) in patients with ALS did not differ from HCs was six times more plausible than the hypothesis that HC and patients would differ (Bayesian independent samples *t*-test, $BF_{01} = 5.92$, error% = 1.380e − 5). This constitutes moderate evidence that a difference between HC and patients with ALS is absent.

Next, we investigated whether cognitive and/or behavioural impairment influenced brain ageing. Here, we observed moderate to extreme evidence favouring the influence of cognitive/behavioural impairment: the strength of the evidence fluctuated by the severity of impairment. The main effect of Strong profile was 524 times more plausible than the hypothesis that fluctuations of PAD score were driven by age, sex or recruitment location (Table 3, $BF_{10} = 524.74$, considered 'extreme evidence'). The ALSci and ALS-FTD patients showed significantly greater brain age

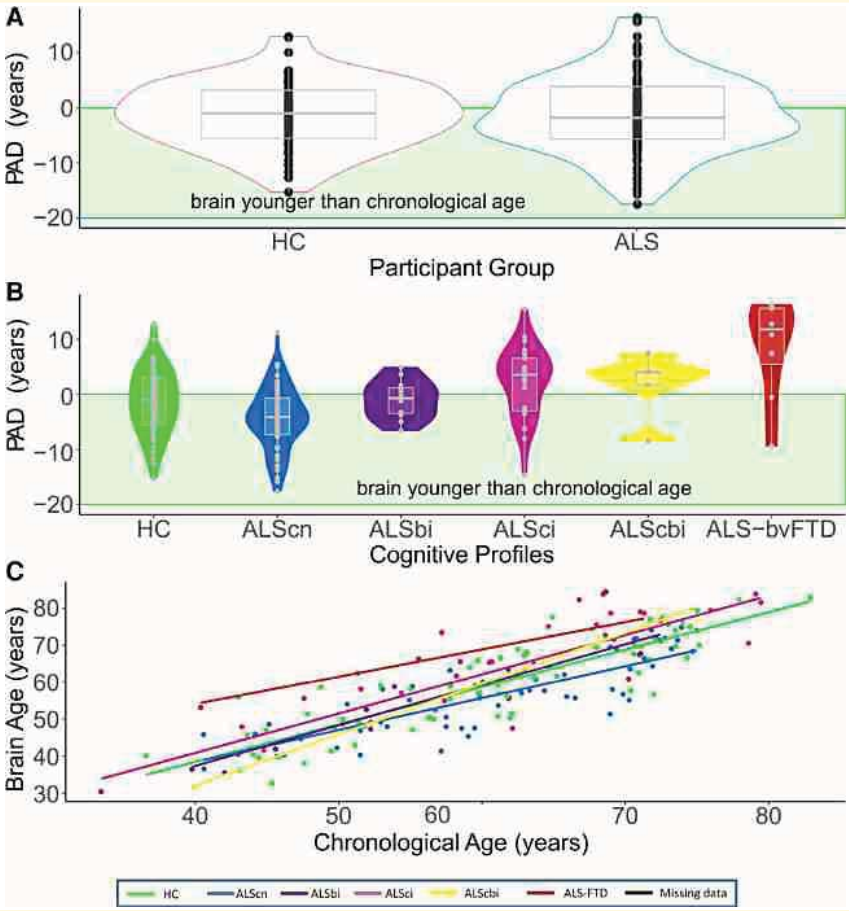


Figure 1 Predicted brain age difference (PAD) is increased in cognitively/behaviourally impaired patients with ALS. (**A**) The multivariate model predicted brain age accurately in our healthy controls (HCs). There was no difference in PAD in patients with ALS *per se* (Bayesian independent samples *t*-test, $BF_{01} = 5.92$, error% = 1.380e − 5, favouring the absence of differences). (**B**) Cognitive/behavioural impairment increased PAD score significantly (ANCOVA main effect, $BF_{10} = 524.74$), while ALScn patients showed significant decreased PAD (ANCOVA *post hoc* test, $BF_{10} = 7.71$ in favour of this difference). (**C**) Chronologic age and predicted brain age correlated strongly and had a very narrow credible interval, suggesting a homogeneous, reliable effect (Pearson's rho for the overall cohort = 0.85, with a 95% credible interval from 0.80 to 0.88; $BF_{10} = 2.19e + 48$).

compared with ALScn patients and HC (Fig. 1B). ALS-FTD patients' brains exhibited strong to extreme evidence for greater added ageing (8.58 ± 9.18 years; mean ± SD), compared with the HC ($BF_{10} = 221.80$, 'extreme evidence'), ALScn ($BF_{10} = 12\,918.68$, 'extreme evidence') and ALSbi groups ($BF_{10} = 10.03$, strong evidence). ALSci patients' brains had second highest PAD (2.27 ± 6.40 years, mean ± SD, Fig. 1B); evidence was modest that this was more pronounced than in the HC's brains ($BF_{10} = 4.55$), and extremely strong evidence compared with ALScn patients' brains ($BF_{10} = 2735.58$). The effects in our data were not strong enough to provide sufficient evidence for or against differences between the ALScbi and ALSbi groups, possibly because these groups exhibited heterogenous effects on PAD as reflected by their large credible intervals including zero (see Supplementary materials). Unexpectedly, we found modest evidence that ALScn patients' brain age was moderately lower than those of the HC group (−4.33 ± 5.79 years, mean ± SD, $BF_{10} = 7.71$). In our data, the hypothesis that ALScn patients have younger appearing brains was seven times more likely than the absence of any differences. Predicted brain age correlated well with chronologic brain age in HCs, in ALScn and ALS-impaired (ci, bi, cbi and FTD; Pearson's rho between 0.66 and 0.99, see Fig. 1C).

The above provides compelling and novel evidence that brain age increases at different speeds across different clinical subgroups of ALS, and that brain age is associated with survival time. We re-ran the above analyses while excluding the 22 patients whose ALS did not meet the El Escorial criteria, and those who had genetic variants of the disease. This did not fundamentally affect the above conclusions with one exception: the difference between ALScn and HC prevailed when either uncertain El Escorial types were excluded ($BF_{10} = 7.71$; Supplementary Fig. 2), or when genetic variants were excluded ($BF_{10} = 7.30$; Supplementary Fig. 3). Therefore, we did not exclude those for further analysis,

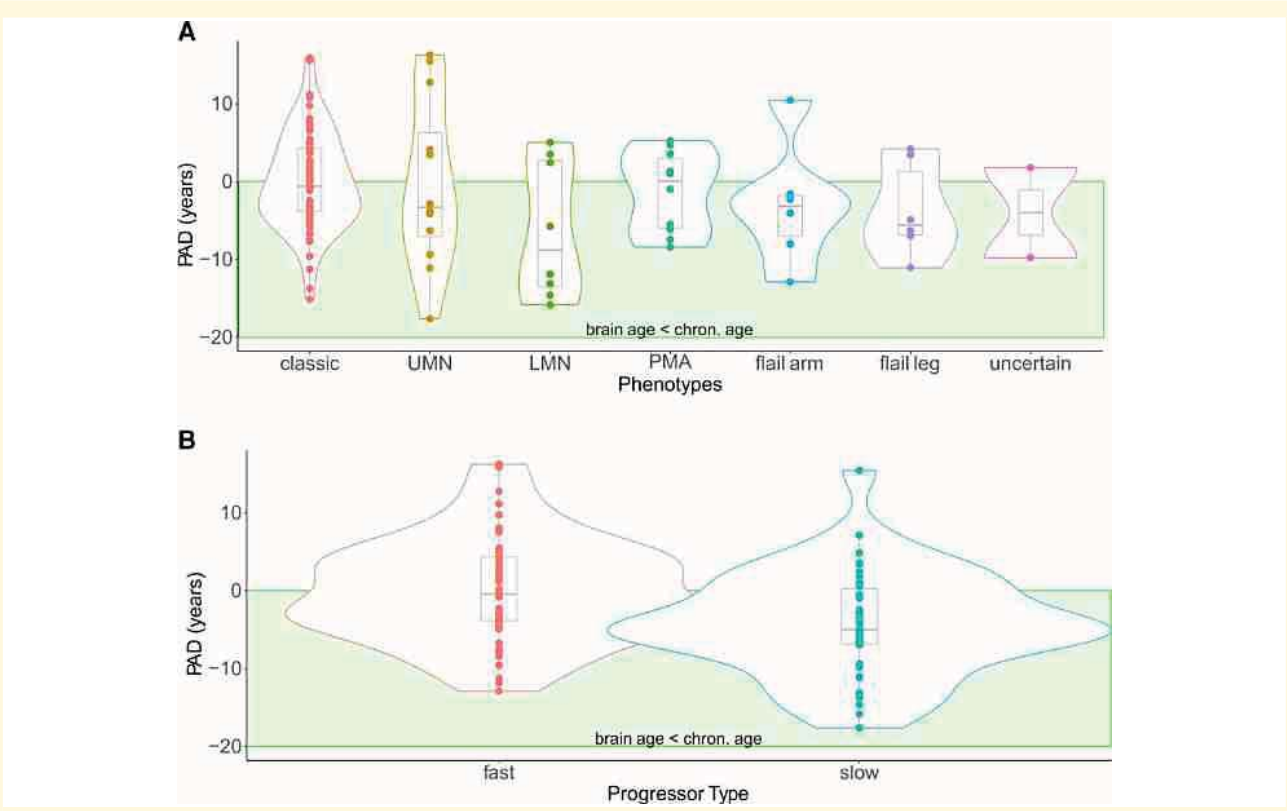


Figure 2 Predicted brain age is not influenced by motor subtypes but by disease progression rate. (**A**) Classical motor subtypes did not influence PAD (ANCOVA prior model probability $P(M) = 2\%$ was reduced to $P(M|data) < 0.0001\%$ a posteriori). (**B**) The comparison of slow (Δ ALSFRS-R <0.5) versus fast disease progression (Δ ALSFRS-R ≥0.5)—measured by (48-current ALSFRS-R score)/months since disease onset—yielded moderate evidence favouring a main effect (ANCOVA, $BF_{10} = 5.52$; *post hoc* directional informed ANCOVA BF = 262.61).

being most likely more representative for typical clinical settings.

Predicted brain age was not influenced by motor subtypes but by disease progression rate

Different motoric phenotypes of ALS—classical ALS, UMN/LMN ALS, flail arm, flail leg, PMA—did not exhibit differences in brain ageing (Fig. 2A): motoric phenotype effect decreased in plausibility from the prior model probability $P(M) = 2\%$ to a posterior model probability $P(M|data) < 0.0001\%$. Consequently, the Bayesian analysis of effects demonstrated that models excluding the motoric phenotype variable were four times better than those including phenotype ($BF_{excl} = 4.60$), and models excluding disease onset site were five times better than those including disease onset site ($BF_{excl} = 4.94$). The plausibility of onset site's effect on brain ageing also decreased from 1.6% to below 0.0001% (Supplementary Fig. 4A). In summary, clinicomotoric aspects of ALS did not affect brain ageing. Our data further supported the absence of correlations between increased brain ageing and age at disease onset (Pearson's $r = 0.03$ with a 95% credible interval ranging from −0.155 to 0.212, $BF_{01} = 8.06$), and disease duration until the time point of MRI investigation (Pearson's $r = -0.127$ with a 95% credible interval between −0.301 and 0.060, $BF_{01} = 3.54$; Supplementary Fig. 5).

We next asked whether upper motor neuron involvement was the key driver of increased PAD score, so we grouped the PMA and LMN groups (including flail-arm and flail-leg syndrome) and compared them with all others. However, the evidence regarding a potential effect of upper motor neuron involvement was inconclusive: our data decreased the effect's plausibility by a factor of 10^4 but at $BF_{10} = 0.77$, no hypothesis was preferable to the other (Supplementary Fig. 4B).

Several clinical studies reported differential therapeutic effects in rapid versus slow-progressing patients with ALS suggesting that rapid disease progression might represent a distinct disease type. Thus, we directionally hypothesized that fast progressors—dichotomized by a monthly decline of ALS functional rating scale revised (ALSFRS-R) ≥0.5—would exhibit increased brain age. There was moderate evidence favouring the main effect of Progressor type (Table 3, $BF_{10} = 5.52$) which we followed up with an informed Bayesian ANOVA to specifically test our directional hypothesis. It was 262.61 times more plausible than the effects of

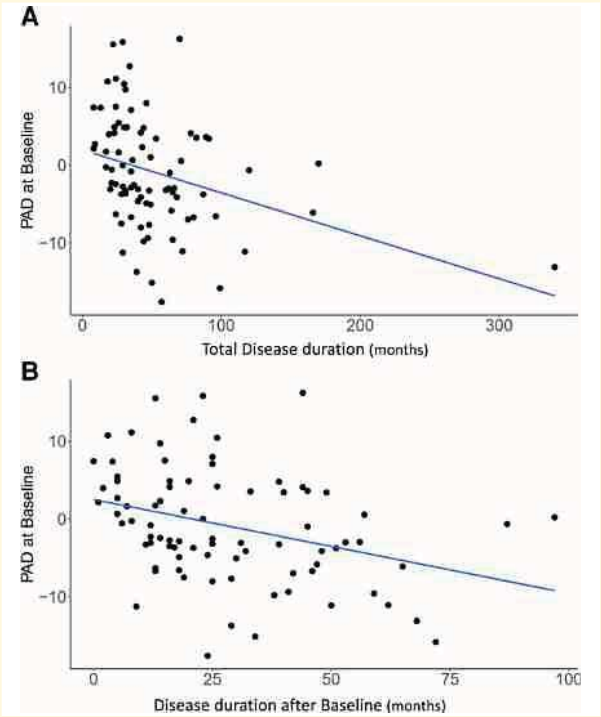


Figure 3 Predicted brain age is a prognostic marker. PAD score negatively correlated with total disease duration (**A**, Kendall's tau = −0.291 with a credible interval from −0.423 to −0.139, $BF_{10} = 250.206$) and disease duration after baseline (=time point of MRI) (**B**, Kendall's tau = −0.272 with a 95% credible interval of −0.405 to −0.120, $BF_{10} = 96.94$).

chronologic age, sex and recruitment location alone. As shown in Fig. 2B, slow progressors' brains were younger than their chronologic age (−4.25 ± 6.75 years). Of note, lower PAD in slow progressors was independent of their Strong profiles (Supplementary Fig. 4C).

The hypothesis that slow-progressing ALScn patients had younger brains than HC was 62 times more plausible than the absence of an effect ($BF_{+0} = 62.45$, Supplentary Fig. 4D). This suggests that these patients might possess the strongest resilience factors protecting them from cognitive/behavioural impairment or from brain atrophy/ageing.

Cognitive/behavioural impairment and disease progression were independent but additive predictors of PAD

A posteriori, the most plausible effects in our data were the co-occurring but independent main effects of Strong profile and progressor type: they increased in probability from $P(M) = 1.6\%$ to $P(M|data) = 24\%$ and were nearly 5000 more likely than the sole influences of chronologic age, sex and recruitment location ($P(M) = 0.016$, $P(M|data) = 0.240$, $BF_{10} = 4803.70$, error% = 3.57). This ANCOVA was also able to discriminate between suitable and unsuitable predictors. Models containing Strong profile were 579 times better than those without this predictor ($P(incl) = 0.500$, $P(incl|data) = 0.998$, $BF_{incl} = 579.20$), and models containing progressor type were seven times better than models without it ($P(incl) = 0.500$, $P(incl|data) = 0.884$, $BF_{incl} = 7.31$) when it came to explaining PAD scores. This informs us that Strong profile was the more plausible predictor for brain ageing but both were independently relevant (Table 3).

PAD correlated with survival time

We next investigated predictive power of brain ageing on disease duration/survival. Survival data were available for 83 patients. Firstly, the correlation between increased PAD and shorter total disease duration was 250 times more plausible than the absence of any correlation (Kendall's tau = −0.291 with a credible interval from −0.423 to −0.139, $BF_{10} = 250.206$; Fig. 3A). The total disease duration was estimated based on patients' memory of their own disease onset. In addition, we correlated the disease duration from time point of MRI to death and PAD. The correlation between older appearing brain and shorter survival was 97 times more plausible than the absence of any correlation (Kendall's tau = −0.272 with a 95% credible interval of −0.405 to −0.120, $BF_{10} = 96.94$; Fig. 3B).

What is the focal representation of increased brain age in ALS?

We were wondering to what degree PAD score correlated with motor cortex atrophy. In addition, we tested with which brain volumes PAD score was associated in controls. Correlation of PAD and whole-brain grey-matter maps showed significantly different patterns between healthy elderly people and patients with ALS. While in healthy elderly people, the focal representation of increased PAD score was mainly seen in midcingulate cortex, rolandic operculum and postcentral gyrus (Voxels >1000; *t*-score >4.5; uncorrected $P < 0.001$; Fig. 4A), patients with ALS showed remarkable focal atrophy in frontotemporal and motor cortex as well as in the thalamus (Fig. 4B). Thus, PAD was associated with motor cortex atrophy in both, HCs and patients with ALS. This also means that motor cortex atrophy was not correlated with PAD score exclusively in patients with ALS (Supplementary Fig. 6A and B). Frontobasal structures distinguished the PAD-VBM correlations between ALScn compared with ALSi (ci, bi, cbi and FTD; Supplementary Fig. 6C) or HC compared with ALSi (Supplementary Fig. 6D).

What are possible resilience factors?

We next investigated which focal brain map patterns contributed best to the younger brain age in slowly progressive patients with ALS. We compared voxel-wide grey-matter

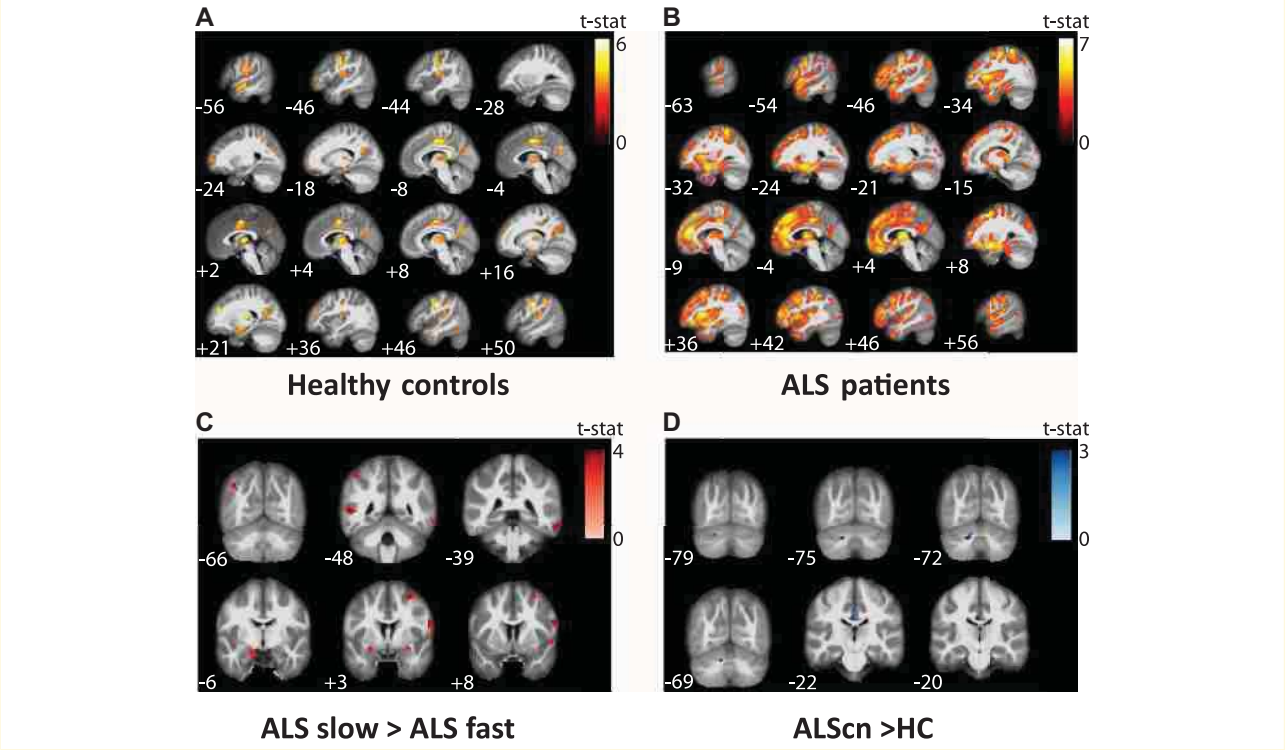


Figure 4 Correlation of PAD with voxel-based morphometry data showed significantly different patterns between healthy elderly people and patients with ALS. (**A**) The focal representation of increased PAD score in healthy controls is significant different to the **(B)** disease-associated focal representation of increased PAD of ALS showing a typical frontotemporal atrophy pattern. (**C**) Comparison of voxel-wide grey-matter volumes between ALS fast and slow progressors (ALS slow > ALS fast). (**D**) Comparison of voxel-wide grey-matter volumes between ALScn cases and controls (ALScn > controls). Significant clusters are displayed with T-score values represented by a colour map. An uncorrected threshold of $P = 0.001$ was used for all the presented illustrations and only clusters with at least 50 voxels extent were retained in the results. All clusters shown in **A** and **B** also passed a more conservative significance threshold of $P = 0.05$ applying false discovery rate (FDR) correction. No clusters in **C** and **D** survived FDR correction. All voxel-based analyses were controlled for total intracranial volume, chronologic age, sex and site of measurement as these were potential nuisance variables.

volumes between ALS fast versus slow progressors and identified significant regional atrophy mainly in the operculum and temporal lobe (Voxels ~1000, *t*-value >4.0; Fig. 4C) in fast compared with slow progressors.

We finally studied the surprisingly younger brain age in ALScn patients and compared voxel-wide grey-matter volumes between ALScn cases and controls. We identified few and only very small significant focal atrophy patterns (Voxels <50, *t*-value >3.5) in ALScn patients compared with HCs. Of note, however, we detected relative increase of grey-matter volume in ALScn patients compared with controls in left Crus II and left Lobule VIIa (Voxels >250, *t*-values >3.0; Fig. 4D).

Discussion

We used volumetric MRI with data-driven machine-learning algorithms to estimate individuals' brain age in patients with ALS and age-matched controls. We had hypothesized that PAD would be increased in motor impairment-only ALS cases and that this effect would even be more pronounced in the presence of additional cognitive and behaviour impairment. PAD was a very stable parameter of the individual, neither affected by age of onset, motor subtypes, disease onset type or disease duration until time point of investigation. Against our a priori hypothesis, we found strong evidence that predicted brain age was not increased in ALS *per se*; however, higher PAD was observed in patients with ALS who were additionally cognitively and/or behaviourally impaired. Surprisingly, predicted brain age was lower in ALScn patients and the subgroup of slowly progressive patients with ALS when compared with HCs. While a significant number of studies reported increased PAD in disease conditions such as Alzheimer's disease, multiple sclerosis, epilepsy or schizophrenia, none have reported reduced/younger brain ages in a disease condition.

To better understand the unexpected results of younger brain age in ALScn compared with controls, we went on to investigate correlations between PAD and grey-matter volume in a VBM analysis. Our initial hypothesis was that the correlate for PAD score variance in ALS was motor cortex atrophy. Indeed, PAD correlated with motor cortex grey matter, but also with large areas outside of motor cortex

Table 3 Summary of the Model Comparisons based on the Bayesian ANCOVA

Model Name	*P*(M)	*P*(M|data)	BF_{10}	Error %
Null model (incl. sex, age, recruitment location)	0.02	5.02e − 5	1.00	
Strong profile + progressor type	0.02	0.24	4803.70	1.970
Strong profile	0.02	0.03	524.74	3.30
Progressor type	0.02	2.566e − 5	5.52	3.43
Disease duration until examination	0.02	8.55e − 5	1.70	3.21
Phenotype	0.02	2.16e − 5	0.43	3.15
Age at onset	0.02	1.18e − 3	0.37	3.49
Onset type	0.02	3.87e − 4	0.12	3.32
LMN versus UMN	0.02	2.75e − 3	0.77	3.42

BF_{10}, Bayes factor in favour of the model compared with the null model; error%, numerical stability of the BF_{10} over 10 000 MCMC iterations; LMN, lower motor neuron dominant; *P*(M), prior probability of this model; *P*(M|data), posterior probability of this model after data analysis; UMN, upper motor neuron dominant.

(Fig. 4 and Supplementary Fig. 6). PAD score was associated with frontotemporal lobe atrophy, consistent with pattern of brain atrophy found in ALS cases with cognitive and/or behavioural impairment. Furthermore, focal temporal lobe atrophy pattern was the morphologic correlate of increased PAD score in fast versus slow progressors (Fig. 4C). Similar atrophy patterns were recently reported in other VBM studies.[23] Together these data inform us that the used machine-learning algorithm was sensitive enough to detect changes typically found in patients with ALS and distinct ALS populations.

One of the key findings was the surprisingly lower brain age of ALScn patients compared with HCs and the increased relative brain volume in cerebellar structures. Of note, Qiu and colleagues[24] reported increased grey-matter volume in cerebellar subregions in a VBM study. Similarly, Zhou and colleagues[25] reported an increased brain functional network connectivity in the cerebellum of patients with ALS. Since both studies excluded patients with an FTD diagnosis and cognitive impairment (MoCA <26), these patients very much resembled our ALScn group. The cerebellum contributes to executive functions such as planning, verbal fluency, abstract reasoning and working memory.[26] All of these functions are typically impaired in ALSi patients. Thus, cerebellar compensation may be specifically important in cognitively unimpaired patients with ALS and thus could be considered as resilience factor against executive dysfunction associated with shortened survival.

Consequently, the question arises as to how this translates, e.g. to C9ORF72-ALS patients, the most common monogenetic form of ALS. In C9ORF72 patients, most abundant dipeptide repeat–associated neuropathology is found in the cerebellum.[27] It is remarkable though that C9ORF72 patients do much more often suffer from cognitive/behavioural impairment and shorter survival than sporadic patients with ALS.[28] In agreement with these data, our findings suggest a cerebellar resilience factor against cognitive/behavioural impairment associated with longer survival in patients with ALS which needs further testing in independent cohorts.

We cannot yet generalize whether motor system neurodegeneration without cognitive affection does not lead to accelerated brain ageing also in other motor neurodegenerative diseases such as Parkinson's disease, or whether this is a specific finding for ALS. Only two studies on Parkinson's disease have been published so far, both reporting 'surprisingly' small increases in PAD,[29] especially when compared with Alzheimer's disease.[30] Of note, neither study distinguished systematically between demented, cognitively impaired and non-demented Parkinson's disease patients. However, they showed a negative correlation between cognitive performance (measured by MoCA) and PAD. Thus, future studies are needed, separating Parkinson's disease patients with and without (mild) cognitive impairment to address this question.

Limitations of the study are the small sample sizes of ALS subgroups, specifically in the ALSbi, –cbi and –FTD subgroups, which is also true for the motoric phenotypes. This did not allow us to analyse all suspected confounding factors such as diets, environmental pollutants, trauma, drug use, etc. Thus, there is considerable heterogeneity in the whole cohort which might explain some of the variances, e.g. in disease duration. However, this distribution reflects the population incidence of motor subtypes and cognitive/behavioural impairment, as ALS presents very heterogeneously.[20] Importantly, small sample sizes serve to detect large or very large effects, as reported here. However, this limitation does not apply to the overall results of a negative PAD in ALScn and positive PAD only in case of additional cognitive and/or behavioural impairment. Nevertheless, larger follow-up studies are warranted to further determine PAD and the underlying processes in the different forms of ALS. Furthermore, the brain-age analysis pipeline yields a single value and its ease of use might make it well suited in routine clinical care. However, it is conceptualized on the whole brain and distinct neuroanatomical information are not available. Consequently, it might not be sensitive enough for every disease entity depending on the spatial patterns of brain atrophy.

Therefore, we performed extensive correlations using VBM analysis. By doing so, we can conclude that we showed here the value of brain-age algorithms in the motor neurodegenerative disease ALS. In addition, we report unexpected findings of younger brain age in patients with ALS without cognitive/behavioural impairments ('cognitive normal') not only if compared with ALSi (=ci, bi, cbi or FTD) patients but even if compared with HCs with possible cerebellar resilience factors against cognitive/behavioural impairment in ALS.

Acknowledgements

The authors acknowledge all patients and healthy controls for their participation in the intersite project. J.M. is funded

by the federal state of Saxony-Anhalt and the European Regional Development Fund (ERDF) in the Center for Behavioral Brain Sciences (CBBS, ZS/2016/04/78113).

Funding

This work was supported in part by the NOMIS foundation, the Boris Canessa ALS foundation and the Hermann und Lilly Schilling-Stiftung für medizinische Forschung im Stifterverband. None of those had any role in the design and execution of this study, nor in the subsequent analyses and interpretation of the data, nor in the decision to publish the results. We were not paid to write this article by a pharmaceutical company or other agency.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain Communications* online.

References

1. Brettschneider J, Del Tredici K, Toledo JB, *et al.* Stages of pTDP-43 pathology in amyotrophic lateral sclerosis. *Ann Neurol*. 2013;74(1):20-38.
2. Prudlo J, Konig J, Schuster C, *et al.* TDP-43 pathology and cognition in ALS: A prospective clinicopathologic correlation study. *Neurology* 2016;87(10):1019-1023.
3. Brownell B, Oppenheimer DR, Hughes JT. The central nervous system in motor neurone disease. *J Neurol Neurosurg Psychiatry* 1970;33(3):338-357.
4. Gorges M, Vercruysse P, Muller HP, *et al.* Hypothalamic atrophy is related to body mass index and age at onset in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* 2017;88(12):1033-1041.
5. Bensimon G, Lacomblez L, Meininger V. A controlled trial of riluzole in amyotrophic lateral sclerosis. ALS/Riluzole study group. *N Engl J Med*. 1994;330(9):585-591.
6. Elamin M, Phukan J, Bede P, *et al.* Executive dysfunction is a negative prognostic indicator in patients with ALS without dementia. *Neurology* 2011;76(14):1263-1269.
7. Abrahams S, Leigh PN, Goldstein LH. Cognitive change in ALS: A prospective study. *Neurology* 2005;64(7):1222-1226.
8. Strong MJ, Abrahams S, Goldstein LH, *et al.* Amyotrophic lateral sclerosis—Frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. *Amyotroph Lateral Scler Frontotemporal Degener*. 2017;18(3–4):153-174.
9. Rascovsky K, Hodges JR, Knopman D, *et al.* Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 2011;134(Pt 9):2456-2477.
10. Collaborators GBDMND. Global, regional, and national burden of motor neuron diseases 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2018;17-(12):1083-1097.
11. Raz N, Rodrigue KM. Differential aging of the brain: Patterns, cognitive correlates and modifiers. *Neurosci Biobehav Rev*. 2006;30(6):730-748.
12. Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW. Mapping cortical change across the human life span. *Nat Neurosci*. 2003;6(3):309-315.
13. Franke K, Gaser C. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Front Neurol*. 2019;10:789.
14. Cole JH, Raffel J, Friede T, *et al.* Longitudinal assessment of multiple sclerosis with the brain-age paradigm. *Ann Neurol*. 2020;88(1):93-105.
15. Cole JH, Marioni RE, Harris SE, Deary IJ. Brain age and other bodily 'ages': Implications for neuropsychiatry. *Mol Psychiatry* 2019;24(2):266-281.
16. Wrigglesworth J, Ward P, Harding IH, *et al.* Factors associated with brain ageing—A systematic review. *BMC Neurol*. 2021;21(1):312.
17. de Lange AG, Anaturk M, Suri S, *et al.* Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *Neuroimage* 2020;222:117292.
18. Cole JH, Ritchie SJ, Bastin ME, *et al.* Brain age predicts mortality. *Mol Psychiatry* 2018;23(5):1385-1392.
19. Schuster C, Kasper E, Dyrba M, *et al.* Cortical thinning and its relation to cognition in amyotrophic lateral sclerosis. *Neurobiol Aging*. 2014;35(1):240-246.
20. Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis. *Nat Rev Neurol*. 2014;10(11):661-670.
21. Ashburner J. A fast diffeomorphic image registration algorithm. *NeuroImage* 2007;38(1):95-113.
22. Temp AGM, Naumann M, Hermann A, Glass H. Applied Bayesian approaches for research in motor neuron disease. *Front Neurol*. 2022;13:796777.
23. Albuquerque M, Andrade H, Silva C, Nucci A, Franca J. Marcondes. Voxel-based morphometry (VBM) study in ALS (amyotrophic lateral sclerosis): Temporal cortical damage as a prognostic marker (P4.097). *Neurology* 2014;82(10 Supplement):P4.097.
24. Qiu T, Zhang Y, Tang X, *et al.* Precentral degeneration and cerebellar compensation in amyotrophic lateral sclerosis: A multimodal MRI analysis. *Hum Brain Mapp*. 2019;40(12):3464-3474.
25. Zhou C, Hu X, Hu J, *et al.* Altered brain network in amyotrophic lateral sclerosis: A resting graph theory-based network study at voxel-wise level. *Front Neurosci*. 2016;10:204.
26. Schmahmann JD, Sherman JC. The cerebellar cognitive affective syndrome. *Brain* 1998;121(Pt 4):561-579.
27. Mackenzie IR, Frick P, Grasser FA, *et al.* Quantitative analysis and clinico-pathological correlations of different dipeptide repeat protein pathologies in C9ORF72 mutation carriers. *Acta Neuropathol*. 2015;130(6):845-861.
28. Byrne S, Elamin M, Bede P, *et al.* Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: A population-based cohort study. *Lancet Neurol*. 2012;11(3):232-240.
29. Eickhoff CR, Hoffstaedter F, Caspers J, *et al.* Advanced brain ageing in Parkinson's disease is related to disease duration and individual impairment. *Brain Commun*. 2021;3(3):fcab191.
30. Beheshti I, Mishra S, Sone D, Khanna P, Matsuda H. T1-weighted MRI-driven brain age estimation in Alzheimer's disease and Parkinson's disease. *Aging Dis*. 2020;11(3):618-628.

Supplement Methodology (Full methods)

Design

This was a two-center prospective, observational cross-sectional and longitudinal study conducted between April 2011 and August 2013. The local ethics committees of both universities approved the study (Rostock: A 2011 56; Magdeburg: 75/11) and all subjects gave written informed consent prior to their inclusion.

Participants

We recruited 182 German participants in Rostock and Magdeburg, Germany. Persons with a history of brain injury, epilepsy or psychiatric illness were excluded. Control participants were screened for cognitive impairment using the Montreal Cognitive Assessment, and excluded if they scored below 26 out of total 30. 70 healthy controls and 112 patients diagnosed with ALS using the revised El Escorial criteria were included [1] (Supplemental Fig. 1). ALS cases were classified into ALS without cognitive/behavior impairments (ALScn), ALS with cognitive impairment (ALSci), ALS with behavior impairment (ALSbi), ALS with cognitive and behavioural impairments (ALScbi) and ALS with frontotemporal dementia (ALS-FTD) following the Strong and Rascovsky criteria[2,3]. Demographic details can be found in Table 1

Clinical and neuropsychological measures

Cognitive testing. Participants underwent full neuropsychological examination in the executive, memory, visuospatial and fluency domains, for details see Kasper, Schuster, Machts, Bittner, Vielhaber, Benecke, Teipel and Prudlo [4]. Where necessary, we corrected cognitive tests for motor impairment [5]. We calculated the patients' standardized z-scores based on the controls' means and standard deviations; z scores $\leq$ -2 were considered as impaired performance.

Behavioural assessment. For behavioural classification, we relied on clinical observations and proxy-rated *Frontal Systems Behavior Scales* [6]. Ratings were transformed to T-scores based on published norms, with T$\geq$65 reflecting behavioural impairment.

Motor impairment. We used the revised ALS-Functional Rating Scale ALSFRS-R, [7], and calculated patients' progression rate δ as: (48-current ALSFRS-R score)/months since disease onset. Participants with $\delta \geq 0.5$ were classified as "fast progressors" (n=56). Participants with $\delta < 0.5$ were considered "slow progressors" (n=38).

MRI acquisition and processing

MRI scanning was performed with two 3T Siemens Magnetom VERIO scanners (Erlangen, Germany) using a 32-channel head coil; one single scanner at each site (Rostock and Magdeburg, Germany). High-resolution T_1-weighted anatomical images were acquired using the magnetization-prepared rapid gradient echo (MPRAGE) sequence with the following parameters: 256x256 image matrix with 192 sagittal slices, FOV 250x250x192mm, voxel size 1x1x1mm³, echo time 4.82ms, repetition time 2500ms, and flip angle 7°. The anatomical T_1-weighted images were segmented into grey matter, white matter and cerebrospinal fluid partitions using the SPM12 toolbox in Matlab 2019a. Then, the *Diffeomorphic Anatomical Registration Through Exponentiated Lie (DARTEL)* algebra algorithm [8] was used in combination with a custom brainAgeR brain template to normalize the T_1-weighted images to the *Montreal Neurological Institute (MNI)* reference coordinate system. The estimated deformation fields were subsequently applied to the grey matter segments to bring them in MNI space as well, followed by modulation to preserve the total amount of grey matter and smoothing with an 8 mm Gaussian kernel for the voxel-based morphometry analysis. In phantom tests according to the American College of Radiology guidelines [9], both sites' scanners met the criteria for geometric accuracy, high contrast spatial resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent signal ghosting and low contrast object detectability.

Brain Age model and predicted Brain Age Difference (PAD)

For Brain Age estimation, we used the brainageR toolbox version 2.1 available at https://github.com/james-cole/brainageR. This model is implemented in R (https://www.r-project.org/), and had been trained on n= 3377 healthy individuals and validated on 857 people. For the prediction of Brain Age, it follows an automated pipeline starting with T1-weighted image segmentation and normalized using SPM12 and smoothing with an 4 mm Gaussian kernel. Then, the spatially normalized grey and white matter segments as well as cerebrospinal fluid segments were loaded into R, masked to exclude voxels with less than 30% tissue/fluid probability, and vectorized to apply a principal component transform. The transformed data was then entered in the pretrained Gaussian progress regression model to obtain the predicted brain age. Finally, the predicted age was subtracted from the chronological age to calculate the "predicted brain age difference" (PAD):

Predicted age difference (PAD) = Estimated Brain Age – Chronological Brain Age

While a positive PAD indicates an older appearing brain, a negative score suggests a younger appearing brain.

Voxel-based analysis of group differences

Complementarily, we performed a whole brain voxel-based morphometry analysis for which the normalized and smoothed grey matter maps were analyzed using Statistical Parametric Mapping (SPM12; http://www.fil.ion.ucl.ac.uk/spm).

First, we run a linear regression model correlating the PAD scores with gray matter volume to determine potential regional pattern of atrophy driving the PAD score. We conducted this analysis separately for the healthy controls and the ALS patients.

Furthermore, we performed group comparisons to investigate the brain volume differences between different clinical subgroups. On one side, we investigated the between group differences between the healthy controls, the ALS non-impaired (ni) and the ALS impaired groups through a full factorial model design. Here, ALS impaired group included the Strong subgroups bi, ci, cbi, and bv-FTD. Notably, the contrast vector was weighted by the size of these cognitive subgroups such that each patient was equally weighted. On the other side, a two-sample T-test design was run between ALS fast and slow progressors, dichotomized by a monthly decline of ALSFRS-R $\geq$0.5.

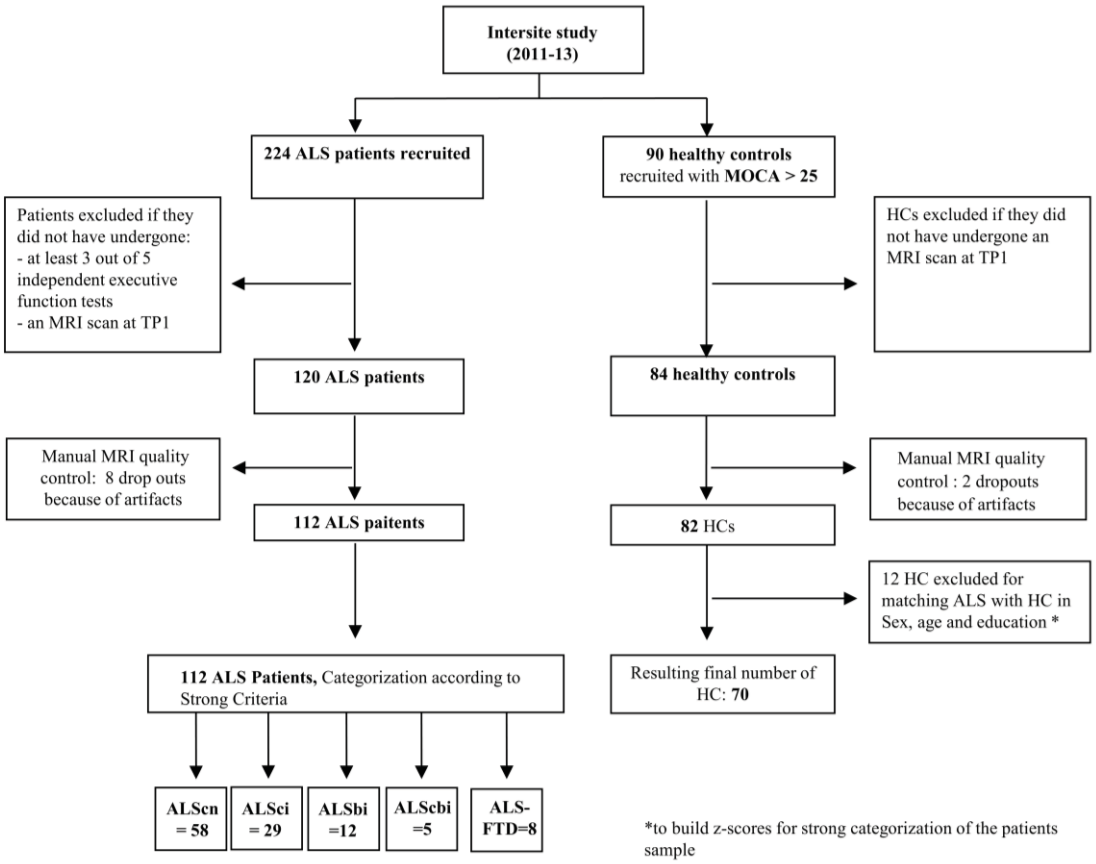
All voxel-based analyses were controlled for total intracranial volume (TIV), chronological age, sex and site as these were potential nuisance variables. The statistical threshold for the analyses was set to an uncorrected $p<0.001$ and only clusters with at least 50 voxels extent were retained in the results.

Statistical Analysis

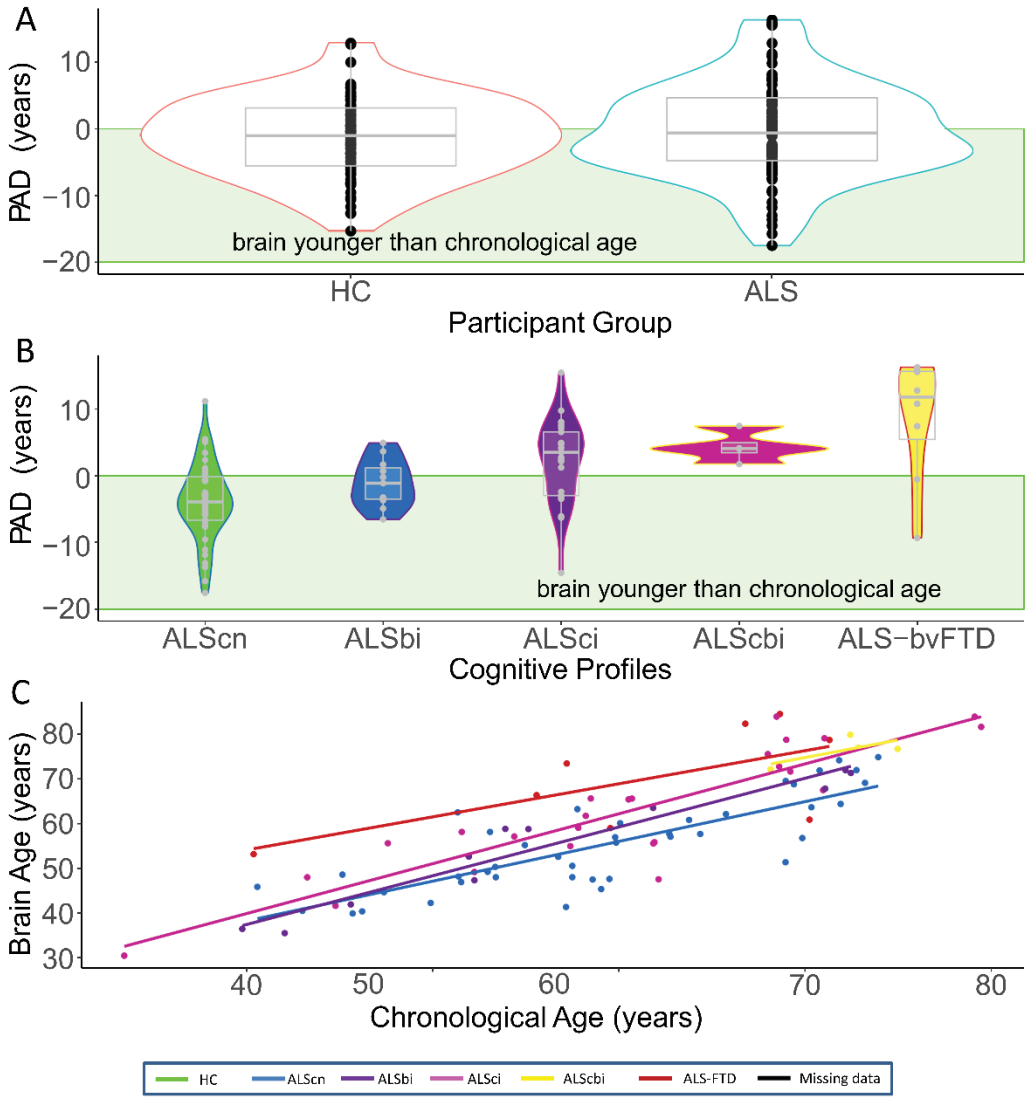
As classical null hypothesis significance testing only enables us to reject the null hypothesis that there are no effects of clinical presentation on PAD, we opted for *Bayes factor hypothesis testing (BFHT)* using an analysis of covariance. This Bayesian approach allows for the estimation of the likelihood of such effects given the observed data and, hence, more directly infer and compare the actual effects. Specifically, we compared the effects of Strong profile, progressor type, phenotype, onset type, disease duration until MRI scanning, and age at disease onset, while controlling for age at MRI, sex and recruitment location by adding them to the null model. We conducted one multi-factorial analysis of covariance (ANCOVA) which compared all these effects against one another, and against the corrected null hypothesis model. A priori, we assumed all models to be equally likely.

We applied default Jeffreys–Zellner–Siow (JZS) priors, with the seed set to 84293. Please see Table 2 for a summary of the statistical measures we will be reporting. All Bayesian analyses were conducted in *Jeffreys's Amazing Statistics Program* (JASP, 0.14.3). JASP was set to report the corrected null model on top, and to compare all other models against it using BF_{10}. Bayes factors do not require thresholding akin to $p<.05$ to determine statistical significance: instead they fall on a continuum ranging from support for the null hypothesis via no support for either hypothesis to support for the alternative hypothesis[10]. Additionally, we can add qualitative descriptors by stating that $BF_{10}>100$ constitutes "extreme evidence" for H_1, $BF_{10}>30$ constitutes "very strong" evidence for H_1, $BF_{10}>10$ constitutes "strong" evidence for H_1 and $BF_{10}>3$ constitutes "moderate" support for H_1.

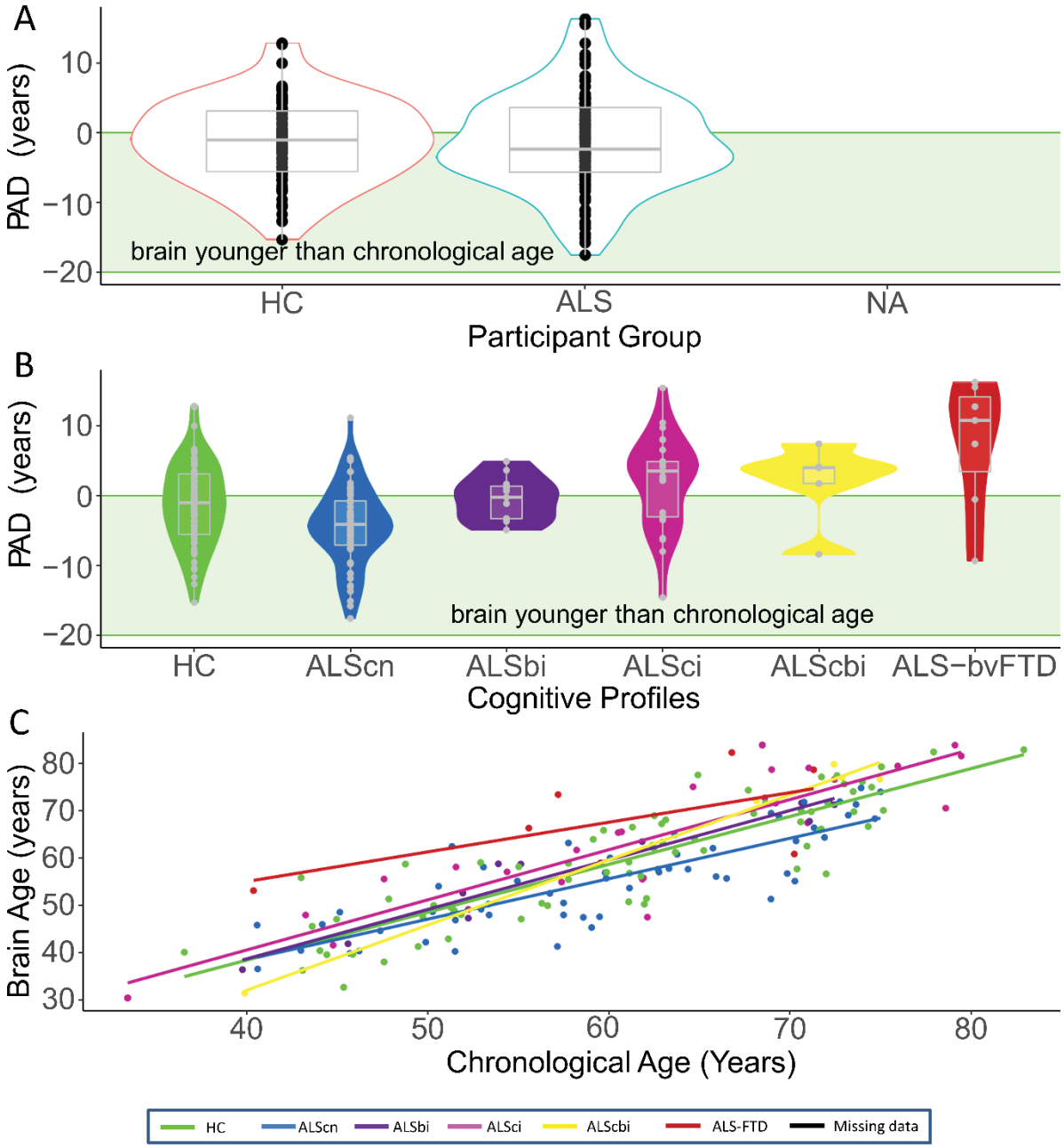
Supplementary Figures



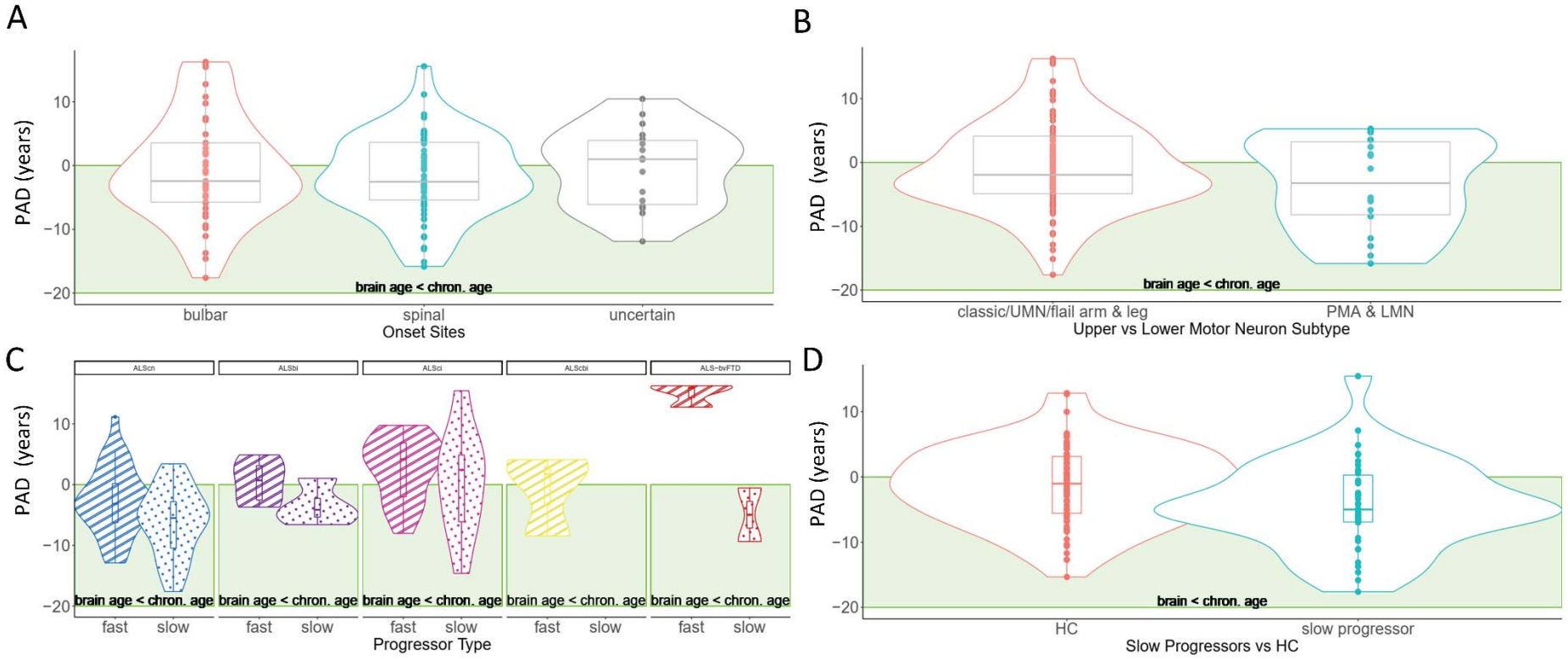
Supplementary Figure 1: Flow chart of the study population.



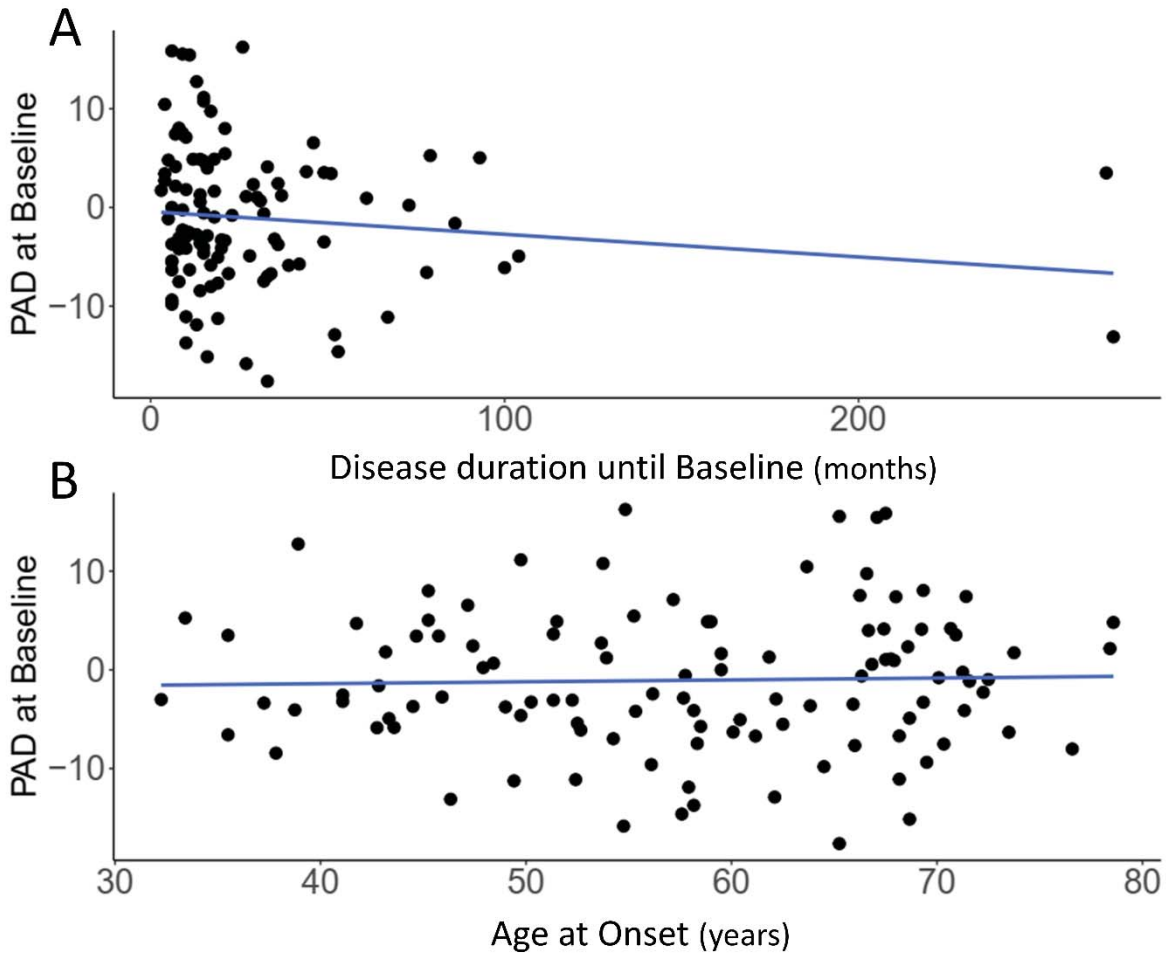
Supplementary Figure 2: Predicted brain age difference (PAD) is increased in cognitively/behaviorally impaired ALS patients also if patients without meeting El Escorial criteria are excluded. (A) There was no difference in PAD in ALS patients per se. **(B)** Cognitive/behavioral impairment increased PAD score significantly, while the difference between ALScn and HC prevailed when uncertain El Escorial types were excluded (BF_{10}=7.71). **(C)** Chronological age and predicted brain age correlated strongly and had a very narrow credible interval, suggesting a homogeneous, reliable effect.



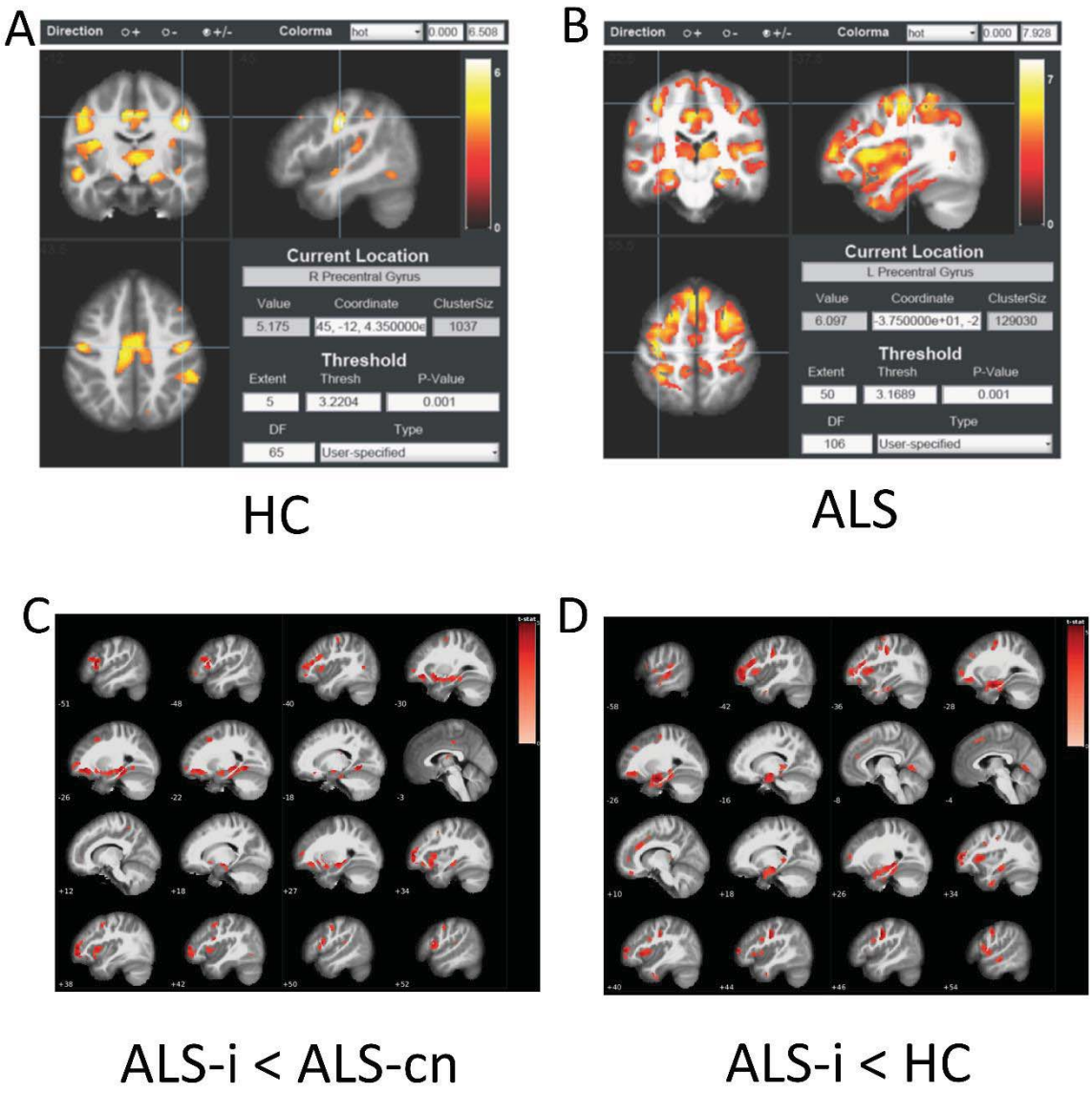
Supplementary Figure 3: Predicted brain age difference (PAD) is increased in cognitively/behaviorally impaired ALS patients also if patients with monogenetic ALS forms were excluded. (A) There was no difference in PAD in ALS patients per se. **(B)** Cognitive/behavioral impairment increased PAD score significantly, while the difference between ALScn and HC prevailed when genetic variants were excluded (BF_{10}=7.30). **(C)** Chronological age and predicted brain age correlated strongly and had a very narrow credible interval, suggesting a homogeneous, reliable effect.



Supplementary Figure 4: Predicted brain age is not influenced by motor subtypes but by disease progression rate. (A) Site of disease onset did not influence PAD (with its prior probability in the ANCOVA decreasing from 1.6% to below 0.0001%). **(B)** Upper motor neuron involvement was also not probable as a predictor of PAD: our data decreased the effect's plausibility by a factor of 10^4. **(C)** The comparison of slow (Δ ALSFRS-R <0.5) *vs.* fast disease progression (Δ ALSFRS-R ≥0.5) – measured by (48-current ALSFRS-R score)/months since disease onset – yielded moderate evidence favouring a main effect in every subgroup of the Strong criteria (ANCOVA combined main effects BF_{10}=4803.70). **(D)** Slowly progressive ALS patients had younger predicted brain age than healthy controls (directional Bayesian independent samples t-test, BF_{+0}=62.45).



Supplementary Figure 5: PAD score is a stable parameter. PAD did neither correlate with disease duration until timepoint of the MRI measurement (=baseline) **(A)** nor with age at onset **(B)**.



Supplementary Figure 6: Correlation of PAD with brain maps showed that motor cortex only partially caused the increased PAD score in ALS. (A-B) Significant clusters are displayed with color map representing T-score values on slices in axial, coronal and sagittal orientation, focusing on the precentral gyrus. Age-associated atrophy pattern in healthy elderly people **(A)** and disease-associated atrophy pattern in ALS patients **(B)** shows involvement of motor cortex and non-motoric regions in both. Same tresholds are used as for Figure 4. Represented current location (cross of lines) represents R/L precentral gyrus, respectively. Note the different maximum T-value and cluster size of the precentral gyrus in ALS and Controls. Disease-associated atrophy pattern in ALS showed larger effect (cluster size) in motor cortex. Furthermore, in ALS patients, significant more regions contributing to the PAD deviance, mainly in frontotemporal structures. **(C-D)** These frontobasal structures contributed to PAD deviance in ALSi patients (ALS impaired < cognitively normal **(C)** and ALS impaired < healthy controls **(D)**).

References

1. Brooks BR, Miller RG, Swash M, Munsat TL, World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord*. Dec 2000;1(5):293-9. doi:10.1080/146608200300079536
2. Rascovsky K, Hodges JR, Knopman D, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*. Sep 2011;134(Pt 9):2456-77. doi:10.1093/brain/awr179
3. Strong MJ, Abrahams S, Goldstein LH, et al. Amyotrophic lateral sclerosis - frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. *Amyotrophic lateral sclerosis & frontotemporal degeneration*. 2017;18(3-4):153–174. doi:10.1080/21678421.2016.1267768
4. Kasper E, Schuster C, Machts J, et al. Dysexecutive functioning in ALS patients and its clinical implications. *Amyotroph Lateral Scler Frontotemporal Degener*. Jun 2015;16(3-4):160-71. doi:doi.org/10.3109/21678421.2015.1026267
5. Abrahams S, Leigh PN, Harvey A, Vythelingum GN, Grisé D, Goldstein LH. Verbal fluency and executive dysfunction in amyotrophic lateral sclerosis (ALS). *Neuropsychologia*. 2000;38(6):734-747. doi:10.1016/s0028-3932(99)00146-3
6. Grace J, Malloy PH. *Frontal Systems Behavior Scale (FrSBe): Professional Manual*. Psychological Assessment Resources (PAR); 2001.
7. Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci*. 1999;169(1-2):13-21. doi:10.1016/s0022-510x(99)00210-5
8. Ashburner J. A fast diffeomorphic image registration algorithm. *NeuroImage*. 2007;38(1):95–113. doi:10.1016/j.neuroimage.2007.07.007
9. American College of Radiology. Phantom Test Guidance for Use of the Large MRI Phantom for the ACR MRI Accreditation Program. 2018.
10. Temp AGM, Naumann M, Hermann A, Glass H. Applied Bayesian Approaches for Research in Motor Neuron Disease. *Front Neurol*. 2022;13:796777. doi:10.3389/fneur.2022.796777

Journal of Alzheimer's Disease 65 (2018) 731–746
DOI 10.3233/JAD-161197
IOS Press

Comparison of Different Hypotheses Regarding the Spread of Alzheimer's Disease Using Markov Random Fields and Multimodal Imaging

Martin Dyrba[a,*], Michel J. Grothe[a], Abdolreza Mohammadi[b], Harald Binder[c], Thomas Kirste[d] and Stefan J. Teipel[a,e] for the Alzheimer's Disease Neuroimaging Initiative[1]
[a]*German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, Rostock, Germany*
[b]*Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands*
[c]*Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University, Mainz, Germany*
[d]*Mobile Multimedia Information Systems Group (MMIS), University of Rostock, Rostock, Germany*
[e]*Clinic for Psychosomatic and Psychotherapeutic Medicine, University Medical Center Rostock, Rostock, Germany*

Accepted 18 April 2017

Abstract. Alzheimer's disease (AD) is characterized by a cascade of pathological processes that can be assessed *in vivo* using different neuroimaging methods. Recent research suggests a systematic sequence of pathogenic events on a global biomarker level, but little is known about the associations and dependencies of distinct lesion patterns on a regional level. Markov random fields are a probabilistic graphical modeling approach that represent the interaction between individual random variables by an undirected graph. We propose the novel application of this approach to study the interregional associations and dependencies between multimodal imaging markers of AD pathology and to compare different hypotheses regarding the spread of the disease. We retrieved multimodal imaging data from 577 subjects enrolled in the Alzheimer's Disease Neuroimaging Initiative. Mean amyloid load (AV45-PET), glucose metabolism (FDG-PET), and gray matter volume (MRI) were calculated for the six principle nodes of the default mode network—a functional network of brain regions that appears to be preferentially targeted by AD. Multimodal Markov random field models were developed for three different hypotheses regarding the spread of the disease: the "intraregional evolution model", the "trans-neuronal spread" hypothesis, and the "wear-and-tear" hypothesis. The model likelihood to reflect the given data was evaluated using tenfold cross-validation with 1,000 repetitions. The most likely graph structure contained the posterior cingulate cortex as main hub region with edges to various other regions, in accordance with the "wear-and-tear" hypothesis of disease vulnerability. Probabilistic graphical models facilitate the analysis of interactions between several variables in a network model and therefore afford great potential to complement traditional multiple regression analyses in multimodal neuroimaging research.

Keywords: Alzheimer's disease, AV45-PET, FDG-PET, Markov random field, mild cognitive impairment, multimodal imaging, probabilistic graphical model

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and provided data but did not participate in analysis or in the writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Correspondence to: Martin Dyrba, German Center for Neurodegenerative Diseases (DZNE), c/o Zentrum für Nervenheilkunde, Gehlsheimer Str. 20, D-18147 Rostock, Germany. Tel.: +49 381 494 9482; Fax: +49 381 494 9472; E-mail: martin.dyrba@dzne.de.

INTRODUCTION

Alzheimer's disease (AD) is characterized by the extracellular deposition of amyloid-β ("plaques"), intracellular aggregation of hyperphosphorylated tau protein ("tangles"), and a progressive degeneration of intra-cortical projecting neurons [1, 2]. Over the last few decades, several neuroimaging markers have been developed to characterize the distinct regional lesion patterns within the continuum of cognitively healthy aging to AD dementia. Studies using structural magnetic resonance imaging (MRI) as an *in vivo* surrogate measure of neuronal loss [3] have consistently reported regional reductions of gray matter volume or cortical thickness in cortical limbic and temporal association areas [4–8]. Fluorodeoxyglucose positron emission tomography (FDG-PET) estimates the local cerebral metabolic rate of glucose consumption [9]. Hypometabolism reflecting neurodegenerative synapse dysfunction was found primarily in temporoparietal, medial temporal, and frontal areas [10–14]. Various tracers have been developed to assess the regional distribution of amyloid-β *in vivo* [15–17]. They show pronounced patterns of amyloid-β deposition in large parts of the neocortical association areas [18–20]. All three of these imaging modalities have been included in the recommendations of the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for AD to define the prodromal or even preclinical stages of the disease [21, 22].

Although the temporal dynamics of these imaging markers have been characterized on a global level in longitudinal clinical studies [23], little is known about the regional interactions of distinct lesion patterns across imaging modalities. Several hypothetical models for the spread of the disease have been proposed: the simplest pathology model is an "intraregional evolution model" [24–26], which assumes a sequence of amyloid-β deposition, hyperphosphorylation and aggregation of tau protein, metabolic dysfunction, and finally, neuronal death, all within a single region and independent of other regions. Similarly, the "trans-neuronal spread" hypothesis [27–29] assumes a prion-like propagation of pathogenic proteins along structural connections in the brain. In contrast, the "wear-and-tear" hypothesis [30, 31] supposes that brain regions with high metabolic demands, for example due to serving as highly interconnected hub regions, such as the precuneus/posterior cingulate cortex included in the default mode network [32–34], are particularly vulnerable to amyloid deposition, neuronal dysfunction, and, later on, neuronal death.

Several studies have investigated the relationship between pathologic patterns using multimodal imaging markers [12, 20, 30, 35–41]. However, these approaches were typically univariate. Thus, the studies 1) assessed the statistical associations between modalities using a voxel-by-voxel comparison [20, 40–41], 2) used estimates of an a priori defined seed region, such as gray matter atrophy of the hippocampus or amyloid deposition in the posterior cingulate cortex, to study its correlation with metabolism or volume throughout the brain [12, 30, 35–38, 42], or 3) used a global index of one modality, such as global cortical amyloid load on amyloid-PET, to stratify the sample into distinct subgroups and to assess regional group differences in the remaining modalities [12, 30, 37]. So far, few studies have investigated the associations and statistical dependencies between lesion patterns of brain regions on a network level. Such integrated analyses require the formulation of an appropriate statistical framework that can efficiently integrate multimodal and multiregional brain variables within the same model, and can account for a priori information on functional and structural connections. Initial approaches mainly used graph-theoretical metrics to analyze correlation patterns between each of two regions to identify cortical hubs or disease epicenters involved in AD [43–45]. Complementary approaches used advanced generative models that focused on one specific hypothesis of lesion propagation [28, 29, 46]. These models successfully predicted regional patterns and longitudinal dynamics of imaging markers and provided information about the individual contribution of hypothesized main factors for lesion propagation. However, these approaches also used complex, fine-tuned models to represent one specific hypothesis regarding the spread of the disease and, thus, cannot be applied generally to other imaging modalities or disease spreading hypotheses without major modifications.

Here we propose the novel application of Markov random fields, a probabilistic graphical network model that allows easy encoding of structural knowledge on the interaction between individual random variables by a graph representation, to study interregional multimodal associations and dependencies and to compare different hypotheses on the spread of AD pathology. Markov random fields are widely used in computer vision for image segmentation and

restoration [47, 48] or in language processing for text sequence labeling [49]. They are related to Bayesian networks and structural equation models, which use directed graphs to represent causal (acyclic) associations. In contrast, Markov random fields use undirected graphs that are able to model acausal (cyclic) interactions between random variables. This makes them an interesting candidate for building statistical models of systems with inherent spatial structures, such as functionally- or anatomically-connected brain regions. In this paper, we go beyond established statistical modeling approaches from the viewpoint of both directed and undirected graphical models. Directed models usually have a highly irregular structure, representing the causal influence of individual, carefully modeled variables. Typical undirected models have a highly regular structure of many variables that have essentially identical local interaction behavior, for instance a grid structure of adjacent voxels. Here we propose a new type of model: undirected with an irregular structure imposed by prior knowledge about functional and structural connectivity. This allows a better representation of the association between functional and structural brain characteristics, as expert knowledge on the causal relations between variables can be integrated that may not be present in the specific sample or that cannot be learned from the data automatically due to noise. A short introduction to the formalism of Markov random fields for discrete variables and the complementary method of Gaussian graphical models is provided in the Supplementary Material. For this study, we were interested in which of the three proposed mechanisms for the spread of the AD was most compatible with the multimodal neuroimaging data and how various brain regions and pathology patterns interacted with each other.

MATERIALS AND METHODS

This section is organized as follows: A basic introduction to Markov random fields for discrete data and Gaussian graphical models for multivariate normal data is provided in the Supplementary Material (Foundations of Markov random fields). The following subsections describe the study sample, image data acquisition and preprocessing, and feature extraction. Finally, we define graph structures based on three hypotheses regarding the spread of AD pathology proposed in the literature.

Subjects and imaging data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations, with the primary goal of testing whether neuroimaging, neuropsychological, and other biological measurements can be used as reliable *in vivo* markers of AD pathogenesis. A complete description of ADNI and up-to-date information is available at http://www.adni-info.org. For this study, 398 subjects with amnestic mild cognitive impairment (aMCI) and 179 cognitively healthy control subjects (CN) were selected from the ADNI-2 extension of the ADNI project, based on the availability of concurrent structural MRI at 3 Tesla, FDG-PET, amyloid-sensitive AV45-PET, and neuropsychological assessments. Detailed inclusion criteria for the diagnostic categories can be found at the ADNI website (http://adni.loni.usc.edu/methods). In brief, aMCI subjects had Mini-Mental State Examination (MMSE) scores between 24 and 30 (inclusive), a subjective memory concern reported by subject, informant, or clinician, plus objective memory loss measured by education-adjusted scores on delayed recall (Wechsler Memory Scale Logical Memory II), a clinical dementia rating (CDR) = 0.5, the absence of significant levels of impairment in other cognitive domains, essentially preserved daily-living activities, and the absence of dementia. Control subjects had MMSE scores between 26 and 30 (inclusive) and a CDR = 0, were non-depressed, non-MCI, and non-demented. Demographics and neuropsychological profiles of the different diagnostic groups used in the present study are summarized in Table 1.

Image data acquisition and processing

ADNI-GO/-2 MRI data were acquired on multiple 3T MRI scanners using scanner-specific T_1-weighted sagittal 3D MPRAGE sequences. To increase signal uniformity across the multicenter scanner platforms, original MPRAGE acquisitions in ADNI underwent standardized image preprocessing correction steps. FDG- and AV45-PET data were acquired on multiple instruments of varying resolution and following different platform-specific acquisition protocols.

Table 1
Sample characteristics

	CN	aMCI
Sample size	179	398
Age (y)	73.8 ± 6.5	71.7 ± 7.7*
Gender (M/F)	88/91	216/182
Education (y)	16.6 ± 2.5	16.1 ± 2.7*
MMSE	29.1 ± 1.2	28.1 ± 1.7*
DR	7.5 ± 4.0	5.0 ± 4.2*

Sample size, demographics, and neuropsychological test performance for each subject group. Numbers indicate group mean and standard deviation or number of subjects in each category for bivariate variables. Asterisks indicate significant difference between aMCI and CN groups ($p < 0.001$) based on two-sample *t*-test. Gender distribution did not differ significantly between aMCI and CN groups ($p = 0.25$, chi-square test). CN, cognitively healthy controls; DR, delayed recall of the 15-item wordlist of the Rey Auditory Verbal Learning Test; F, female; aMCI, amnestic mild cognitive impairment; M, male; MMSE, Mini-Mental State Examination.

Similar to the MRI data, PET data in ADNI were also subject to standardized image preprocessing correction steps, with the aim of increasing data uniformity across the multicenter acquisitions. More detailed information on the different imaging protocols employed across ADNI sites and the standardized image preprocessing steps for MRI and PET acquisitions can be found on the ADNI website (http://adni.loni.usc.edu/methods). As previously described [37, 41, 50, 51], imaging data were processed using the Statistical Parametric Mapping software (SPM8, Wellcome Trust Center for Neuroimaging) and the Voxel-Based Morphometry toolbox (VBM8, http://dbm.neuro.uni-jena.de/vbm) implemented in MATLAB R2013a (MathWorks, Natick, MA, USA). Initially, MRI scans were automatically segmented into gray matter, white matter, and cerebrospinal fluid partitions of 1.5 mm isotropic voxel size using the segmentation routine of the VBM8 toolbox. The resulting gray and white matter partitions of each subject in native space were then high-dimensionally registered to an aging/AD-specific reference template from a previous study [52] using the DARTEL algorithm [53]. Individual flow fields obtained from the DARTEL registration to the reference template were used to warp the gray matter segments, and voxel values were modulated for volumetric changes introduced by the high-dimensional normalization, such that the total amount of gray matter volume present before warping was preserved. All gray matter maps passed a visual inspection for overall segmentation and registration accuracy. Each subject's FDG- and AV45-PET scans were rigidly co-registered to a skull-stripped version of the corresponding structural MRI scan. Then, the PET scans were warped to the aging/AD-specific reference space without modulation of voxel values using the DARTEL flow fields of the corresponding MRI scans. To better characterize the aMCI subjects, we used the previously established thresholds for amyloid-positivity based on the standard uptake value ratios (SUVR), that is the global cortical signal divided by the mean uptake of the whole cerebellum, defined as $SUVR_{Cer} = 1.17$ [54, 55].

Feature extraction

For the present study, we selected six major regions of the default mode network (DMN) that are preferentially affected in AD and show alterations across various imaging modalities [6, 12, 30, 31, 56–62]. To not be biased with respect to any of the modalities under consideration, we selected a functionally-defined atlas derived from resting-state functional MRI [63] containing six clusters located in the posterior cingulum cortex (PCC), medial prefrontal cortex (MPC), left and right inferior parietal cortices (IPL, IPR), and in the left and right hippocampi (HPL, HPR) (Fig. 1). Before regional mean values were extracted, a stringent gray matter mask was applied to the DMN atlas which was derived from the aging/AD-specific reference template by applying a threshold of at least 50% of gray matter within each single voxel. Then, we calculated the mean gray matter volumes and mean FDG- and AV45-PET values within these clusters. Subsequently, regional gray matter volumes were proportionally scaled by total intracranial volume [64–66], regional FDG-PET values were proportionally scaled to pons uptake [12, 18, 37, 67], and regional AV45-PET values were proportionally scaled to whole-cerebellum uptake [12, 18, 67]. To be able to directly compare the different modalities, all values were normalized using the control group as reference. The so-called *W*-scores are analogous to *Z*-scores but are adjusted for specific covariates [7, 20]; age, gender, and education in the present case. These factors have been reported to influence brain volume, amyloid-β deposition, and metabolism previously [68–71]. Further, as for instance age highly interacts with the alterations caused by AD [69], we followed the advice from [72] to use reference data from cognitively healthy subjects not included in the primary analysis to determine and remove the effects of the covariates. Like *Z*-scores, *W*-scores have a mean value of 0 and a standard deviation of 1 in the

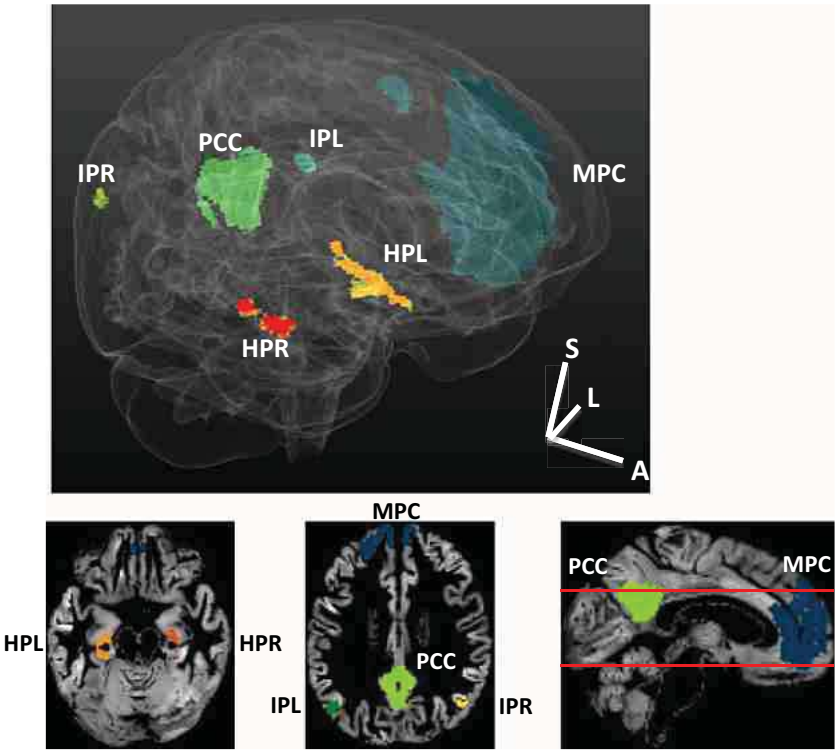


Fig. 1. Clusters of the default mode network, obtained using resting-state functional MRI. HPL/HPR, left and right hippocampus; IPL/IPR, left and right inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

control group, and values of +1.65 and –1.65 correspond to the 95th and 5th percentiles, respectively. To calculate the *W*-scores, regression models were estimated for the control group using age, gender, and education as independent variables and the mean value of each region as dependent variable. Then, *W*-scores were computed using $W = \left(x_{ij} - e_{ij}\right) / s_{res,j}$; with x_{ij} being the ith subject's raw value for region j; e_{ij} being the value expected for region j in the control group for the ith subject's age, gender, and education; and $s_{res,j}$ being the standard deviation of the residuals for region j in controls [7, 20]. The values for AV45-PET were reversed so that negative *W*-scores indicated pathology for all modalities, i.e., lower gray matter volume, lower glucose metabolism, but higher amyloid-β deposition. Finally, all data were dichotomized into the categories *pathologic* and *normal* using the 10th percentile of the controls as threshold [73, 74]. This step was necessary as Markov random fields model the association between discrete states instead of continuous variables (see below). We did not consider an extended interval-scaled discretization of the data as each additional factor level leads to a quadratic increase in the number of model parameters for each edge. To assess the influence of the threshold on the model fit, we additionally used the thresholds $W = -1$ and $W = -1.5$ corresponding to the 16th and 7th percentile to dichotomize the data and repeated the analyses.

Statistical modeling

Graph structures representing different hypotheses on the spread of AD

As a requirement for the Markov random field analysis, we needed to define alternative graph structures a priori. In order to study the associations between modalities, we derived specific graph structures representing the different hypotheses on the spread of AD (Fig. 2). We then compared the model fit for each pair of modalities: amyloid-β deposition/metabolism; metabolism/gray matter volume; and amyloid-β deposition/gray matter volume. Each model consisted of two sub-networks, one that represented the associations between regions within the same imaging modality, and another part that modeled the associations between the two imaging modalities. Model A represented a simple

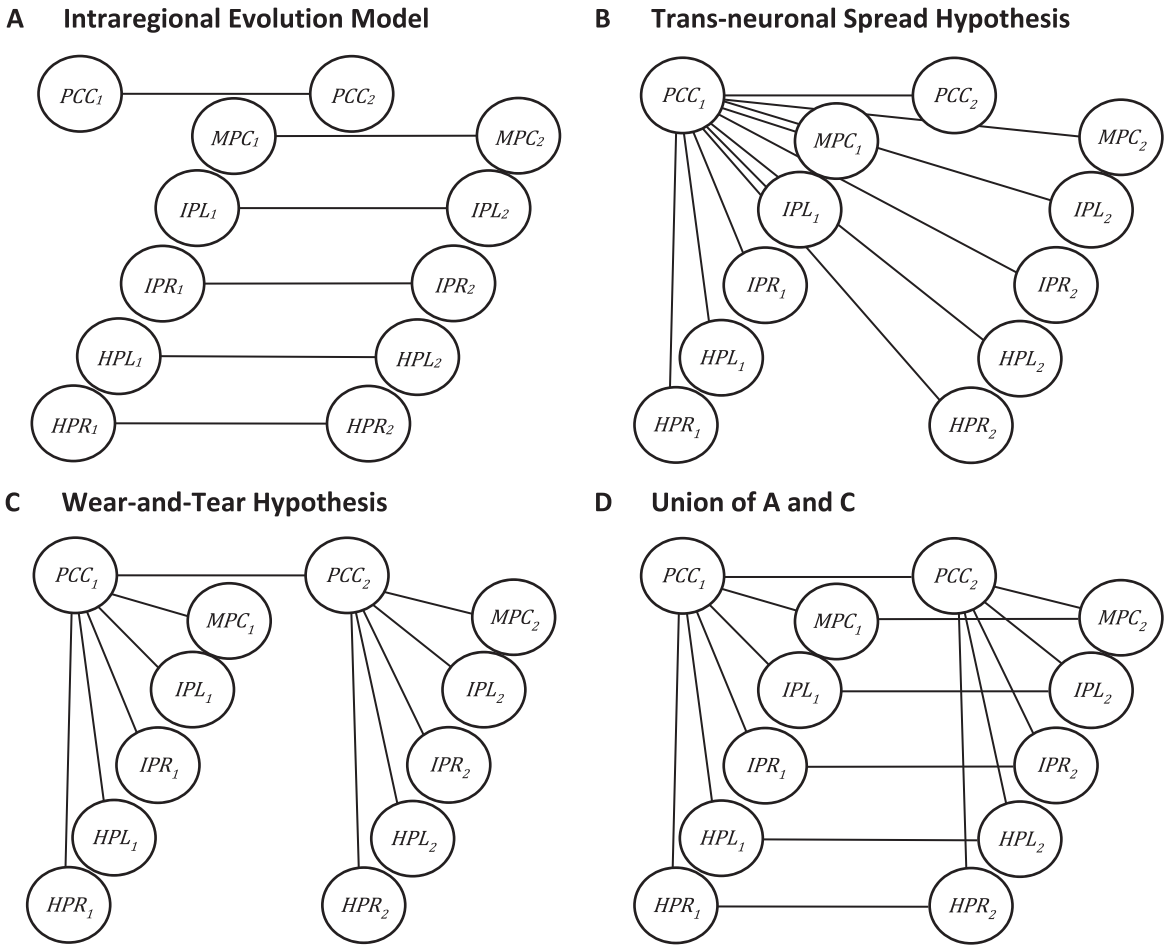


Fig. 2. Representative Markov random field models for the bi-modal analysis. Exemplary graph structures for the different hypotheses regarding disease spreading: (A) intraregional evolution model, (B) trans-neuronal spread hypothesis with the PCC as hub node, (C) wear-and-tear hypothesis with the PCC as hub nodes, and (D) union of structures (A) and (C) representing a combination of both models. For a further explanation refer to the Materials and Methods section. HPL/HPR, left and right hippocampus; IPL/IPR, left and right inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

"intraregional evolution model" [24–26] for which there was no statistical association between the regions, but an association between two pathologies (imaging modalities) in the same region (Fig. 2A). Model B was adapted from the "trans-neuronal spread" hypothesis [27–29] for which the pathology measured by the first modality in a disease-specific "epicenter" [35, 44], in Fig. 2B illustrated as the posterior cingulate cortex (PCC_1), triggered the pathology in all other connected regions of the same image modality. To differentiate the model for this hypothesis from the other models, we additionally assumed that the pathology of the "epicenter" (PCC_1) in the first modality triggered the pathology of all other regions in the second modality (Fig. 2B). Model C assumed the "wear-and-tear" hypothesis [30, 31] for which the pathology of the main hub, in Fig. 2C the posterior cingulate cortex (PCC_1), triggered the pathology in all other regions of the same modality plus the pathology of the hub region in the second modality, in Fig. 2C again the posterior cingulate cortex (PCC_2). The hub region of the second modality in turn triggered the pathology of all other regions in that modality (Fig. 2C). Finally, we supposed that combinations of two models might occur; that means that two processes run concurrently. For instance, the union of the models A and C, formally A∪C (Fig. 2D), assuming both local (within the same region) pathology propagation across modalities, as well as regional spread of pathology originating in the main hub region. For completeness, each possible default mode network node was selected as hub

region for models B, C and the pairwise combinations. However, we restricted the graph structure to only include one specific region as hub to avoid a combinatorial explosion of alternative graph structures. We excluded the right inferior parietal and right hippocampus cluster, as both clusters were only half of the size of their left counterparts in the functionally defined atlas [63]. The values for both regions were highly correlated ($r > 0.76$) with their counterparts in the left hemisphere.

Markov random field modeling

The Markov random field analyses were performed using the Undirected Graphical Modeling (UGM) toolbox (http://www.cs.ubc.ca/~schmidtm/Software/UGM.html, release 2013) implemented in MATLAB R2013a (MathWorks, Natick, MA, USA). To assess the stability of the model likelihood and individual parameters, we used tenfold cross-validation approach with 1,000 repetitions. For each iteration, Markov random field models were trained using maximum likelihood estimation for the training proportion of the data, and finally, the predictive deviance (log-likelihood) was obtained for the test proportion of the data. The performance of two graph structures was compared by calculating an empirical p-value from the proportion of cross-validation iterations [75] with $D = -2log\,(L_{M2}/L_{M1}) = -2\,(log\,(L_{M2}) - log\,(L_{M1})) > 0$, with the likelihood L for two different graph structures M_1 and M_2. We assessed the consistency of the likelihoods obtained for the different thresholds to dichotomize the data using the Pearson correlation coefficient of the vectors containing the average log-likelihood for the various graph structures and hub nodes: $L = -\log\left[L_A, L_{B,HPL}, \ldots, L_{D,PCC}\right]$. More detailed information about this modeling approach can be found in the Supplementary Material.

Gaussian graphical model analysis

In addition to the primary analysis of this study comparing Markov random field models with different graph structures that represent selected hypotheses on the spread of AD (models A–C above), we also used a data-driven approach where Gaussian graphical models learned the most likely graph structure from the data itself. The resulting graph structures were then compared with the manually specified models. For Markov random fields with discrete variables, graph structure learning is computationally expensive and often intractable without enforcing several assumptions about the underlying network structure and in addition it suffers from a high vulnerability to overfitting [76, Ch. 20.7]. Therefore, we used Gaussian graphical models to derive the most likely graph structure from the data. More specifically, we used the R package *flare* [77] (version 1.5.0) that implements a tuning-insensitive approach for optimally estimating large undirected graphs (*TIGER*). For assessing the stability of the resulting graphs, we again employed a ten-fold cross-validation scheme with 1,000 repetitions: The precision matrix that provides the graph structure was estimated for the training partition of the data using the raw *W*-scores instead of the dichotomized data and an internal tenfold cross-validation for selecting the optimal regularization parameter lambda that controls the density of the graph. Finally, we assessed the similarity of the Gaussian graphical models and the manually derived models A to D derived from the hypotheses regarding the spread of AD using the Jaccard similarity coefficient: $J = \left|E_i \cap E_j\right| / \left|E_i \cup E_j\right|$, with the sets of edges E_i and E_j for the Graphs i and j, respectively. J scales between 0 if both graphs do not share any edge and 1 if both graphs perfectly match each other.

RESULTS

Demographics and neuropsychological profiles of the study sample are summarized in Table 1. A proportion of 54% ($n = 219$) of the aMCI subjects exceeded the threshold of $SUVR_{Cer} = 1.10$ indicating amyloid-positivity and prodromal AD [54, 55]. Although age and education only slightly differed by 2.1 and 0.5 years, respectively, this difference reached statistical significance ($p < 0.001$, two-tailed, two-sample *t*-test) (Table 1). As expected, control subjects and aMCI patients significantly differed in MMSE and delayed recall scores of the 15-item wordlist of the Rey Auditory Verbal Learning Test [78, 79] (Table 1).

The results for the Markov random field models are given in Table 2 and Fig. 3. Table 2 contains the log-likelihood of the test proportion of the data in the cross-validation. Best model fits as indicated by lowest log-likelihood were obtained for different graph structures depending on the specific pair of modalities (Table 2). For the pair amyloid-β deposition/metabolism, the combined models B∪C provided best model fit, followed by both C and A∪C (Table 2). Further, highest likelihood was obtained using the posterior cingulate cortex as hub for these

Table 2
Negative log-likelihood for different hypotheses of disease spreading

Modalities	Model A	Model B	Model C	Model A∪B	Model A∪C	Model B∪C	Hub node
Amyloid –	253 ± 21*	238 ± 21*	224 ± 20*	236 ± 21*	221 ± 20*	224 ± 20*	HPL
Metabolism	–	217 ± 21*	206 ± 20	217 ± 21*	206 ± 20	206 ± 20	IPL
	–	213 ± 20*	199 ± 19	213 ± 20*	197 ± 19	198 ± 19	MPC
	–	209 ± 20*	196 ± 19	209 ± 20*	196 ± 19	**195 ± 19**	PCC
Amyloid –	248 ± 17*	231 ± 17*	225 ± 17*	231 ± 17*	225 ± 17*	225 ± 17*	HPL
Gray matter volume	–	213 ± 18	213 ± 17	213 ± 18	213 ± 18	213 ± 18	IPL
	–	208 ± 17	209 ± 17	207 ± 17	208 ± 17	208 ± 17	MPC
	–	**205 ± 17**	206 ± 17	**205 ± 17**	206 ± 17	**205 ± 17**	PCC
Metabolism –	230 ± 19*	218 ± 19	212 ± 19	216 ± 18	**211 ± 19**	212 ± 19	HPL
Gray matter volume	–	218 ± 18	219 ± 18	216 ± 18	216 ± 18	218 ± 18	IPL
	–	215 ± 18	215 ± 18	214 ± 18	213 ± 18	216 ± 18	MPC
	–	215 ± 18	215 ± 18	213 ± 18	213 ± 18	214 ± 18	PCC

Lower numbers represent better model fit for the test proportion of the data in the cross-validation. Lowest numbers are printed in bold letters. Asterisk indicates significant difference of the log-likelihood in comparison to best models ($p < 0.05$, empirical p-value based on cross-validation). A, intraregional evolution model; B, trans-neuronal spread hypothesis; C, wear-and-tear hypothesis (the respective network structures are given in Fig. 2); HPL, left hippocampus; IPL, left inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

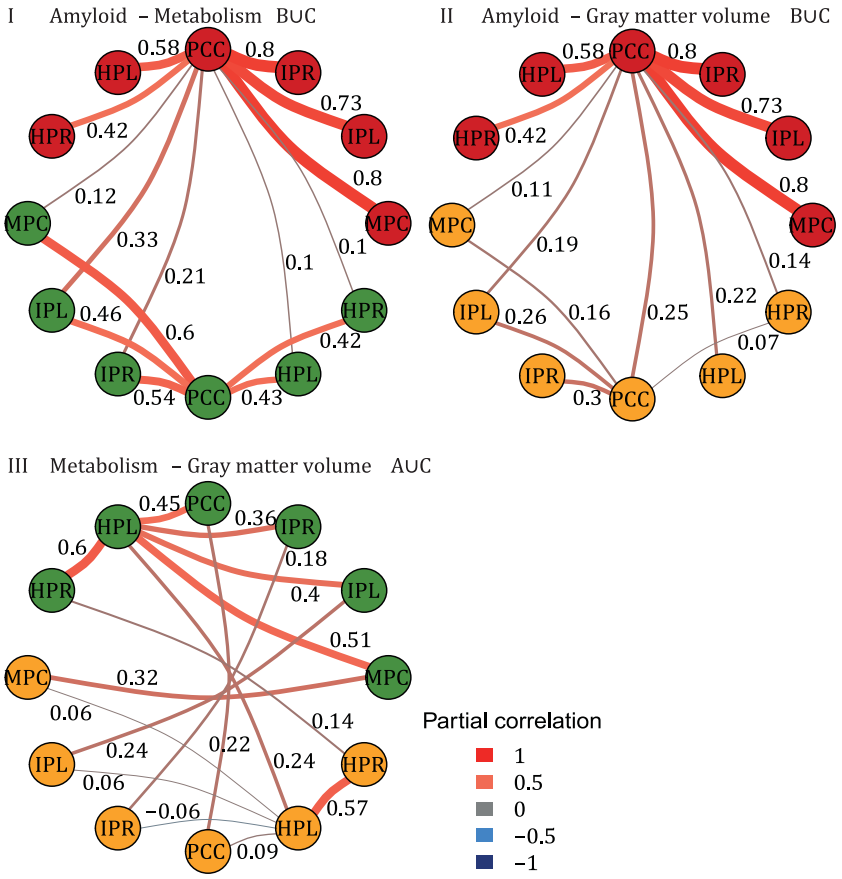


Fig. 3. Graph structures and partial correlation between nodes for the best Markov random field models. Based on the dichotomous data, thresholded at the 10th percentile. Edges represent the mean partial correlation across the 10000 cross-validation iterations. Edges with non-significant partial correlation were removed ($p < 0.05$). Red, amyloid-β; green, fluorodeoxyglucose metabolism; orange, gray matter; HPL/HPR, left and right hippocampus; IPL/IPR, left and right inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

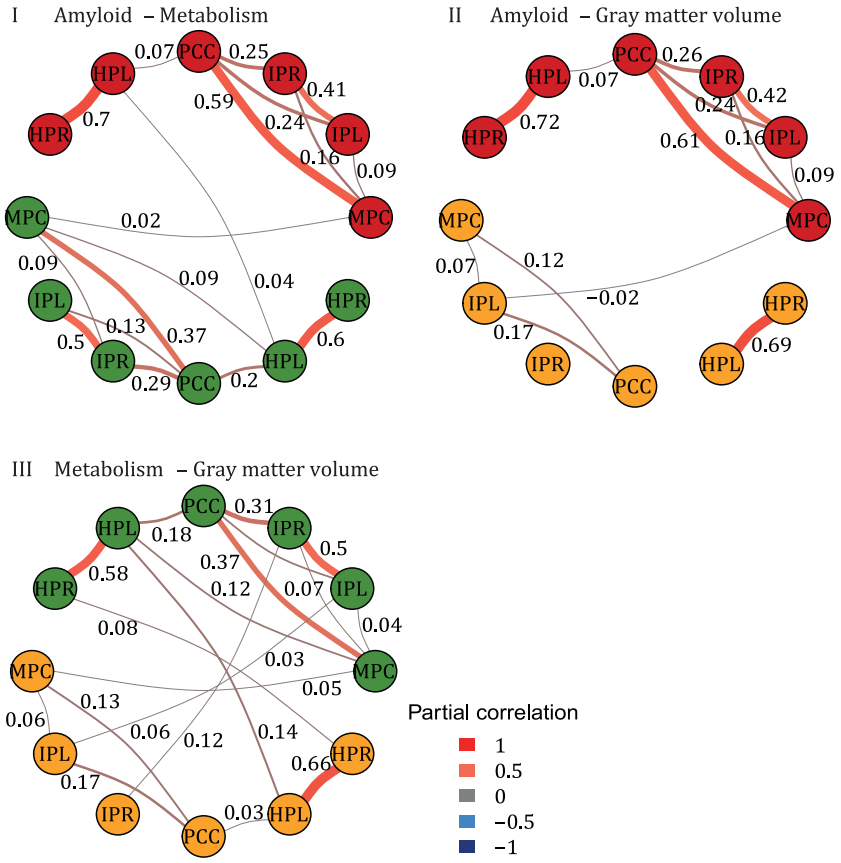


Fig. 4. Graph structures and partial correlation between nodes obtained from Gaussian graphical models for the raw W-scores. Edges represent the mean partial correlation across the 10000 cross-validation iterations, thresholded at a significance level of $p < 0.05$. Red, amyloid-β; green, fluorodeoxyglucose metabolism; orange, gray matte; HPL/HPR, left and right hippocampus; IPL/IPR, left and right inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

models. The best-performing model provided significantly better model fit than models A, B, A∪B, or any model using the left hippocampus cluster as hub node (Table 2). For the pair amyloid-β deposition/gray matter volume, best model fit was achieved by B, A∪B, and B∪C (Table 2). Notably, the remaining models C and A∪C performed only slightly worse. For this pair of modalities, again, the posterior cingulate cortex provided the best hub node, which yielded significantly better fits than model A or any graph structure including the left hippocampus cluster as hub (Table 2). For the combination metabolism/gray matter volume, best model fit was obtained for model A∪C using the left hippocampus cluster as hub node (Table 2). The difference in likelihood between alternative graph structures only reached statistical significance when compared to model A. We obtained highly concordant results when repeating all Markov random field analyses with the *W*-score thresholds –1 and –1.5 compared to the 10th percentile (data not shown). The likelihoods obtained for the alternative thresholds were correlated with the values given in Table 2 with $r \geq 0.93$ (Pearson coefficient).

The Gaussian graphical models obtained for the raw *W*-scores are displayed in Fig. 4. In summary, the algorithm estimated only few edges between different modalities in comparison to a high number of edges between the brain regions within the same modality. Seven of the eight (88%) detected edges between different modalities connect the same brain region, resembling the structure of model A (Fig. 4). Table 3 contains the Jaccard similarity coefficients for the graph structures obtained from the graph learning algorithm and the graphical models derived from hypotheses regarding the spread of AD. For all pairs of modalities, the learned graph structure

Table 3
Jaccard similarity coefficient for the graph structures learned from the data and the manually specified models

Modalities	Model A	Model B	Model C	Model A∪B	Model A∪C	Model B∪C	Hub node
Amyloid –	0.08 ± 0.01*	0.14 ± 0.01*	0.26 ± 0.01*	0.15 ± 0.01*	0.27 ± 0.01*	0.22 ± 0.01*	HPL
Metabolism	–	0.10 ± 0.00*	0.22 ± 0.01*	0.15 ± 0.01*	0.27 ± 0.02*	0.18 ± 0.01*	IPL
	–	0.17 ± 0.01*	$\mathbf{0.42 \pm 0.02}$	0.19 ± 0.01*	0.40 ± 0.02	0.34 ± 0.02*	MPC
	–	0.14 ± 0.00*	0.36 ± 0.02	0.19 ± 0.01*	0.40 ± 0.02	0.30 ± 0.02*	PCC
Amyloid –	0.01 ± 0.02*	0.13 ± 0.02*	0.27 ± 0.04	0.11 ± 0.02*	0.22 ± 0.03*	0.22 ± 0.03*	HPL
Gray matter volume	–	0.12 ± 0.01*	0.29 ± 0.04	0.11 ± 0.02*	0.25 ± 0.03*	0.23 ± 0.03*	IPL
	–	0.17 ± 0.01*	0.29 ± 0.03	0.15 ± 0.02*	0.25 ± 0.03*	0.24 ± 0.02*	MPC
	–	0.17 ± 0.01*	$\mathbf{0.40 \pm 0.03}$	0.15 ± 0.02*	0.33 ± 0.03*	0.32 ± 0.02*	PCC
Metabolism –	0.22 ± 0.02*	0.13 ± 0.01*	0.23 ± 0.02*	0.27 ± 0.02*	0.38 ± 0.03*	0.20 ± 0.02*	HPL
Gray matter volume	–	0.13 ± 0.01*	0.26 ± 0.03*	0.27 ± 0.02*	0.41 ± 0.03*	0.22 ± 0.02*	IPL
	–	0.21 ± 0.01*	0.30 ± 0.02*	0.35 ± 0.02*	0.46 ± 0.03	0.25 ± 0.01*	MPC
	–	0.16 ± 0.02*	0.34 ± 0.03*	0.33 ± 0.03*	$\mathbf{0.53 \pm 0.04}$	0.29 ± 0.03*	PCC

Mean and standard deviation of the Jaccard similarity coefficient for comparing the set of edges obtained from Gaussian graphical models and the manually specified models for the different hypotheses regarding the spread of Alzheimer's disease. Highest similarity coefficients are printed in bold letters. Asterisk indicates significant difference of the Jaccard index in comparison to the model with highest similarity ($p < 0.05$, empirical p-value based on cross-validation). A, intraregional evolution model; B, trans-neuronal spread hypothesis; C, wear-and-tear hypothesis (the respective network structures are given in Fig. 2); HPL, left hippocampus; IPL, left inferior parietal cortex; MPC, medial prefrontal cortex; PCC, posterior cingulate cortex.

was most similar to models C or A∪C, respectively (Table 3). Similarity was highest using the medial prefrontal cortex as hub node for the pair amyloid-β/metabolism. For amyloid-β deposition/gray matter volume and metabolism/gray matter volume, highest similarity was obtained when the posterior cingulate cortex served as hub node (Table 3).

DISCUSSION

Comparison of hypotheses regarding the spread of AD

Amyloid-β deposition and glucose metabolism

For the modality pair amyloid-β deposition/glucose metabolism, best Markov random field model fits were obtained for models C, A∪C, and B∪C with the posterior cingulate cortex cluster as main hub node (Table 2). In comparison to the other models, model B∪C combined edges within each modality as well as edges across modalities originating from amyloid deposition in the posterior cingulate cortex (Fig. 3I). These results suggest that the observed data best reflect a combination of the "trans-neuronal spread" [27–29] hypothesis, and the "wear-and-tear" hypothesis of regional susceptibility [30, 31], which assume a regional progression of lesions among highly active hub regions within and across modalities. These results match the competing evidence found for both hypotheses in the literature [27–31, 44]. In comparison to the study by Bischof et al. [80], which reported diverging directions of associations between amyloid load and metabolism, our analyses estimated all correlations to be strictly positive, i.e., pathologic amyloid deposition leading to hypometabolism (Fig. 3I). Other studies obtained similar results to ours [81], whereas Cohen et al. [82] reported contradicting results, i.e., pathologic amyloid deposition leading to hypermetabolism. The diverging results are commonly explained by different stages of AD, that means that initially amyloid-β deposition leads to hypermetabolism as compensatory effect [82]; while at more progressed stages when tau tangles accumulate, hypometabolism co-occurs [80, 83]. This hypothesis is consistent with our sample, where only 25 of the 398 aMCI subjects (6%) showed hypermetabolism, but only one of them actually showed a pathologic level of amyloid-β deposition. In contrast, hypometabolism was present in 185 aMCI subjects (46%), with 105 of those additionally having pathologic levels of amyloid-β deposition.

Amyloid-β deposition and gray matter volume

For the modality pair amyloid-β deposition/gray matter volume, Model B and the combinations A∪B and B∪C, favoring the "trans-neuronal spread" hypothesis as essential part, obtained a slightly better likelihood than models C and A∪C (Table 2). This suggests that again combinations of different hypotheses on the spread of AD provide nearly equally valid explanations for the observed pathology patterns that relate amyloid-β deposition and gray matter volume. Interestingly, the hub node posterior cingulate cortex provided best model fit, although gray matter atrophy was most present in

the hippocampi. However, as visible in Fig. 3II and III, the atrophy pattern of the hippocampi was highly correlated between both the left and right hippocampus ($r_\phi = 0.57$, Fig. 3III), but not so much with the pathology pattern of the remaining gray matter volume ($r_\phi < 0.1$, Fig. 3III). Instead, the correlation between the volume level of the posterior cingulate cortex cluster and the other regions except the hippocampi was substantially larger ($r_\phi \geq 0.16$, Fig. 3II). In comparison, the correlation between the amyloid-β pathology of the posterior cingulate cortex cluster and gray matter volume pathology was estimated to be of roughly the same magnitude ($r_\phi \geq 0.11$, Fig. 3II). These results are in line with previous studies using mainly linear regression models to assess the association between amyloid-β deposition and volume [38, 41, 84, 85]; however, it has to be noted that these results were partly obtained based on the same ADNI data as used in our analyses.

Glucose metabolism and gray matter volume

For the modality pair metabolism/gray matter volume, model A∪C provided slightly better fit than models C and B∪C, using the left hippocampus cluster as hub node (Table 2). When looking at the estimated edge weights for these models, we observed a high intraregional association between metabolism and volume ($r_\phi \geq 0.14$, Fig. 3III). These associations were substantially higher than between the left hippocampal volume and the volume of the other dorsal regions ($|r_\phi| \leq 0.09$, Fig. 3III). This can be explained by the current sample, in which most subjects showed selective memory impairment and therefore, atrophy was less pronounced in other regions than the hippocampi. Our results match several previous studies using univariate regression/correlation models [20, 41, 42, 86].

Graph learning

We complementarily employed hypothesis-free graph learning available for Gaussian graphical models. Here, estimated graph structures provided highest Jaccard similarity with models C (amyloid-β/metabolism, amyloid-β/gray matter volume) and A∪C (metabolism/gray matter volume), respectively (Table 3). Highest similarity was further obtained using the medial prefrontal cortex (amyloid-β/metabolism) or the posterior cingulate cortex cluster (amyloid-β/gray matter volume, metabolism/gray matter volume) as hub nodes (Table 3). In contrast to the Markov random field analyses for the dichotomized data, the Gaussian graphical models yielded a significantly lower number of edges between the modalities amyloid-β/metabolism and amyloid-β/gray matter volume (Fig. 4I and II versus Fig. 3I and II). This discrepancy between the Markov random field models and Gaussian graphical models could be explained by the high collinearity within the modalities amyloid-β and metabolism, such that the partial correlation between the different modalities was estimated to be non-significant. This multicollinearity is known to provide a serious source for model overfitting [72, 87, 88]. For the modality pair metabolism/gray matter volume, our results substantially differ and the estimated graph structure and associations showed a higher concordance between both modeling approaches (Fig. 4III versus Fig. 3III). Interestingly, edges between the two modalities connected the different types of pathology within the same brain regions (Fig. 4III), in compliance with the intraregional evolution model A, and in concordance with the common notion that neuronal injury is reflected by hypometabolism, which later progresses to neuronal death as represented by atrophy in volumetric MRI [20, 23, 41, 42, 60].

Summary

In our Markov random field analyses, the graphical models were derived from previous hypotheses regarding the spread of AD and specified to provide an easy and intuitive representation of the pathophysiological process. In this context, while being better for approximating the covariance structure from the given training data, graph learning as implemented in Gaussian graphical models may obtain results that are hard to interpret or that are influenced by noise in the training data. Therefore, we combined both approaches in order to be able to compare the results and complementary utilize the information obtained from these methodologies. In total, our results can be seen as confirmatory with respect to the literature of the last decade that places AD in the context of a network pathology [28, 29, 31, 43, 44, 46, 89, 90].

Methodological characteristics

Markov random fields and Gaussian graphical models assess statistical associations between various random variables based on correlation patterns in given training data and provide a graph structure that represents associated (i.e., statistically dependent) variables. The parameter learning process is formally a convex optimization problem that can be solved

efficiently using iterative gradient-descent algorithms. To avoid overfitting of the optimum set of parameters we used a repeated tenfold cross-validation approach with 1000 repetitions. For the Gaussian graphical model toolbox *flare* we additionally used an internal tenfold cross-validation for estimating the graph structure. Automated learning of the graph structure within the Markov random field model for discrete variables is theoretically possible but computationally intractable for data sets with more than 30 nodes [88, Ch. 17.4.1]. Another shortcoming of graphical models is the need to use either discrete or normally distributed continuous data. Although approaches for learning mixed models have been proposed recently [91], these are conceptually not easy to handle and suffer from computational challenges such as requiring a much larger sample size than the traditional approaches to obtain stable and reliable estimates [91]. For the Markov random field analyses, we applied the dichotomization of the variables into normal and pathologic categories using the 10th percentile of cognitively normal subjects as threshold, which has been commonly used in the literature [73, 74]. The additional analyses varying this threshold yielded essentially identical results with respect to the likelihood of different graph structures. Although dichotomization of variables is known to alter the covariance of the data [92], it substantially improves the interpretability of the results in a clinical context. We complimentarily applied Gaussian graphical models to the raw *W*-score data in order to circumvent possible biases arising from dichotomizing the variables and to be able to compare the results. Fine-tuned generative models, such as employed by Raj et al. [46] and Iturria-Medina et al. [28], rely on differential equation models and require careful definition of the functional relations between individual variables in order to estimate the most probable parameterization of the whole system. Descriptive models, such as used in the present study, may be used for identifying relevant associations between the various brain lesions and ultimately improve our understanding of the pathophysiological process of AD.

Limitations

An important shortcoming of the present study is that the study sample of aMCI subjects may not only contain subjects with an underlying AD pathology. Approximately half of our sample exceeded the threshold of amyloid-positivity [54, 55] that was previously established to indicate pathological amyloid deposition based on the 95% confidence interval upper limit of observed signal in a group of healthy young adults who are highly unlikely to exhibit cortical amyloid pathology [54, 55]. However, we decided to not restricting the sample to amyloid-β positive cases only, as constraining the global amyloid-β amount directly influences the observed patterns of regional amyloid-β deposition which were examined in this work. We focused on six major regions of the default mode network for the current study. These regions are preferentially affected in AD and show alterations across various imaging modalities [6, 12, 30, 31, 56–58, 60, 62, 93]. Although other brain regions may substantially contribute as well, we decided to use the same regions-of-interest for all three modalities a priori to simplify the graph structures for the analyses in this proof-of-concept study. To not be biased with respect to any of the modalities under consideration, we selected a functionally-defined atlas derived from resting-state functional MRI [63]. Stringent masking of the original atlas labels to include at least 50% of gray matter led to partly small, but very specific regions-of-interest for the subsequent analyses. Consequently, the estimated gray matter volume may be slightly biased due to the small size of the inferior-parietal clusters. However, if we used for instance an anatomical atlas, we would lose the regional specificity. In future work, we will obtain separate regions-of-interest for each modality based on a meta-analytic approach. Further, we used only cross-sectional baseline data from the ADNI database. Future studies may utilize longitudinal data to include the information on sequential and temporal interactions between regions and imaging modalities in the statistical analyses. Additionally, the integration of tau PET data may contribute substantial information that could enhance the link between amyloid-β deposition and neuronal dysfunction.

Conclusions

With their simple and intuitive representation of statistical associations using a graph model, Markov random fields and Gaussian graphical models offer convenient frameworks for studying the associations of distinct lesion patterns in neurologic diseases. They afford great potential to complement traditional multiple regression analyses in studying statistical associations in complex multimodal datasets.

ACKNOWLEDGMENTS

This project was supported by the Rostock Massive Data Research Facility (RMDRF) funded by the German Research Foundation (DFG) (FKZ INST 264/128-1 FUGG).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institute of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California. We would like to thank the ADNI (http://adni.loni.usc.edu/) and the Functional Imaging in Neuropsychiatric Disorders Lab (http://findlab.stanford.edu/) investigators for publicly sharing their valuable neuroimaging data.

Authors' disclosures available online (http://j-alz.com/manuscript-disclosures/16-1197r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-161197.

REFERENCES

[1] Morrison JH, Scherr S, Lewis DA, Campbell MJ, Bloom FE (1986) The laminar and regional distribution of neocortical somatostatin and neuritic plaques: Implications for Alzheimer's disease as a global neocortical disconnection syndrome. In *The Biological Substrates of Alzheimer's Disease*, Scheibel AB, Weschler AF, eds. Academic Press, New York, NY, pp. 115-131.

[2] D'Amelio M, Rossini PM (2012) Brain excitability and connectivity of neuronal assemblies in Alzheimer's disease: From animal models to human findings. *Prog Neurobiol* **99**, 42-60.

[3] Bobinski M, de Leon MJ, Wegiel J, Desanti S, Convit A, Saint Louis LA, Rusinek H, Wisniewski HM (2000) The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. *Neuroscience* **95**, 721-725.

[4] Singh V, Chertkow H, Lerch JP, Evans AC, Dorr AE, Kabani NJ (2006) Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. *Brain* **129**, 2885-2893.

[5] Li C, Wang J, Gui L, Zheng J, Liu C, Du H (2011) Alterations of whole-brain cortical area and thickness in mild cognitive impairment and Alzheimer's disease. *J Alzheimers Dis* **27**, 281-290.

[6] Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Möller HJ, Hampel H (2007) Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage* **38**, 13-24.

[7] Jack CR Jr, Petersen RC, Xu YC, Waring SC, O'Brien PC, Tangalos EG, Smith GE, Ivnik RJ, Kokmen E (1997) Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* **49**, 786-794.

[8] Risacher S, Saykin A, West JD, Shen L, Firpi H, McDonald B, Alzheimer's Disease Neuroimaging Initiave (ADNI) (2009) Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res* **6**, 347-361.

[9] Ido T, Wan CN, Casella V, Fowler JS, Wolf AP, Reivich M, Kuhl DE (1978) Labeled 2-deoxy-D-glucose analogs. ^{18}F-labeled 2-deoxy-2-fluoro-D-glucose, 2-deoxy-2-fluoro-D-mannose and ^{14}C-2-deoxy-2-fluoro-D-glucose. *J Labelled Comp Radiopharm* **14**, 175-183.

[10] Friedland RP, Budinger TF, Ganz E, Yano Y, Mathis CA, Koss B, Ober BA, Huesman RH, Derenzo SE (1983) Regional cerebral metabolic alterations in dementia of the Alzheimer type: Positron emission tomography with [18F]fluorodeoxyglucose. *J Comput Assist Tomogr* **7**, 590-598.

[11] Mielke R, Kessler J, Szelies B, Herholz K, Wienhard K, Heiss WD (1998) Normal and pathological aging–findings of positron-emission-tomography. *J Neural Transm (Vienna)* **105**, 821-837.

[12] Kljajevic V, Grothe MJ, Ewers M, Teipel S, Alzheimer's Disease Neuroimaging Initiative (2014) Distinct pattern of

hypometabolism and atrophy in preclinical and predementia Alzheimer's disease. *Neurobiol Aging* **35**, 1973–1981.

[13] Minoshima S, Giordani B, Berent S, Frey KA, Foster NL, Kuhl DE (1997) Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Ann Neurol* **42**, 85-94.

[14] Drzezga A, Lautenschlager N, Siebner H, Riemenschneider M, Willoch F, Minoshima S, Schwaiger M, Kurz A (2003) Cerebral metabolic changes accompanying conversion of mild cognitive impairment into Alzheimer's disease: A PET follow-up study. *Eur J Nucl Med Mol Imaging* **30**, 1104-1113.

[15] Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, Bergström M, Savitcheva I, Huang GF, Estrada S, Ausén B, Debnath ML, Barletta J, Price JC, Sandell J, Lopresti BJ, Wall A, Koivisto P, Antoni G, Mathis CA, Långström B (2004) Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Ann Neurol* **55**, 306-319.

[16] Rowe CC, Ackerman U, Browne W, Mulligan R, Pike KL, O'Keefe G, Tochon-Danguy H, Chan G, Berlangieri SU, Jones G, Dickinson-Rowe KL, Kung HP, Zhang W, Kung MP, Skovronsky D, Dyrks T, Holl G, Krause S, Friebe M, Lehman L, Lindemann S, Dinkelborg LM, Masters CL, Villemagne VL (2008) Imaging of amyloid beta in Alzheimer's disease with 18F-BAY94-9172, a novel PET tracer: Proof of mechanism. *Lancet Neurol* **7**, 129-135.

[17] Wong DF, Rosenberg PB, Zhou Y, Kumar A, Raymont V, Ravert HT, Dannals RF, Nandi A, Brasić JR, Ye W, Hilton J, Lyketsos C, Kung HF, Joshi AD, Skovronsky DM, Pontecorvo MJ (2010) In vivo imaging of amyloid deposition in Alzheimer disease using the radioligand 18F-AV-45 (Flobetapir [corrected] F 18). *J Nucl Med* **51**, 913-920.

[18] Lowe VJ, Kemp BJ, Jack CR Jr, Senjem M, Weigand S, Shiung M, Smith G, Knopman D, Boeve B, Mullan B, Petersen RC (2009) Comparison of 18F-FDG and PiB PET in cognitive impairment. *J Nucl Med* **50**, 878-886.

[19] Villeneuve S, Rabinovici GD, Cohn-Sheehy BI, Madison C, Ayakta N, Ghosh PM, La Joie R, Arthur-Bentil SK, Vogel JW, Marks SM, Lehmann M, Rosen HJ, Reed B, Olichney J, Boxer AL, Miller BL, Borys E, Jin LW, Huang EJ, Grinberg LT, DeCarli C, Seeley WW, Jagust W (2015) Existing Pittsburgh Compound-B positron emission tomography thresholds are too high: Statistical and pathological evaluation. *Brain* **138**, 2020-2033.

[20] La Joie R, Perrotin A, Barré L, Hommet C, Mézenge F, Ibazizene M, Camus V, Abbas A, Landeau B, Guilloteau D, de La Sayette V, Eustache F, Desgranges B, Chételat G (2012) Region-specific hierarchy between atrophy, hypometabolism, and β-amyloid (Aβ) load in Alzheimer's disease dementia. *J Neurosci* **32**, 16265–16273.

[21] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) 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* **7**, 270-279.

[22] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280-292.

[23] Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ (2013) Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol* **12**, 207-216.

[24] Hardy J, Allsop D (1991) Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol Sci* **12**, 383-388.

[25] Castello MA, Soriano S (2014) On the origin of Alzheimer's disease. Trials and tribulations of the amyloid hypothesis. *Ageing Res Rev* **13**, 10-12.

[26] Drachman DA (2014) The amyloid hypothesis, time to move on: Amyloid is the downstream result, not cause, of Alzheimer's disease. *Alzheimers Dement* **10**, 372-380.

[27] Hallbeck M, Nath S, Marcusson J (2013) Neuron-to-neuron transmission of neurodegenerative pathology. *Neuroscientist* **19**, 560-566.

[28] Iturria-Medina Y, Sotero RC, Toussaint PJ, Evans AC, Alzheimer's Disease Neuroimaging Initiative (2014) Epidemic spreading model to characterize misfolded proteins propagation in aging and associated neurodegenerative disorders. *PLoS Comput Biol* **10**, e1003956.

[29] Raj A, LoCastro E, Kuceyeski A, Tosun D, Relkin N, Weiner M, for the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2015) Network diffusion model of progression predicts longitudinal patterns of atrophy and metabolism in Alzheimer's disease. *Cell Rep* **10**, 359-369.

[30] Buckner RL, Snyder AZ, Shannon BJ, LaRossa G, Sachs R, Fotenos AF, Sheline YI, Klunk WE, Mathis CA, Morris JC, Mintun MA (2005) Molecular, structural, and functional characterization of Alzheimer's disease: Evidence for a relationship between default activity, amyloid, and memory. *J Neurosci* **25**, 7709-7717.

[31] Buckner RL, Sepulcre J, Talukdar T, Krienen FM, Liu H, Hedden T, Andrews-Hanna JR, Sperling RA, Johnson KA (2009) Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease. *J Neurosci* **29**, 1860-1873.

[32] Greicius MD, Srivastava G, Reiss AL, Menon V (2004) Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proc Natl Acad Sci U S A* **101**, 4637-4642.

[33] Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001) A default mode of brain function. *Proc Natl Acad Sci U S A* **98**, 676-682.

[34] Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME (2005) The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A* **102**, 9673-9678.

[35] Seeley WW, Crawford RK, Zhou J, Miller BL, Greicius MD (2009) Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62**, 42-52.

[36] Villain N, Fouquet M, Baron JC, Mézenge F, Landeau B, de La Sayette V, Viader F, Eustache F, Desgranges B, Chételat G (2010) Sequential relationships between grey matter and white matter atrophy and brain metabolic abnormalities in early Alzheimer's disease. *Brain* **133**, 3301-3314.

[37] Grothe MJ, Heinsen H, Amaro E Jr, Grinberg LT, Teipel SJ (2016) Cognitive correlates of basal forebrain atrophy and associated cortical hypometabolism in mild cognitive impairment. *Cereb Cortex* **26**, 2411-2426.

[38] Chang YT, Huang CW, Chang YH, Chen NC, Lin KJ, Yan TC, Chang WN, Chen SF, Lui CC, Lin PH, Chang CC (2015) Amyloid burden in the hippocampus and default mode network: Relationships with gray matter volume and cognitive performance in mild stage Alzheimer disease. *Medicine (Baltimore)* **94**, e763.

[39] Young AL, Oxtoby NP, Huang J, Marinescu RV, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC (2015) Multiple orderings of events in disease progression. In *Inf Process Med Imaging*, Ourselin S, Alexander DC, Westin C-F, Cardoso MJ, eds. Springer International Publishing, pp. 711-722.

[40] Altmann A, Ng B, Landau SM, Jagust WJ, Greicius MD, Alzheimer's Disease Neuroimaging Initiative (2015) Regional brain hypometabolism is unrelated to regional amyloid plaque burden. *Brain* **138**, 3734-3746.

[41] Grothe MJ, Teipel SJ, Alzheimer's Disease Neuroimaging Initiative (2016) Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks. *Hum Brain Mapp* **37**, 35-53.

[42] Teipel S, Grothe MJ, Alzheimer s Disease Neuroimaging Initiative (2016) Does posterior cingulate hypometabolism result from disconnection or local pathology across preclinical and clinical stages of Alzheimer's disease? *Eur J Nucl Med Mol Imaging* **43**, 526-536.

[43] Stam CJ (2014) Modern network science of neurological disorders. *Nat Rev Neurosci* **15**, 683-695.

[44] Zhou J, Gennatas ED, Kramer JH, Miller BL, Seeley WW (2012) Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron* **73**, 1216-1227.

[45] Sepulcre J, Sabuncu MR, Becker A, Sperling R, Johnson KA (2013) In vivo characterization of the early states of the amyloid-beta network. *Brain* **136**, 2239-2252.

[46] Raj A, Kuceyeski A, Weiner M (2012) A network diffusion model of disease progression in dementia. *Neuron* **73**, 1204-1215.

[47] Wang C, Komodakis N, Paragios N (2013) Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey. *Comput Vis Image Underst* **117**, 1610-1627.

[48] Ashburner J, Friston KJ (2005) Unified segmentation. *Neuroimage* **26**, 839-851.

[49] Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, Brodley CE, Danyluk A, eds. Morgan Kaufmann, San Francisco, pp. 282-289.

[50] Grothe MJ, Ewers M, Krause B, Heinsen H, Teipel SJ, Alzheimer's Disease Neuroimaging Initiative (2014) Basal forebrain atrophy and cortical amyloid deposition in nondemented elderly subjects. *Alzheimers Dement* **10**, S344-S353.

[51] Teipel S, Heinsen H, Amaro E Jr, Grinberg LT, Krause B, Grothe M, Alzheimer's Disease Neuroimaging Initiative (2014) Cholinergic basal forebrain atrophy predicts amyloid burden in Alzheimer's disease. *Neurobiol Aging* **35**, 482-491.

[52] Grothe M, Heinsen H, Teipel S (2013) Longitudinal measures of cholinergic forebrain atrophy in the transition from healthy aging to Alzheimer's disease. *Neurobiol Aging* **34**, 1210-1220.

[53] Ashburner J (2007) A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95-113.

[54] Clark CM, Schneider JA, Bedell BJ, Beach TG, Bilker WB, Mintun MA, Pontecorvo MJ, Hefti F, Carpenter AP, Flitter ML, Krautkramer MJ, Kung HF, Coleman RE, Doraiswamy PM, Fleisher AS, Sabbagh MN, Sadowsky CH, Reiman EP, Zehntner SP, Skovronsky DM, AV45-AV07 Study Group (2011) Use of florbetapir-PET for imaging beta-amyloid pathology. *JAMA* **305**, 275-283.

[55] Fleisher AS, Chen K, Liu X, Roontiva A, Thiyyagura P, Ayutyanont N, Joshi AD, Clark CM, Mintun MA, Pontecorvo MJ, Doraiswamy PM, Johnson KA, Skovronsky DM, Reiman EM (2011) Using positron emission tomography and florbetapir F18 to image cortical amyloid in patients with mild cognitive impairment or dementia due to Alzheimer disease. *Arch Neurol* **68**, 1404-1411.

[56] Koch W, Teipel S, Mueller S, Benninghoff J, Wagner M, Bokde AL, Hampel H, Coates U, Reiser M, Meindl T (2012) Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiol Aging* **33**, 466-478.

[57] Ewers M, Frisoni GB, Teipel SJ, Grinberg LT, Amaro E Jr, Heinsen H, Thompson PM, Hampel H (2011) Staging Alzheimer's disease progression with multimodality neuroimaging. *Prog Neurobiol* **95**, 535-546.

[58] Ewers M, Sperling RA, Klunk WE, Weiner MW, Hampel H (2011) Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends Neurosci* **34**, 430-442.

[59] Dyrba M, Grothe MJ, Kirste T, Teipel SJ (2015) Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum Brain Mapp* **36**, 2118-2131.

[60] Teipel S, Drzezga A, Grothe MJ, Barthel H, Chételat G, Schuff N, Skudlarski P, Cavedo E, Frisoni GB, Hoffmann W, Thyrian JR, Fox C, Minoshima S, Sabri O, Fellgiebel A (2015) Multimodal imaging in Alzheimer's disease: Validity and usefulness for early detection. *Lancet Neurol* **14**, 1037-1053.

[61] Teipel SJ, Bokde AL, Meindl T, Amaro E Jr, Soldner J, Reiser MF, Herpertz SC, Möller H-J, Hampel H (2010) White matter microstructure underlying default mode network connectivity in the human brain. *Neuroimage* **49**, 2021-2032.

[62] Bokde AL, Ewers M, Hampel H (2009) Assessing neuronal networks: Understanding Alzheimer's disease. *Prog Neurobiol* **89**, 125-133.

[63] Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD (2012) Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex* **22**, 158-165.

[64] Chan D, Fox NC, Scahill RI, Crum WR, Whitwell JL, Leschziner G, Rossor AM, Stevens JM, Cipolotti L, Rossor MN (2001) Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann Neurol* **49**, 433-442.

[65] Jenkins R, Fox NC, Rossor AM, Harvey RJ, Rossor MN (2000) Intracranial volume and Alzheimer disease: Evidence against the cerebral reverse hypothesis. *Arch Neurol* **57**, 220-224.

[66] Müller MJ, Greverus D, Weibrich C, Dellani PR, Scheurich A, Stoeter P, Fellgiebel A (2007) Diagnostic utility of hippocampal size and mean diffusivity in amnestic MCI. *Neurobiol Aging* **28**, 398-403.

[67] Jagust WJ, Landau SM, Alzheimer's Disease Neuroimaging Initiative (2012) Apolipoprotein E, not fibrillar β-amyloid,

reduces cerebral glucose metabolism in normal aging. *J Neurosci* **32**, 18227-18233.

[68] Barnes J, Ridgway GR, Bartlett J, Henley SM, Lehmann M, Hobbs N, Clarkson MJ, MacManus DG, Ourselin S, Fox NC (2010) Head size, age and gender adjustment in MRI studies: A necessary nuisance? *Neuroimage* **53**, 1244-1255.

[69] Chételat G, La Joie R, Villain N, Perrotin A, de La Sayette V, Eustache F, Vandenberghe R (2013) Amyloid imaging in cognitively normal individuals, at-risk populations and preclinical Alzheimer's disease. *Neuroimage Clin* **2**, 356-365.

[70] Hu Y, Xu Q, Li K, Zhu H, Qi R, Zhang Z, Lu G (2013) Gender differences of brain glucose metabolic networks revealed by FDG-PET: Evidence from a large cohort of 400 young adults. *PLoS One* **8**, e83821.

[71] Hsieh TC, Lin WY, Ding HJ, Sun SS, Wu YC, Yen KY, Kao CH (2012) Sex- and age-related differences in brain FDG metabolism of healthy adults: An SPM analysis. *J Neuroimaging* **22**, 21-27.

[72] Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, García Marquéz JR, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27-46.

[73] Knopman DS, Jack CR Jr, Wiste HJ, Weigand SD, Vemuri P, Lowe V, Kantarci K, Gunter JL, Senjem ML, Ivnik RJ, Roberts RO, Boeve BF, Petersen RC (2012) Short-term clinical outcomes for stages of NIA-AA preclinical Alzheimer disease. *Neurology* **78**, 1576-1582.

[74] Jack CR Jr, Knopman DS, Weigand SD, Wiste HJ, Vemuri P, Lowe V, Kantarci K, Gunter JL, Senjem ML, Ivnik RJ, Roberts RO, Rocca WA, Boeve BF, Petersen RC (2012) An operational approach to National Institute on Aging-Alzheimer's Association criteria for preclinical Alzheimer disease. *Ann Neurol* **71**, 765-775.

[75] Ojala M, Garriga GC (2010) Permutation tests for studying classifier performance. *J Mach Learn Res* **11**, 1833-1863.

[76] Koller D, Friedman N (2009) *Probabilistic graphical models: Principles and techniques*, MIT Press, Cambridge, MA.

[77] Li X, Zhao T, Yuan X, Liu H (2015) The flare Package for High Dimensional Linear Regression and Precision Matrix Estimation in R. *J Mach Learn Res* **16**, 553-557.

[78] Rey A (1964) *L'examen clinique en psychologie*, Presses universitaires de France.

[79] Schmidt M (1996) *Rey Auditory Verbal Learning Test: RAVLT: A Handbook*, Western Psychological Services.

[80] Bischof GN, Jessen F, Fliessbach K, Dronse J, Hammes J, Neumaier B, Onur O, Fink GR, Kukolja J, Drzezga A, van Eimeren T (2016) Impact of tau and amyloid burden on glucose metabolism in Alzheimer's disease. *Ann Clin Transl Neurol* **3**, 934-939.

[81] Li Y, Rinne JO, Mosconi L, Pirraglia E, Rusinek H, DeSanti S, Kemppainen N, Någren K, Kim B-C, Tsui W, de Leon MJ (2008) Regional analysis of FDG and PIB-PET images in normal aging, mild cognitive impairment, and Alzheimer's disease. *Eur J Nucl Med Mol Imaging* **35**, 2169-2181.

[82] Cohen AD, Price JC, Weissfeld LA, James J, Rosario BL, Bi W, Nebes RD, Saxton JA, Snitz BE, Aizenstein HA, Wolk DA, DeKosky ST, Mathis CA, Klunk WE (2009) Basal cerebral metabolism may modulate the cognitive effects of Abeta in mild cognitive impairment: An example of brain reserve. *J Neurosci* **29**, 14770-14778.

[83] Sepulcre J, Schultz AP, Sabuncu M, Gomez-Isla T, Chhatwal J, Becker A, Sperling R, Johnson KA (2016) In vivo tau, amyloid, and gray matter profiles in the aging brain. *J Neurosci* **36**, 7364-7374.

[84] Huijbers W, Mormino EC, Schultz AP, Wigman S, Ward AM, Larvie M, Amariglio RE, Marshall GA, Rentz DM, Johnson KA, Sperling RA (2015) Amyloid-β deposition in mild cognitive impairment is associated with increased hippocampal activity, atrophy and clinical progression. *Brain* **138**, 1023-1035.

[85] Trzepacz PT, Hochstetler H, Yu P, Castelluccio P, Witte MM, Dell'Agnello G, Degenhardt EK, Alzheimer's Disease Neuroimaging Initiative (2016) Relationship of hippocampal volume to amyloid burden across diagnostic stages of Alzheimer's disease. *Dement Geriatr Cogn Disord* **41**, 68-79.

[86] Araque Caballero MÁ, Brendel M, Delker A, Ren J, Rominger A, Bartenstein P, Dichgans M, Weiner MW, Ewers M, Alzheimer's Disease Neuroimaging Initiative (ADNI) (2015) Mapping 3-year changes in gray matter and metabolism in Aβ-positive nondemented subjects. *Neurobiol Aging* **36**, 2913-2924.

[87] Teipel SJ, Kurth J, Krause B, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative (2015) The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment—Beyond classical regression. *Neuroimage Clin* **8**, 583-593.

[88] Hastie TJ, Tibshirani RJ, Friedman JH (2013) *The elements of statistical learning: Data mining, inference, and prediction*, Springer, New York, NY.

[89] Kim DJ, Skosnik PD, Cheng H, Pruce BJ, Brumbaugh MS, Vollmer JM, Hetrick WP, O'Donnell BF, Sporns O, Puce A, Newman SD (2011) Structural network topology revealed by white matter tractography in cannabis users: A graph theoretical analysis. *Brain Connect* **1**, 473-483.

[90] Teipel SJ, Stahl R, Dietrich O, Schoenberg SO, Perneczky R, Bokde ALW, Reiser MF, Möller HJ, Hampel H (2007) Multivariate network analysis of fiber tract integrity in Alzheimer's disease. *Neuroimage* **34**, 985-995.

[91] Lee JD, Hastie TJ (2015) Learning the structure of mixed graphical models. *J Comp Graph Stat* **24**, 230-253.

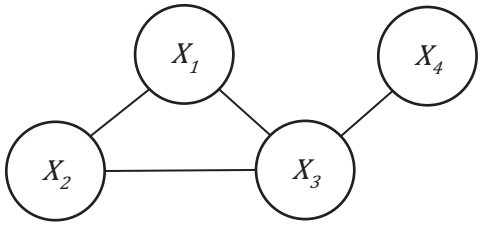
[92] MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. *Psychol Methods* **7**, 19-40.

[93] Dyrba M, Grothe M, Kirste T, Teipel SJ (2015) Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Human Brain Mapp* **36**, 2118-2131.

Supplementary Material

A Foundations of Markov random fields

Probabilistic graphical models combine probability theory and graph theory towards an intuitive and powerful formalism for modeling and solving inference problems in various scientific and engineering fields. The structure of Markov random fields is defined by an undirected graph $= (V, E)$, with the set of nodes $\{X_1, X_2, \dots, X_n\} \in V$ representing random variables, and the set of undirected edges $E = \{\{X_{i1}, X_{j1}\}, \dots, \{X_{im}, X_{jm}\}\}$ representing statistical associations between each pair of variables (Supplementary Figure 1). The model holds an independence assumption, commonly referred to as *local Markov property*, which imposes that a variable is conditionally independent from any other variable given its direct neighbors. For example, in Supplementary Figure 1, node X_1 is conditionally independent from X_4 given its neighbors X_2 and X_3, formally $X_1 \perp X_4 | X_2, X_3$, while node X_3 is conditionally dependent on the three other nodes X_1, X_2 and X_4.



Supplementary Figure 1. Simple Markov random field model.
Vertices X_i represent random variables and edges represent statistical associations between two variables.

A.1 Classical Markov random field for discrete data

A Markov random field can be written as a log-linear model $P(X_1, \dots, X_n) = \exp(\sum_{i=1}^{n} w_i f_i)/Z$, with $P(X_1, \dots, X_n)$ being the joint probability of finding that the random variables X_i take on particular values. In general, the feature functions can represent arbitrary types of associations. But here, for reasons of convenience, the feature functions shall be limited to represent either unary indicator functions $f_l \in \{0,1\}$ representing the discrete state of a random variable X_i (0=pathologic, 1=normal) or binary indicator functions representing value combinations of two connected nodes X_i and X_j. The weights $w_l \in \mathbb{R}$ set the model parameters that are being estimated by the model fitting algorithm and which subsequently can be converted to provide probabilities of individual variable's states or the statistical associations between two variables. In Supplementary Figure 1, let the random variables X_3 and X_4 each have two discrete states 0 and 1. Then, the edge $\{X_3, X_4\}$ is represented by four weights $w_{ab}, ab \in \{00,01,10,11\}$, that is one weight for each combination of states of the two nodes: $w_{ab} \sim P(X_i = a, X_j = b|G)$. In case of the edge $\{X_3, X_4\}$, the weights might correspond to the empirical contingency table for the two random variables X_3 and X_4 as the node X_4 is separated from the rest of the graph. Note that in general the weights deviate from just calculating frequencies of observations in the training data as the whole graphical network is taken into account by the parameter estimation algorithm. In particular, the parameter estimation algorithm infers the partial correlation of two connected variables given their direct neighbors [1, 2]. The partition function $Z = \sum_{X_1,\dots,X_n} \exp(-\sum_{l=1}^{n} w_l f_l)$ is the sum over all possible assignments of values to the network's random variables $X_1, \dots, X_n$ and is used as normalization constant. The partition function Z can be seen as the computational bottleneck of the algorithm as it needs to iterate over all possible realizations of random variables, which is in general computationally expensive. However, modern computers provide enough resources to analyze the relatively small networks that can be derived for multimodal imaging data in a region of interest approach. Common tasks and applications for Markov random fields include (i) modeling of the graph structure, obtained, for instance, from prior knowledge or hypotheses, (ii) parameter learning of the weights w_l using an iterative optimization algorithm maximizing the model likelihood for a given training dataset, and (iii) inference to obtain the probabilities for the network or partial configurations for another dataset. A more extended formal definition of Markov random fields can be found in Wang, et al. [3] and Koller and Friedman [4, Ch. 4].

A.2 Gaussian graphical models for continuous normal data

An alternative class of models sometimes referred to as Gaussian Markov random fields or more commonly Gaussian graphical models assess the interaction between several normally distributed variables: $p(\mathrm{x}) \propto \exp(-\frac{1}{2}(\mathrm{x}-\mu)^T\Sigma^{-1}(\mathrm{x}-\mu))$. Therefore, the inverse covariance matrix $\Theta = \Sigma^{-1}$ is estimated, commonly called information or precision matrix. If a pair of variables conditionally independent, that means when the partial correlation between the two variables is zero when correcting for all the other variables, this will be reflected as a zero entry in the precision matrix [4, Ch. 7.1]. The objective function for being minimized is given as $l(\Theta) = \log\det\Theta - \mathrm{trace}(\hat{\Sigma}\Theta)$, with the log-likelihood of the data l, and the sample covariance matrix $\hat{\Sigma}$ [5, Ch. 17.3.1, Formula 17.11]. To overcome the problem of dimensionality when there are only few samples available in comparison to a much higher number of variables, Gaussian graphical models typically apply a sparsity assumption and only determine the most likely interactions between variables. When considering large networks with a high number of variables, matrix inversion may be much too costly and vulnerable to noise in the specific sample [6-8], such that regularization approaches are commonly used to apply the sparsity constraint on the entries of the precision matrix [6-8]. For example, the *graphical lasso* employs the objective function $l(\Theta) = \log\det\Theta - \mathrm{trace}(\hat{\Sigma}\Theta) - \lambda\|\Theta\|_1$, with $\|\Theta\|_1$ being the L_1 norm of Θ, that is the sum of absolute values of the elements of Θ, and the penalty parameter $\lambda \in \mathbb{R}^+$ that controls sparsity [5, Ch. 17.3.2, Formula 17.21]. The negative of $l(\Theta)$ provides a convex optimization problem that is efficiently solvable using iterative optimization approaches, such as gradient descent algorithms [5, Ch. 17.3.2]. These techniques have been applied previously for genomic data analysis [8, 9] and functional connectivity analysis based on resting-state functional MRI [10]. A more extended formal definition of Markov random fields can be found in Hastie, et al. [5, Ch. 17.3] and Koller and Friedman [4, Ch. 7.3].

B Markov random field modeling

The Markov random field analyses were performed using the *Undirected Graphical Modeling* (UGM) toolbox (*http://www.cs.ubc.ca/~schmidtm/Software/UGM.html*, release 2013) implemented in MATLAB R2013a (MathWorks, Natick, MA, USA). UGM allowed the modeling of arbitrary discrete undirected graphical models and implemented various exact and approximate algorithms for parameter estimation (training) and inference. The weight vector was estimated using the iterative quasi-Newton gradient-descending limited-memory *Broyden–Fletcher–Goldfarb–Shanno* algorithm [11], an expectation–maximization algorithm that was applied for each of the different Markov random field models individually. To assess the stability of the model likelihood and individual parameters, we used tenfold cross-validation approach with 1,000 repetitions. The algorithm estimated both the distribution of states for single nodes, as well as the associations between connected nodes by adjusting the edge weights, such that the model likelihood was maximized with respect to the observed correlation patterns of the training data. In this approach, maximum likelihood estimation represented a convex optimization problem such that the iterative model fitting algorithm converged at the optimal set of parameters [4, p.948]. To reduce the number of model parameters, we simplified the edge weights to represent the probability of concordance and discordance of the two connected nodes instead of using the full conditional probability table. To summarize the estimated concordance/discordance between two nodes, we used the phi coefficient r_ϕ, which is mathematically identical to the Pearson correlation coefficient for two binary variables. Finally, the model likelihood was obtained for the independent test data. To compare the model fit we calculated the differences of log-likelihood for the validation proportion of the data for each iteration of the cross-validation. Therefore, we ensured that the cross-validation training/test sets were identical for all analyses. Using the difference of the log-likelihood (equal to the ratio of likelihood) is commonly used in the likelihood ratio test, which employs the test statistic $D = -2log(L_{M2}/L_{M1}) = -2(log(L_{M2}) - log(L_{M1}))$, with the likelihood L for two models M_1 and M_2. In our case we did not estimate the level of confidence for D using the chi-square distribution, but directly took the proportion of cross-validation iterations with $D > 0$ as an empirical P-value [12].

References

[1] Ravikumar P, Wainwright MJ, Lafferty JD (2010) High-dimensional Ising model selection using ℓ 1 -regularized logistic regression. *Ann Stat* **38**, 1287-1319.

[2] Lee JD, Hastie TJ (2015) Learning the structure of mixed graphical models. *J Comp Graph Stat* **24**, 230-253.

[3] Wang C, Komodakis N, Paragios N (2013) Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey. *Comput Vis Image Underst* **117**, 1610-1627.

[4] Koller D, Friedman N (2009) *Probabilistic graphical models: Principles and techniques*, MIT Press, Cambridge, MA.

[5] Hastie TJ, Tibshirani RJ, Friedman JH (2013) *The elements of statistical learning: Data mining, inference, and prediction*, Springer, New York, NY.

[6] Ravikumar P, Wainwright MJ, Raskutti G, Yu B (2011) High-dimensional covariance estimation by minimizing ℓ1-penalized log-determinant divergence. *Electron J Stat* **5**, 935-980.

[7] Cai TT, Ren Z, Zhou HH (2016) Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron J Stat* **10**, 1-59.

[8] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

[9] Mohammadi A, Wit EC (2015) Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal* **10**, 109-138.

[10] Ryali S, Chen T, Supekar K, Menon V (2012) Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *Neuroimage* **59**, 3852-3861.

[11] Kelley CT (1999) *Iterative methods for optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[12] Ojala M, Garriga GC (2010) Permutation tests for studying classifier performance. *J Mach Learn Res* **11**, 1833-1863.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fnagi.2020.00099

Gaussian Graphical Models Reveal Inter-Modal and Inter-Regional Conditional Dependencies of Brain Alterations in Alzheimer's Disease

Martin Dyrba[1], Reza Mohammadi[2]*, Michel J. Grothe[1], Thomas Kirste[3], and Stefan J. Teipel[1,4] on behalf of the Alzheimer's Disease Neuroimaging Initiative*

[1] German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany, [2] Department of Operation Management, Amsterdam Business School, University of Amsterdam, Amsterdam, Netherlands, [3] Mobile Multimedia Information Systems Group (MMIS), University of Rostock, Rostock, Germany, [4] Clinic for Psychosomatics and Psychotherapeutic Medicine, Rostock University Medical Center, Rostock, Germany

Alzheimer's disease (AD) is characterized by a sequence of pathological changes, which are commonly assessed *in vivo* using various brain imaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET). Currently, the most approaches to analyze statistical associations between regions and imaging modalities rely on Pearson correlation or linear regression models. However, these models are prone to spurious correlations arising from uninformative shared variance and multicollinearity. Notably, there are no appropriate multivariate statistical models available that can easily integrate dozens of multicollinear variables derived from such data, being able to utilize the additional information provided from the combination of data sources. Gaussian graphical models (GGMs) can estimate the conditional dependency from given data, which is conceptually expected to closely reflect the underlying causal relationships between various variables. Hence, we applied GGMs to assess multimodal regional brain alterations in AD. We obtained data from $N = 972$ subjects from the Alzheimer's Disease Neuroimaging Initiative. The mean amyloid load (AV45-PET), glucose metabolism (FDG-PET), and gray matter volume (MRI) were calculated for each of the 108 cortical and subcortical brain regions. GGMs were estimated using a Bayesian framework for the combined multimodal data and the resulted conditional dependency networks were compared to classical covariance networks based on Pearson correlation. Additionally, graph-theoretical network statistics were calculated to determine network alterations associated with disease status. The resulting conditional dependency matrices were much sparser ($\approx$ 10% density) than Pearson correlation matrices ($\approx$ 50% density). Within imaging modalities, conditional dependency networks yielded clusters connecting anatomically adjacent regions. For the associations between different modalities, only few region-specific connections were detected. Network measures such as small-world coefficient were significantly altered across diagnostic groups, with a biphasic u-shape trajectory, i.e., increased small-world coefficient in early mild cognitive impairment (MCI), similar values in late MCI, and decreased values in AD dementia patients compared to cognitively normal controls. In conclusion, GGMs removed commonly

OPEN ACCESS

Edited by:
Muthuraman Muthuraman, University Medical Center of the Johannes Gutenberg University Mainz, Germany

Reviewed by:
Paul Gerson Unschuld, University of Zurich, Switzerland
Yeo Jin Kim, Hallym University, South Korea

***Correspondence:**
Martin Dyrba
martin.dyrba@dzne.de
Reza Mohammadi
a.mohammadi@uva.nl

Received: 28 October 2019
Accepted: 24 March 2020
Published: 21 April 2020

Citation:
Dyrba M, Mohammadi R, Grothe MJ, Kirste T and Teipel SJ (2020) Gaussian Graphical Models Reveal Inter-Modal and Inter-Regional Conditional Dependencies of Brain Alterations in Alzheimer's Disease. Front. Aging Neurosci. 12:99. doi: 10.3389/fnagi.2020.00099

shared variance among multimodal measures of regional brain alterations in MCI and AD, and yielded sparser matrices compared to correlation networks based on the Pearson coefficient. Therefore, GGMs may be used as alternative to thresholding-approaches typically applied to correlation networks to obtain the most informative relations between variables.

Keywords: Alzheimer's disease, mild cognitive impairment, conditional dependency networks, Gaussian graphical models, graph-theoretical analysis, small-world network

1. INTRODUCTION

Alzheimer's disease (AD) is characterized by a range of pathological brain alterations that can be assessed *in vivo* using various neuroimaging methods, including MRI and PET. Several studies suggest that information obtained from combining different imaging modalities could provide reliable markers of cerebral reserve capacity and might be used to predict and monitor the evolution of AD and its relative impact on cognitive domains in pre-clinical, prodromal, and dementia stages of AD [see e.g., reviews (Teipel S. et al., 2015; Teipel et al., 2016)]. However, there is still an unmet need for appropriate analysis methods for assessing statistical associations between individual brain regions and between different pathology markers derived from multiple neuroimaging modalities.

Up to date, multimodal studies employ one of the following approaches:

(i) Correlation of pathology maps on a voxel level (La Joie et al., 2012; Altmann et al., 2015; Grothe and Teipel, 2016);
(ii) linear regression analysis with a-priori specified regions-of-interest (Buckner et al., 2005; Seeley et al., 2009; Villain et al., 2010; Kljajevic et al., 2014; Chang et al., 2015; Grothe et al., 2016; Teipel and Grothe, 2016);
(iii) stratification of subjects into distinct groups (e.g., amyloid positive/negative) to compare differences in other imaging modalities (Buckner et al., 2005; Kljajevic et al., 2014; Grothe et al., 2016);
(iv) comparison of graph-theoretical measures and statistics between modalities (Stam et al., 2006; Buckner et al., 2009; Zhou et al., 2012; Sepulcre et al., 2013, 2017); and
(v) estimation of generative models for comparing spreading mechanisms of amyloid-β deposition and its contribution to neurodegeneration (Dyrba et al., 2017; Iturria-Medina et al., 2017; Torok et al., 2018).

Commonly employed statistical models, such as linear regression analysis, provide limited ability to assess the interactions between dozens of variables in the same model, as they cannot derive reliable estimates regarding the individual contribution of highly collinear predictors and suffer from variance inflation (Dormann et al., 2013). Calculation of covariance/connectivity matrices based on the Pearson correlation between each pair of variables has led to practical problems in deriving meaningful results, i.e., these matrices are commonly thresholded to an a-priori defined density and binarized (Buckner et al., 2009; Zhou et al., 2012; Sepulcre et al., 2013). More recently, summary statistics based on graph-theory have been proposed (Watts and Strogatz, 1998; Stam et al., 2006) and are currently widely applied (Buckner et al., 2009; Zhou et al., 2012; Sepulcre et al., 2013, 2017). However, this approach has been criticized, as for instance, group differences in small-worldness of the brain network might be sensitive to the specific density threshold (Hlinka et al., 2017; Mårtensson et al., 2018).

We suggest the application of Gaussian graphical models (GGMs), which are able to estimate the *partial* correlation between various multicollinear predictors (Hastie et al., 2013, chapter 7.3). GGMs yield sparse conditional dependency matrices, that are conceptually expected to closer reflect the underlying causal relationships (Koller and Friedman, 2009, chapter 21.7; Bontempi and Flauder, 2015). This makes GGMs an interesting candidate for studying properties of the brain network; an example is illustrated in **Figure 1**. The partial correlation derived from GGMs is conceptually similar to the partial correlation obtained from a series of linear regression models, which estimate the statistical association of the dependent and independent variables while controlling for the confounding variables. Additionally, GGMs extend this concept by estimating the partial correlation matrix as a set of coupled regression problems, in contrast to separate regression problems modeled by traditional linear regression (Meinshausen and Bühlmann, 2006; Hastie et al., 2013, chapter 7.3). Technically, GGMs are naively realized by matrix inversion of the covariance matrix. In more robust and efficient approaches, regularization techniques (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011; Ryali et al., 2012; Cai et al., 2013; Wang et al., 2016) or efficient sampling schemes (Mohammadi and Wit, 2015, 2019) are applied.

In this paper, we tested the applicability and clinical utility of GGMs to reveal the conditional dependency structure between regional pathology measures. For this purpose, we assessed inter-regional statistical associations within and between three main imaging markers of Alzheimer's disease using GGMs based on a whole-cortex parcellation of the brain. The assessed imaging markers included amyloid-β deposition (florbetapir/AV45-PET), glucose metabolism (FDG-PET), and gray matter volume (T_1-weighted MRI). Based on our previous results with only six representative brain regions (Dyrba et al., 2017), we hypothesized that regional amyloid deposition has low contribution to gray matter atrophy, whereas hypometabolism was expected to be stronger related to atrophy. Further, we expected a few hub-nodes influencing pathology in other regions. For graph-theoretical measures, we expected a linear trajectory of decreasing clustering

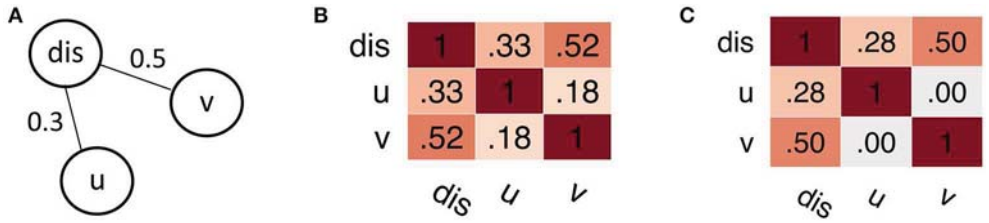


FIGURE 1 | Simple example for spurious correlations. **(A)** True dependency graph. The node *u* is statistically independent from *v* given the node *dis*, formally $p(u, v|dis) = p(u|dis)p(v|dis)$. **(B)** Pearson correlation matrix, showing a "spurious" correlation between nodes u and v. Notably, when considering only *u* and *v* alone, the independence assumption does not hold; formally $p(u, v) \neq p(u)p(v)$. **(C)** Partial correlation matrix derived from Gaussian graphical models. Using this model, we can approximately recover the underlying dependency structure, with $u \perp v|dis \implies cor(u, v|dis) = 0$.

coefficient and increasing path length with stronger disease severity, as previously reported in the literature for connectivity analyses based on Pearson correlation (He et al., 2008; Yao et al., 2010; Li et al., 2012; Morbelli et al., 2012; Tijms et al., 2013; Pereira et al., 2016; John et al., 2017; Titov et al., 2017).

2. MATERIALS AND METHODS

2.1. Study Participants

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations, with the primary goal of testing whether neuroimaging, neuropsychological, and other biological measurements can be used as reliable *in vivo* markers of Alzheimer's disease pathogenesis. A complete description of ADNI and up-to-date information is available at http://www.adni-info.org. For this study, 529 subjects with amnestic mild cognitive impairment (MCI), 189 patients with Alzheimer's dementia (AD), and 254 cognitively healthy control subjects (CN) were selected from the ADNI-GO, ADNI-2, and ADNI-3 extensions of the ADNI project, based on the availability of concurrent structural MRI, FDG-PET, amyloid-sensitive AV45-PET, and neuropsychological assessments. In ADNI, two MCI subgroups exist, which only differ by the less severe impairment of memory function for *early MCI* (EMCI) compared to *late MCI* (LMCI) subjects. Detailed inclusion criteria for the diagnostic categories can be found at the ADNI website (http://adni.loni.usc.edu/methods, ADNI2 manual page 27). Demographics and neuropsychological profiles of the different diagnostic groups are summarized in **Table 1**.

TABLE 1 | Sample characteristics.

	CN	EMCI	LMCI	AD
Sample size (female)	254(130)	309(135)	220(93)	189(80)
Age (SD)	75.4 ± 6.6	71.6 ± 7.5*	74.1 ± 8.1	75.0 ± 8.0
Education (SD)	16.4 ± 2.7	16.0 ± 2.6	16.2 ± 2.8	15.9 ± 2.7
MMSE (SD)	29.1 ± 1.2	28.3 ± 1.6*	27.6 ± 1.9*	22.6 ± 3.2*
Delayed recall (SD)	7.6 ± 4.1	5.7 ± 4.0*	3.2 ± 3.7*	0.8 ± 1.9*

Gender distribution did not differ significantly between groups ($P = 0.15$, chi-square test). Asterisks indicate significant difference between groups ($P < 0.05$) based on pairwise two-sample t-test with CN as reference group. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; MMSE, Mini-Mental State Examination; delayed recall, number of remembered words out of a 15-item wordlist of the Rey Auditory Verbal Learning Test.

2.2. Imaging Data and Feature Extraction

ADNI-GO/2 MRI, FDG- and AV45-PET data were downloaded from the ADNI image archive. ADNI-GO/2 MRI data were acquired on multiple 3T MRI scanners using scanner-specific T1-weighted sagittal 3D MP-RAGE/IR-SPGR sequences. To increase signal uniformity across the multicenter scanner platforms, original T1 acquisitions underwent standardized image preprocessing correction steps (http://adni.loni.usc.edu/methods/mri-tool/mri-pre-processing/). FDG- and AV45-PET data were acquired on multiple instruments of varying resolution and following different platform-specific acquisition protocols. Similar to the MRI data, PET data in ADNI were also subject to standardized image preprocessing correction steps, with the aim of increasing data uniformity across the multicenter acquisitions (http://adni.loni.usc.edu/methods/pet-analysis-method/pet-analysis/). Imaging data were processed by using statistical parametric mapping (SPM8, Wellcome Centre for Human Neuroimaging, University College London) and the VBM8 toolbox (Structural Brain Mapping Group, University of Jena) implemented in MATLAB R2013b (Math-Works, Natick, MA) as previously described in Grothe et al. (2016) and Grothe and Teipel (2016). First, MRI T1 scans were segmented into gray matter, white matter, and cerebrospinal fluid partitions using the segmentation routine of the VBM8 toolbox. Then, the resulting gray matter and white matter segments were spatially normalized to an aging/AD-specific reference template (Grothe et al., 2013) using the DARTEL algorithm. Additionally, voxel values of the normalized gray matter segments were modulated for volumetric changes introduced by the high-dimensional normalization, such that the total amount of gray matter volume present before warping was preserved. Each subject's FDG- and AV45-PET scans were rigidly coregistered to the corresponding skull-stripped T1 scan. Then, the PET scans were corrected for partial volume effects using a three-compartment model

and the MRI-derived tissue segments (Müller-Gärtner et al., 1992; Gonzalez-Escamilla et al., 2017). Corrected PET scans were spatially normalized (without modulation) by applying the deformation fields of the T1-weighted scans. All original data and normalized scans were visually inspected to ensure a high quality of the data. Subsequently, mean gray matter volumes and mean FDG-/AV45-PET uptake values were calculated for 108 cortical and subcortical regions defined by the Harvard-Oxford atlas (Desikan et al., 2006) after projecting the atlas to the aging/AD-specific reference space and removing voxels with a gray matter probability of <50% in the aging/AD template. Finally, regional gray matter volumes were proportionally scaled by total intracranial volume (TIV), regional FDG-PET values were proportionally scaled to pons uptake, and regional AV45-PET values were proportionally scaled to whole-cerebellum uptake. To be able to directly compare the different modalities, all regional values were normalized using the congitively normal subjects as reference group (La Joie et al., 2012). As described previously (Dyrba et al., 2017), we used the so-called *W*-scores, which are analogous to *Z*-scores but are adjusted for specific covariates; age, gender, and education in the present case. Like *Z*-scores, *W*-scores have a mean value of 0 and a standard deviation of 1 in the control group, and values of +1.65 and −1.65 correspond to the 95th and 5th percentiles, respectively. To calculate the *W*-scores, regression models were estimated for the control group using age, gender, and education as independent variables and the mean value of each region as dependent variable. Then, *W*-scores were computed using $W = (x_{ij} - e_{ij})/s_{res,j}$; with x_{ij} being the *i*th subject's raw value for region *j*; e_{ij} being the value expected for region *j* in the control group for the *i*th subject's age, gender, and education; and $s_{res,j}$ being the standard deviation of the residuals for region *j* in controls.

2.3. Statistical Modeling

Graphical models provide an effective way for describing statistical patterns in multivariate data and for estimating the conditional dependency between the various brain regions and imaging modalities based on GGMs (Lauritzen, 1996; Mohammadi and Wit, 2015). For data following a multivariate normal distribution, undirected GGMs are commonly used. In these graphical models, the graph structure is directly characterized by the precision matrix, i.e., the inverse of the covariance matrix: non-zero entries in the precision matrix show the edges in the conditional dependency graph. Notably, simple inversion of the covariance matrix usually does not work in real world data sets, as already slight noise in the empirical data causes the precision matrix to contain almost no zero entries. To overcome this problem, regularization techniques or efficient sampling algorithms have been proposed that reduce the effect of noise by additionally employing a sparsity assumption and, thus, only detect the most probable conditional dependencies. For our analyses, we employed a computationally efficient Bayesian framework implemented in the R package BDgraph. More specifically, this framework implements a continuous-time birth-death Markov process for estimating the most probable graph structure and edge weights that correspond to the observed partial correlations (Mohammadi and Wit, 2015, 2019). For this study, BDgraph was substantially extended by multi-threaded parallel processing and marginal pseudo-likelihood approximation to speed up computations.

2.4. Experimental Setup

First, we estimated GGMs based on the combined data of EMCI, LMCI, and AD patients to study the conditional dependency between brain regions and modalities. Second, we estimated GGMs for each diagnostic group separately to assess alterations

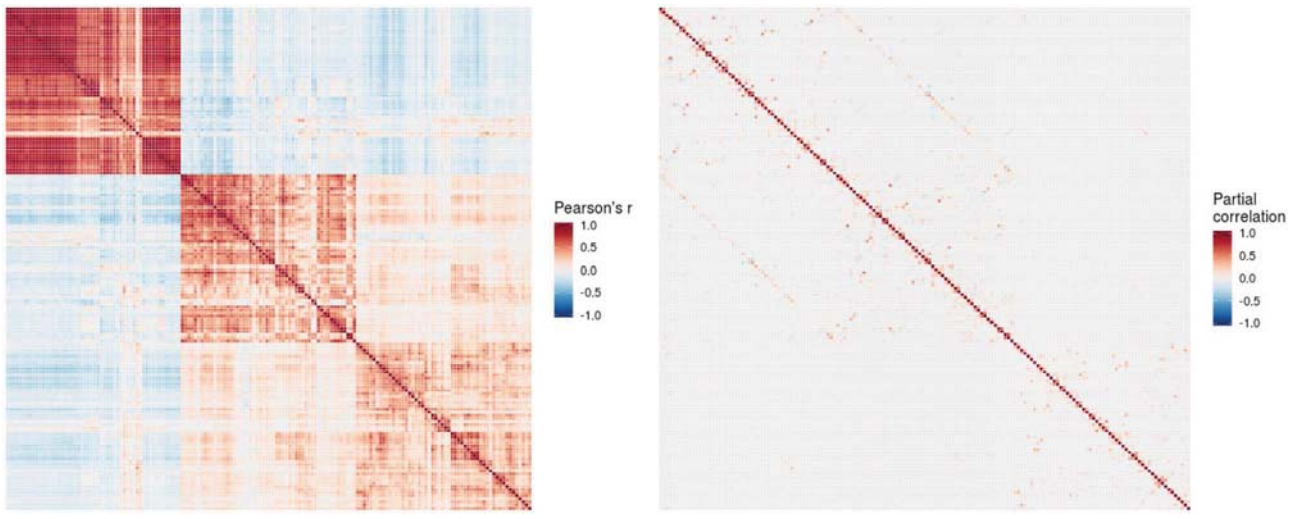


FIGURE 2 | Pearson correlation matrix **(left)** and partial correlation matrix **(right)** for the three imaging modalities (left hemisphere data only) estimated for the combined data of EMCI, LMCI, and AD patients. For better readability, each individual block of the partial correlation matrix is shown in **Figures 3**–**5** and **Figures S2–S4**. EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia.

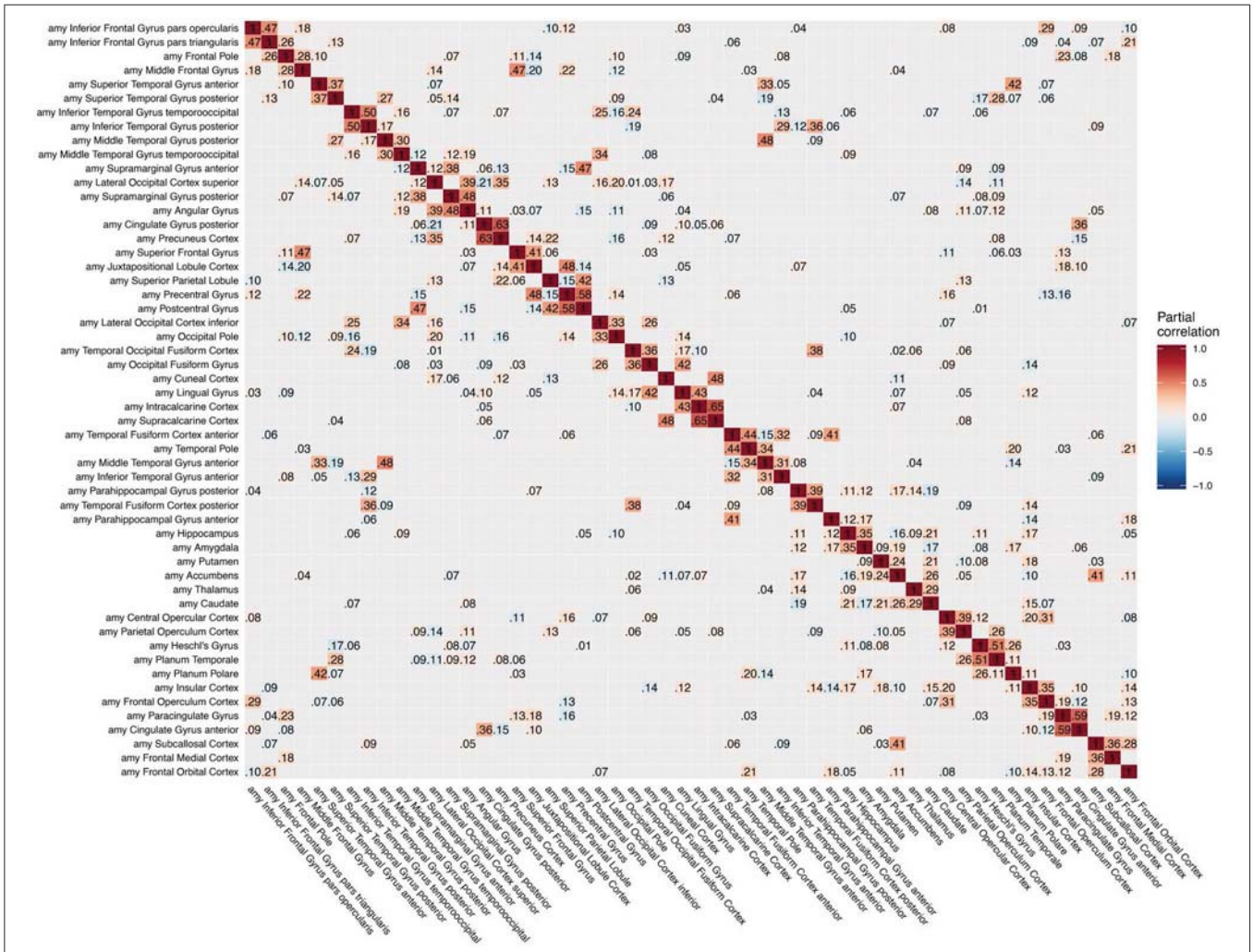


FIGURE 3 | Partial correlation matrix for amyloid-β deposition in the left hemisphere estimated for the combined data of EMCI, LMCI, and AD patients. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; amy, amyloid-β.

of the graph structures. For the combined model, regional W-scores of all MCI and AD patients ($N = 718$) and all three imaging modalities were entered. Initially, we took all 108 cortical and subcortical regions included in the Harvard-Oxford atlas (Desikan et al., 2006) into consideration, corresponding to $P = 3 * 108 = 324$ variables. The sampling process included 1,000,000 burn-in iterations[1], starting from a random estimate for the inverse covariance matrix and converging to estimates with higher posterior probability giving the training data. The burn-in iterations were then discarded, and subsequently 150,000 sampling iterations followed to obtain the estimates for the inverse covariance matrix. Because results were showing a strong left–right hemisphere symmetry, we repeated model estimation including only the 54 regions in the left hemisphere ($P = 3 * 54 = 162$ variables) to increase model stability. From the final model we set a probability threshold of $P_{avg} > 0.5$ for selecting the edges, with the notion that a specific edge was considered to be present if it existed in at least half of all model iterations (Madigan et al., 1996). For the second analysis of group differences, we estimated individual GGMs for each group based on the multimodal data of the left hemisphere. Sampling was again performed with 1,000,000 burn-in iterations followed by 150,000 sampling iterations.

For comparison, these analyses were also repeated (i) using data of the right hemisphere to validate the results and (ii) using the traditional approach of constructing correlation networks based on the Pearson correlation coefficient.

2.5. Graph-Theoretical Analyses

To assess group differences of the estimated graph structure we calculated the three graph-theoretical measures that are most commonly reported in the literature; clustering coefficient,

[1]For Markov chain Monte Carlo (MCMC) methods, burn-in refers to the practice of discarding an initial portion of the Markov chain sample, so that the chain can reach a stationary distribution. Thus, the effect of randomly chosen initial values on the posterior inference is minimized.

characteristic path length, and their ratio, the small-world coefficient. The path length quantifies the distance of connections between two nodes along the shortest path. The weighted characteristic path length is the average minimum distance between a node $i \in N$ and all other nodes, $L_i = \sum_{j \in N, j \neq i} d_{ij}/(n-1)$, where $d_{ij} = \sum_{a_{uv} \in g_{i \leftrightarrow j}} \omega_{uv}$ is the shortest weighted path length between i and j, $g_{i \leftrightarrow j}$ defines the shortest path, and ω_{uv} defines the distance between two nodes. Here, the distance matrix was defined as $\Omega = 1 - abs(\Theta)$, that is one minus the absolute pair-wise partial correlation as derived from the GGMs or the absolute Pearson coefficient, respectively (Rubinov and Sporns, 2010). The weighted clustering coefficient indicates the interconnectedness of neighboring nodes $C_i = 2t_i/(k_i(k_i-1))$, where $t_i = 0.5 \sum_{j,h \in N} (\omega_{ij}\omega_{ih}\omega_{jh})^{1/3}$ is the geometric mean of triangles around node i, and where $k_i = \sum_{j \in N} a_{ij}$ is the number of nodes connected to node i (Onnela et al., 2005; Rubinov and Sporns, 2010). k_i is often referred to as the *degree* of the node i, and the link status $a_{ij} = 1$ if node i is connected to another node j, or $a_{ij} = 0$ otherwise. The small-world coefficient is defined as the ratio of the clustering coefficient C and characteristic path length L in comparison to a random network, $S = (C/C_{rand})/(L/L_{rand})$, with $S \gg 1$ in small-world networks (Rubinov and Sporns, 2010). To simplify calculations, we omitted defining a random network to estimate C_{rand} and L_{rand}, and directly took the ratio $S_i = C_i/L_i$ for group comparisons. Notably, we later report the distribution of graph measures for single regions, as the dependency measures were derive from the whole group of subjects. Graph metrics were compared between diagnostic groups using analysis of variance (ANOVA) and Tukey's honest significant difference tests.

3. RESULTS

3.1. Conditional Dependency of Alzheimer's Pathology

The conditional dependency matrix obtained using the GGM approach for all region of the left hemisphere is given in **Figure 2** (right). For the partial correlation between all pairs of brain regions, we obtained 960 significant associations (7% network

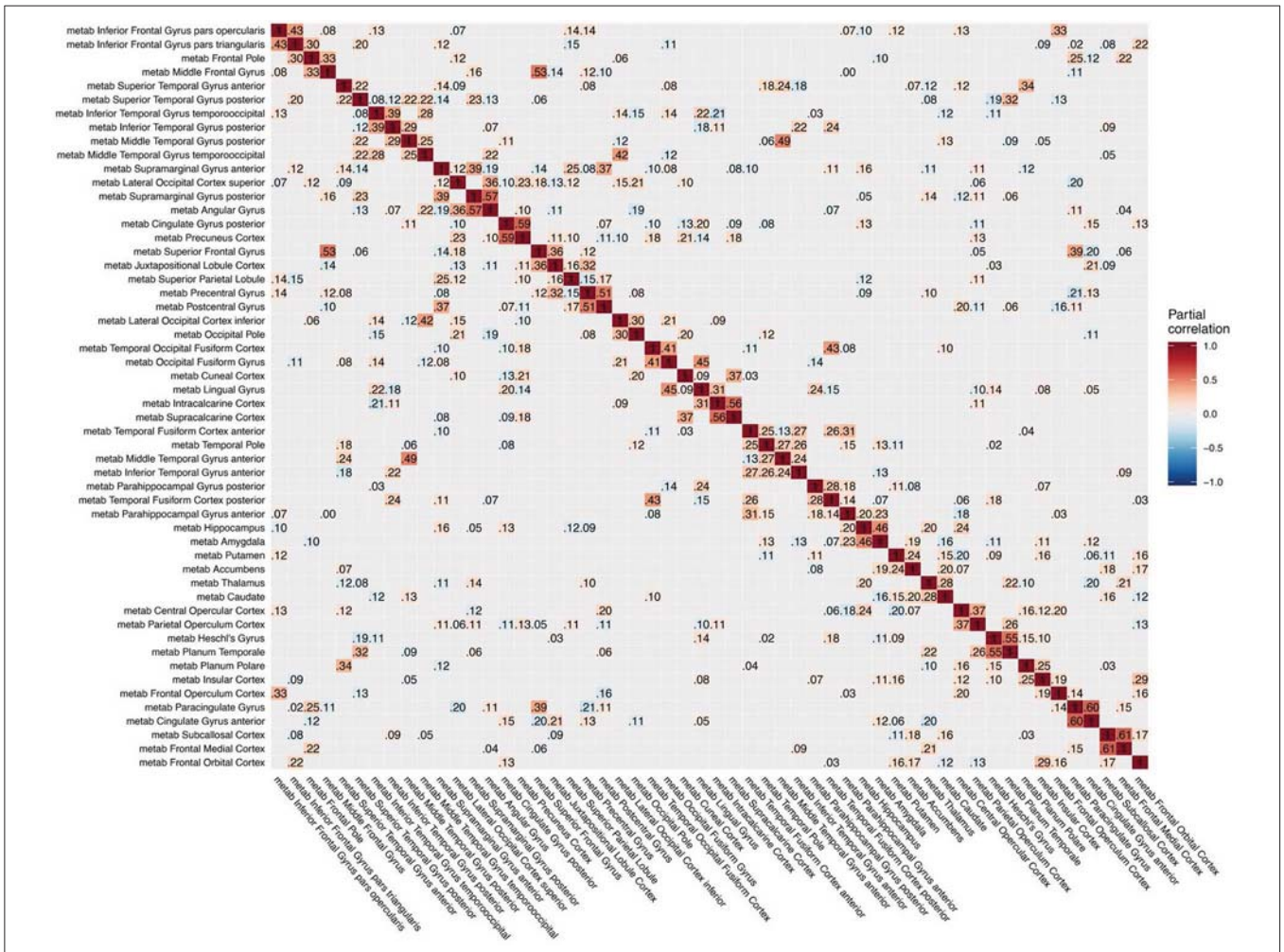


FIGURE 4 | Partial correlation matrix for glucose metabolism in the left hemisphere estimated for the combined data of EMCI, LMCI, and AD patients. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; metab, glucose metabolism.

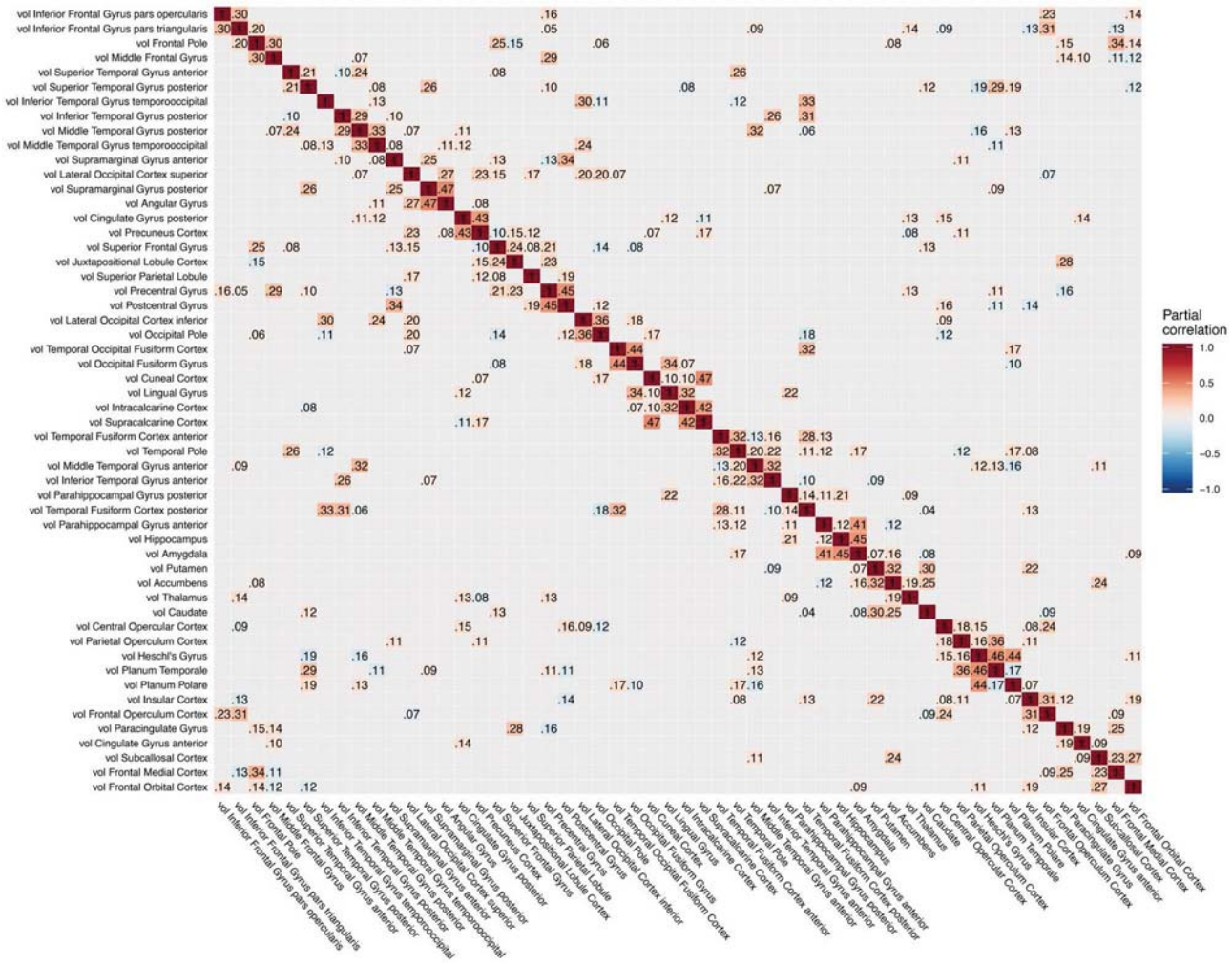


FIGURE 5 | Partial correlation matrix for gray matter volume in the left hemisphere estimated for the combined data of EMCI, LMCI, and AD patients. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; vol, gray matter volume.

density) surviving the posterior probability threshold of $P > 0.5$ (see **Figure S1** showing the probability of links). For comparison, the Pearson correlation matrix is given in **Figure 2** (left). We obtained approximately 6,000 significant Pearson correlations ($P < 0.05$, Bonferroni corrected), corresponding to a network density of 46% of the total number of possible edges.

For intra-modal associations, i.e., within the same imaging modality, brain regions directly adjacent to each other formed smaller clusters of high partial correlation around the main diagonal (**Figures 3**–**5**). When considering inter-modal associations, i.e., between different imaging modalities, we obtained a consistent pattern of significant positive intra-regional conditional dependency for the pairs amyloid-β deposition and metabolism with a mean partial correlation of $\rho = 0.21$ for 43 significant associations. These are visible as the higher intensities in the diagonal of **Figure S2**. Between amyloid-β and gray matter volume as well as between metabolism and gray matter volume, only few significant intra-regional associations were found (**Figures S3, S4**).

3.2. Group Comparison of the Graph Structures

When estimating separate models for each diagnostic group based on the multimodal data, graph structures derived from Pearson and partial correlation matrices (**Figures 6**–**8**) both differed in their density, leading to significant alterations of the clustering coefficient, characteristic path length, and small-world coefficient (**Figure 9** and **Figure S9**). Detailed graph statistics stratified by individual regions and diagnostic groups are provided in **Figures S5**–**S7**.

We observed a biphasic trajectory of the graph measures. This means that the clustering coefficient and small world coefficient initially increases when comparing early MCI and CN participants (**Figure 9**). When Alzheimer's disease progresses, i.e., in the late MCI and dementia groups, both measures decrease again, with late MCI being approximately on the same level as CN participants (**Figure 9**). The characteristic path length showed a similar pattern across groups, but with inverted directionality. All blocks showed significant differences in mean between groups,

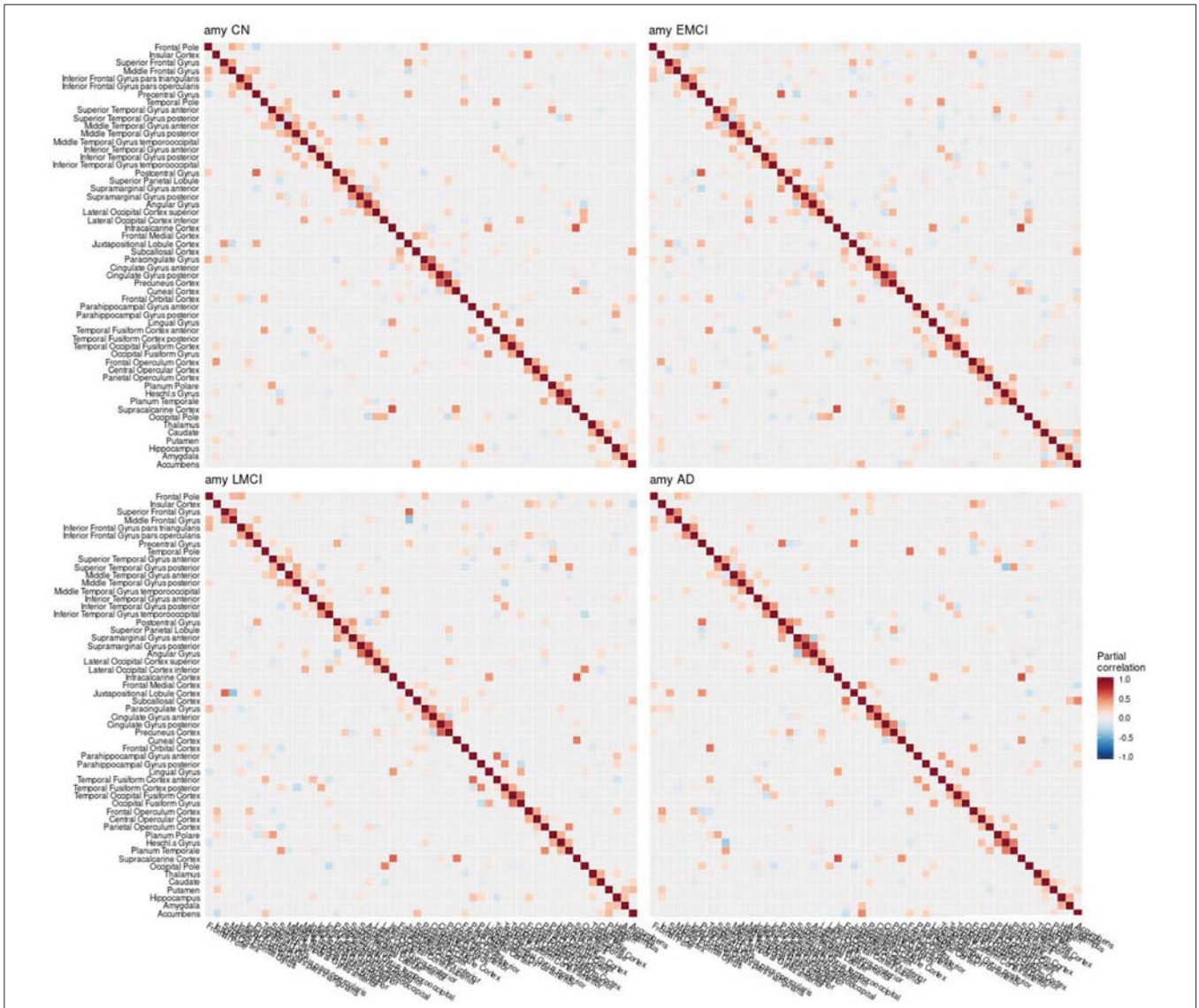


FIGURE 6 | Partial correlation matrix for amyloid-β in the left hemisphere by group. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; amy, amyloid-β.

one-way analysis of variance (ANOVA), $df = 215$, $F \geq 4$, $p < 0.01$, $\eta^2 \geq 0.055$. Detailed results are provided in **Table S2**. *P*-values for Tukey's honest significant difference tests are provided in **Table 2** and **Table S1**. Graph statistics obtained from the right hemisphere data (**Figure S8**) were largely consistent with strongest agreement for the characteristic path length metric.

4. DISCUSSION

4.1. Conditional Dependency Between Brain Regions

The GGMs estimated the strongest conditional dependencies mainly *within* imaging modalities. We expected adjacent brain regions to form clusters with high inter-cluster similarity for amyloid-β deposition (**Figure 3**), as it is known to have low variability in spatial distribution and, therefore, is often used as a dichotomic variable after applying a certain threshold to the global amyloid tracer uptake (Chételat et al., 2013; Landau et al., 2013; Grothe et al., 2017) or as four-stage variable derived from a linear spreading pattern (Grothe et al., 2017; Sakr et al., 2019). We also found such clustering patterns for metabolism (**Figure 4**) and gray matter volume (**Figure 5**), matching previous studies on metabolism and gray matter covariance networks based on Pearson correlation (Yao et al., 2010; Carbonell et al., 2016; Pereira et al., 2016) or principal component analysis (Di and Biswal, 2012; Spetsieris et al., 2015; Savio et al., 2017). Clusters of high covariance have been found in the lateral and medial parietal lobe, lateral frontal lobe, and lateral and medial temporal lobe,

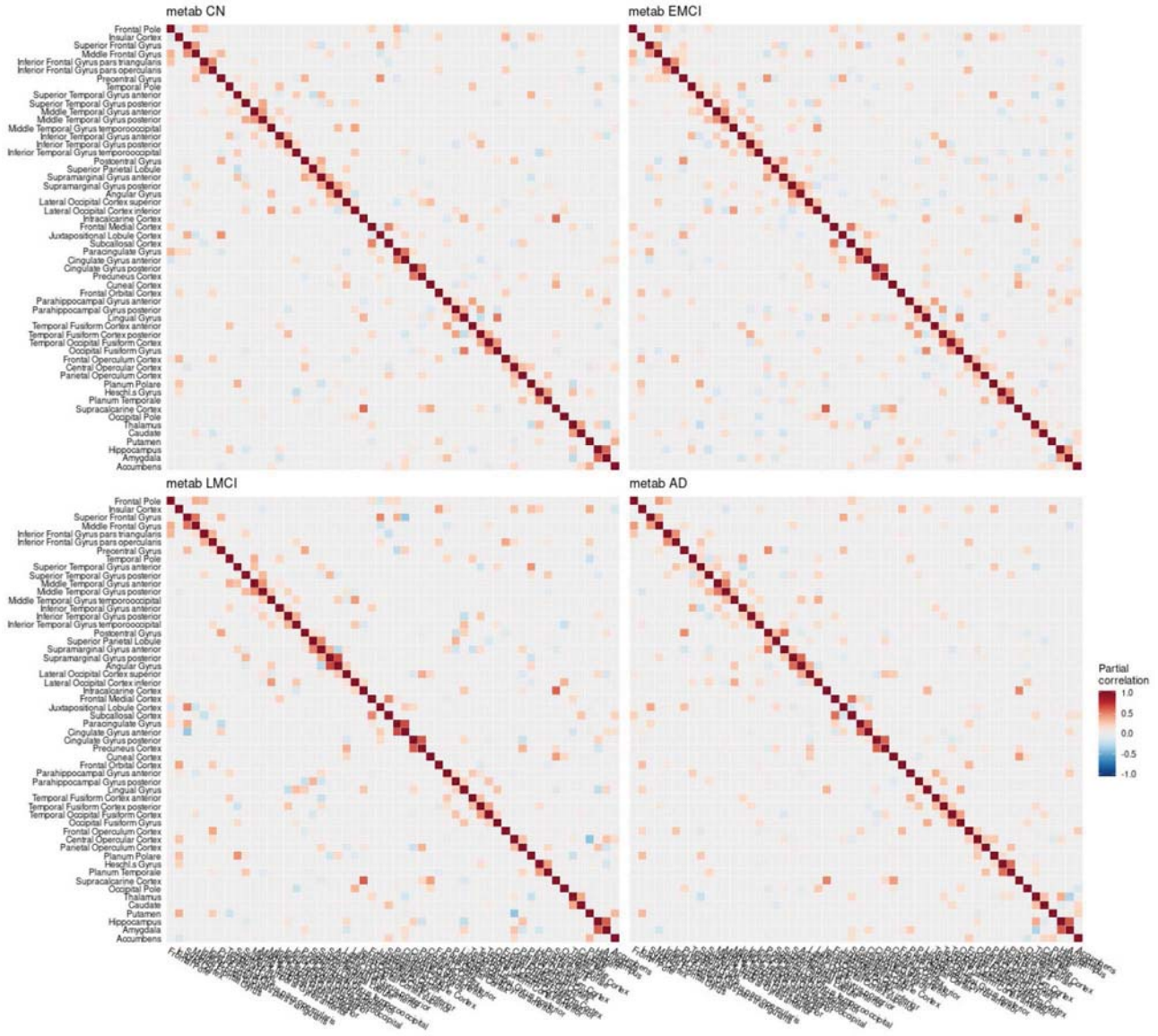


FIGURE 7 | Partial correlation matrix for glucose metabolism in the left hemisphere by group. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; metab, glucose metabolism.

and had been associated with simultaneous growth during brain development, functional co-activation, and axonal connectivity in the literature (Gong et al., 2012; Alexander-Bloch et al., 2013).

Our analyses yielded only few and relatively weak associations *between* different modalities (**Figures S2–S4**), except for the direct intra-regional dependency between amyloid-β and metabolism as well as between amyloid and gray matter volume (diagonal of **Figures S2, S4**), which matched our previous analysis with six selected regions of interest (Dyrba et al., 2017). The positive dependency between amyloid-β and metabolism was strongest in the early MCI group and matches previous results for partial correlation obtained from linear regression models (Altmann et al., 2015). This previous study reported a markedly reduced number and strength of negative associations between regional amyloid-β and metabolism when correcting for global amyloid load. They concluded that the negative association between amyloid deposition and metabolism is more related to the global amyloid level than to the distinct regional level. The pattern of intra-regional dependency between amyloid-β and metabolism as well as between amyloid-β and gray matter volume was strongest in the early MCI group, which could refer to the early phase of the disease and, therefore, a high variation in regional amyloid-β deposition and a strong contribution of the amyloid level on both metabolism and volume (Drzezga et al., 2011; Carbonell et al., 2016). Notably, conditional dependencies between metabolism and volume were obtained only for few

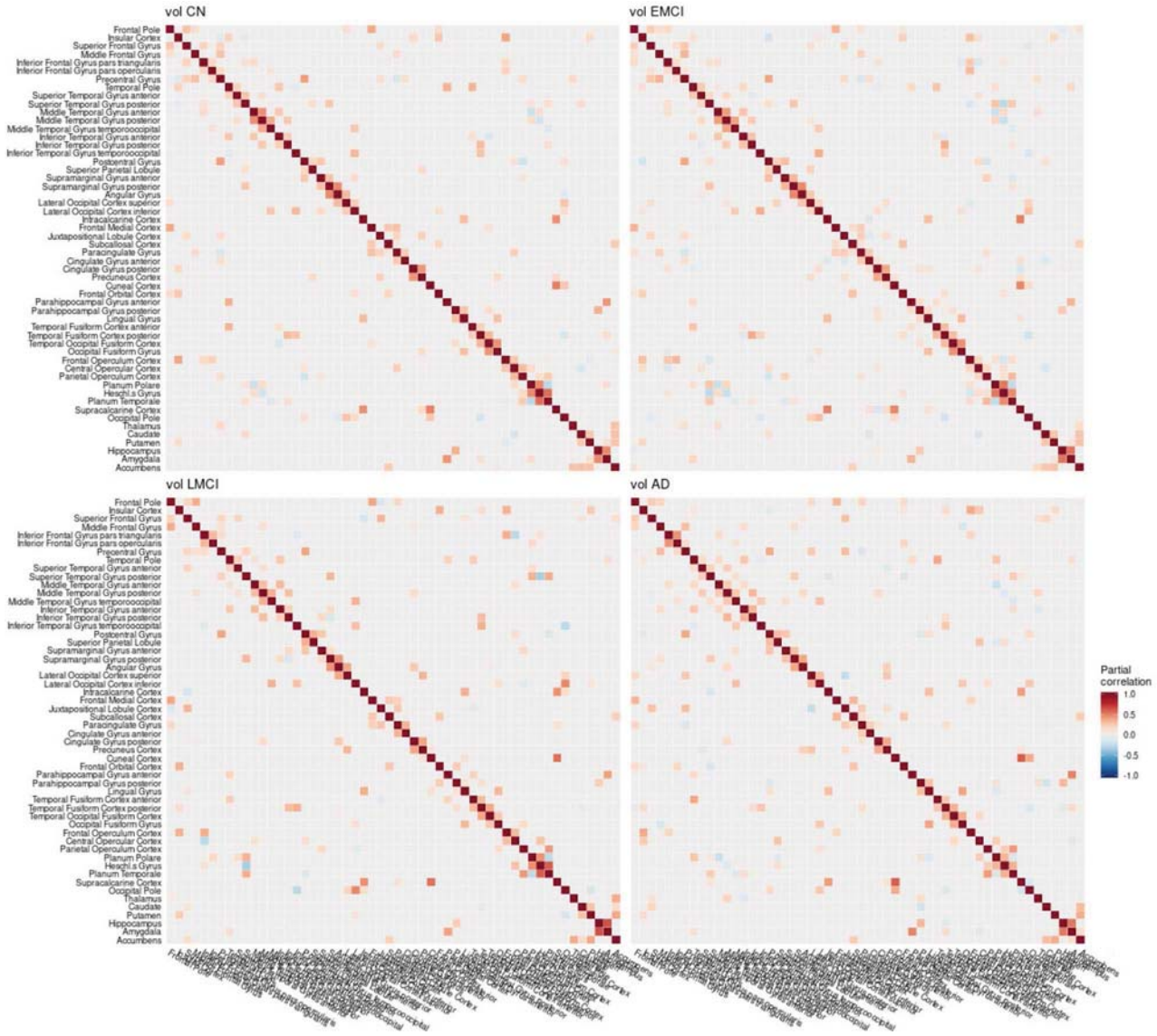


FIGURE 8 | Partial correlation matrix for gray matter volume in the left hemisphere by group. Averaged over 10 repetitions. Associations of lowest magnitude were not present in all iterations. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; vol, gray matter volume.

regions including hippocampus and putamen, but not for other expected regions such as posterior cingulate cortex (Teipel and Grothe, 2016) (**Figure S3**).

4.2. Alterations of Graph Measures

Various studies reported a network disruption of AD in comparison to cognitively healthy controls for gray matter volume (He et al., 2008; Yao et al., 2010; Li et al., 2012; Tijms et al., 2013; John et al., 2017) and glucose metabolism (Morbelli et al., 2012; Titov et al., 2017), and intermediate levels for volume in MCI (Yao et al., 2010; Pereira et al., 2016); which we could replicate in our sample (**Figure S9**). However, it has to be noted that for Pearson correlation matrices usually high thresholds are applied to obtain sparser graphs. Chung et al. (2016) and Voevodskaya et al. (2017) reported a high influence of the selected graph density threshold on the graph measures, leading to divergent increases and decreases of the global clustering coefficient metric. To circumvent such problems, we used weighted versions of the graph measures (Rubinov and Sporns, 2010) and proposed GGMs to obtain sparse conditional dependency matrices. Our results suggest that graph statistics for regional dependency networks follow a biphasic trajectory in the course of AD, a pattern that was recently also reported for cortical thinning and mean diffusivity (Montal et al., 2018) and resting-state fMRI connectivity (Schultz et al., 2017).

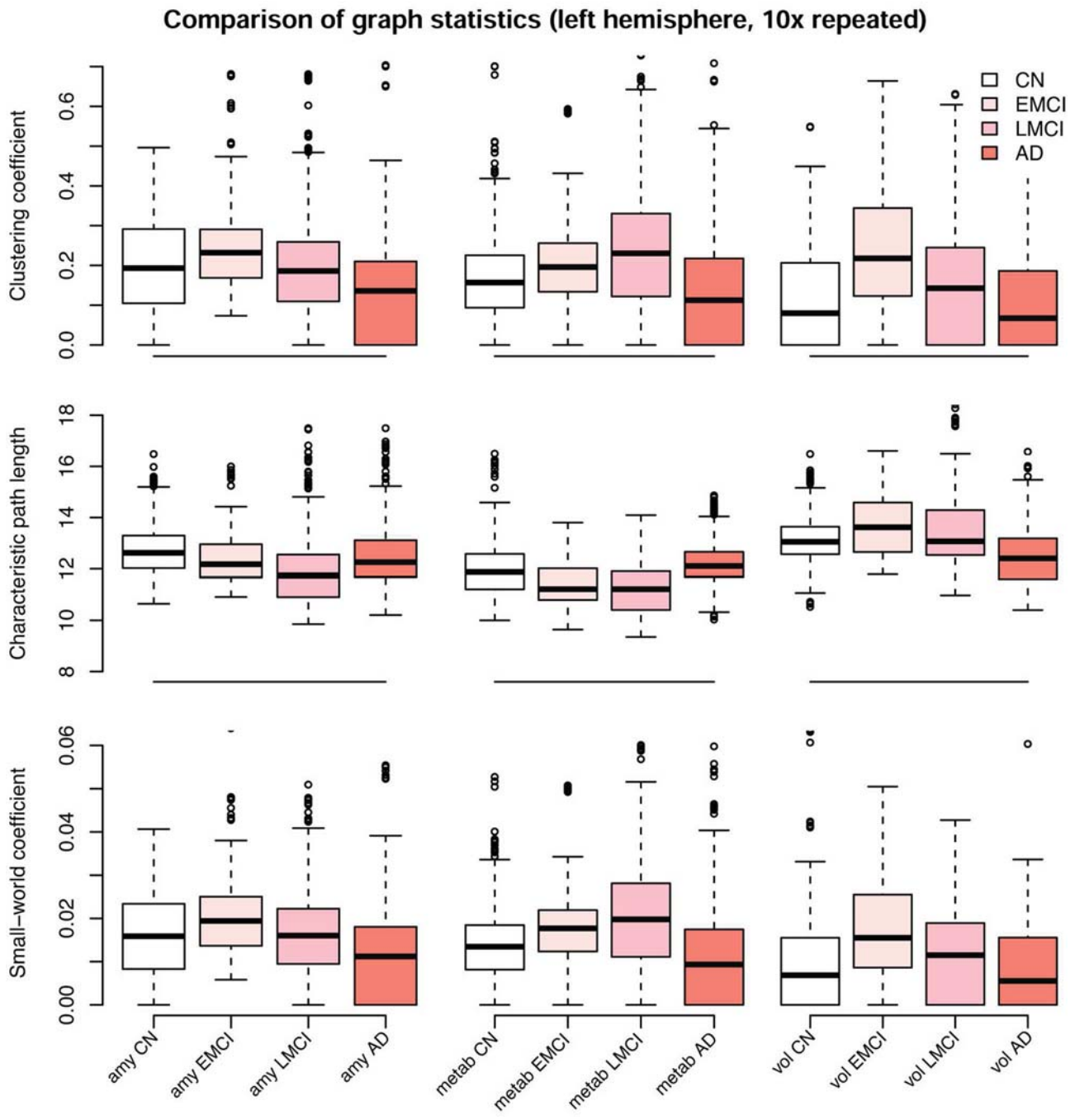


FIGURE 9 | Comparison of graph statistics for the partial correlation matrices of the left hemisphere stratified by diagnostic group and image modality. Estimates based on Gaussian graphical models using multimodal neuroimaging data. The distribution of the weighted clustering coefficient, characteristic weighted path length, and small-world coefficient for individual brain regions is shown. Boxes display median, first and third quartile of the distributions, and whiskers indicate ±1.5×interquartile range. All blocks showed significant differences in mean between groups, one-way analysis of variance (ANOVA), $df = 215$, $F > 4$, $p < 0.01$. *P*-values for Tukey's honest significant difference tests are given in **Table 2**. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia; amy, amyloid-β; metab, glucose metabolism; vol, gray matter volume.

In the current study, the EMCI group displayed the strongest alterations of network structure with an increase of the clustering coefficient, which may relate to the process of amyloid accumulation taking place in several regions simultaneously in this group increasing the intra-cluster correlation. For amyloid-β and volume, LMCI subjects showed a clustering coefficient

TABLE 2 | *P*-values for the group comparison of partial correlation graph statistics (**Figure 9**).

		Amyloid-β			Metabolism			Volume		
		EMCI	**LMCI**	**AD**	**EMCI**	**LMCI**	**AD**	**EMCI**	**LMCI**	**AD**
Clustering coefficient	CN	0.167	0.999	0.178	0.323	0.021	0.718	0.009	0.977	0.999
	EMCI		0.183	< 0.001		0.630	0.031		0.030	0.012
	LMCI			0.162			< 0.001			0.990
Path length	CN	0.264	< 0.001	0.630	0.015	0.001	0.357	0.106	0.664	0.005
	EMCI		0.189	0.922		0.884	< 0.001		0.667	< 0.001
	LMCI			0.044			< 0.001			< 0.001
Small-world coefficient	CN	0.101	0.940	0.301	0.184	0.002	0.701	0.011	0.967	0.987
	EMCI		0.313	< 0.001		0.411	0.011		0.042	0.029
	LMCI			0.096			< 0.001			0.999

Adjusted P*-values from Tukey's honest significant difference tests, controlling for family-wise error rate within each comparison block. CN, cognitively healthy elderly controls; EMCI/LMCI, early and late amnestic mild cognitive impairment; AD, Alzheimer's dementia.*

and small-world coefficient comparable to controls, in contrast to metabolism, where this group showed strongest deviation from the other groups (**Table 2**). The lowest alterations of graph measures were obtained for the gray matter network.

GGMs were recently applied as clustering algorithm for brain networks in a few other single-modality applications. de Vos et al. (2017) found them useful for increasing group separation between AD and controls compared to classical Pearson correlation networks in resting-state functional connectivity. Titov et al. (2017) compared metabolic networks for the differential diagnosis between AD and frontotemporal lobar degeneration (FTLD). They also proposed an algorithm to estimate if an individual subject shows a more AD or FTLD pattern of regional metabolism. Munilla et al. (2017) systematically evaluated the influence of the number of subjects and the regularization strength on the GGM stability and graph structure. They found that the estimated GGM graph structure and small-world coefficient converged to a stable level when including 40 or more subjects in their study sample. For regularization-based approximation of GGMs, they showed that the probability of an edge to exist in the estimated graph structure almost linearly corresponds to the magnitude of their partial correlation. Thus, this finding confirms our initial decision, that sampling-based Bayesian estimation of the graph structure might be more useful for detecting even low associations.

4.3. Limitations

It has to be noted that our methodological framework can currently only be applied as a group statistic but not for individual subjects. Therefore, GGMs can be used for exploratory analyses as alternative to Pearson correlation networks, and may aid generating new hypotheses about the interrelation of clinical variables or feature selection. Then, derived hypotheses can be validated using classical statistical methods such as regression or mediation analysis.

Another limitation is the high uncertainty in the statistical model to estimate the partial correlations. This is due to the theoretically hard problem of matrix inversion on the one hand, and due to the high number of possible graph edges in comparison to the sample size on the other hand. Thus, the model might be fragile with respect to the obtained values and requires large training samples to get stable results. Here, we repeated the model estimation on the whole data for ten times to observe the effect on model stability, which was yielding largely consistent results for strong links with high partial correlation, but getting more variable for weaker links with low partial correlation. Replicating the results using the right hemisphere data also yielded largely consistent results with highest agreement for the characteristic path length metric. Apparent deviation in clustering coefficient and consequently in small-world coefficient (= ratio of both) might be explained by the asymmetry of the brain and the lateralization reported for Alzheimer's disease in the literature (e.g., stronger left hippocampus atrophy in ADNI) (Grothe and Teipel, 2016; Wei, 2018). However, our findings still need to be replicated in independent cohorts.

We observed a saturation of the conditional dependency network when adding many variables. This means, the model parameters might strongly change when having only few variables in the model and adding another variable; in contrast to very stable estimates of larger models with dozens of variables, which are hardly altered when adding another variable. Actually, this problem is well-known for linear regression models and related to multicollinearity in the data (O'brien, 2007; Dormann et al., 2013; Teipel S. J. et al., 2015). Recent developments in stochastic block models may help to overcome these limitations, as they try to infer the underlying clustering block structure and separately estimate statistical associations within and between clusters (Sun et al., 2014; Hosseini and Lee, 2016).

4.4. Conclusion

We applied GGMs to assess inter-modal and inter-regional dependencies of high-dimensional multimodal neuroimaging data of AD-related brain alterations. Our results showed that conditional dependency networks estimated by GGMs provide useful information within imaging modalities and could be used as alternative to Pearson-correlation networks. Nonetheless, GGMs did not detect some expected associations between

modalities and, therefore, may have limited applicability for large-scale data with dozens of variables.

DATA AVAILABILITY STATEMENT

MRI and PET data being used in this study can be retrieved from ADNI (http://adni.loni.usc.edu/data-samples/access-data/). Processed imaging data and extracted regional mean values are available from the corresponding authors upon request. The R package BDgraph can be downloaded from CRAN (https://cran.r-project.org/web/packages/BDgraph) or GitHub (https://github.com/cran/BDgraph).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ADNI internal review board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MD, RM, TK, and ST designed the study. MG and MD preprocessed the imaging data. MD and RM conducted the statistical analyses. MG, TK, and ST aided in interpreting the results. MD drafted the first version of the manuscript. All authors revised the manuscript and contributed to the final version.

FUNDING

This project was supported by the Rostock Massive Data Research Facility (RMDRF) funded by the German Research Foundation (DFG), grant number FKZ INST 264/128-1 FUGG.

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institute of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and provided data but did not participate in analysis or in the writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2020.00099/full#supplementary-material

REFERENCES

(2018). Left lateralized cerebral glucose metabolism declines in amyloid-beta positive persons with mild cognitive impairment. *NeuroImage Clin.* 20, 286–296. doi: 10.1016/j.nicl.2018.07.016

Alexander-Bloch, A., Raznahan, A., Bullmore, E., and Giedd, J. (2013). The convergence of maturational change and structural covariance in human cortical networks. *J. Neurosci.* 33, 2889–2899. doi: 10.1523/JNEUROSCI.3554-12.2013

Altmann, A., Ng, B., Landau, S. M., Jagust, W. J., and Greicius, M. D. (2015). Regional brain hypometabolism is unrelated to regional amyloid plaque burden. *Brain* 138(Pt 12), 3734–3746. doi: 10.1093/brain/awv278

Bontempi, G., and Flauder, M. (2015). From dependency to causality: a machine learning approach. *J. Mach. Learn. Res.* 16, 2437–2457. doi: 10.5555/2789272.2912076

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., et al. (2009). Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci.* 29, 1860–1873. doi: 10.1523/JNEUROSCI.5062-08.2009

Buckner, R. L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, A. F., et al. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J. Neurosci.* 25, 7709–7717. doi: 10.1523/JNEUROSCI.2177-05.2005

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* 100, 139–156. doi: 10.1093/biomet/ass058

Carbonell, F., Zijdenbos, A. P., McLaren, D. G., Iturria-Medina, Y., and Bedell, B. J. (2016). Modulation of glucose metabolism and metabolic connectivity by beta-amyloid. *J. Cereb. Blood Flow Metab.* 36, 2058–2071. doi: 10.1177/0271678X16654492

Chang, Y.-T., Huang, C.-W., Chang, Y.-H., Chen, N.-C., Lin, K.-J., Yan, T.-C., et al. (2015). Amyloid burden in the hippocampus and default mode network: relationships with gray matter volume and cognitive performance in mild stage Alzheimer disease. *Medicine* 94:e763. doi: 10.1097/MD.0000000000000763

Chételat, G., La Joie, R., Villain, N., Perrotin, A., de La Sayette, V., Eustache, F., et al. (2013). Amyloid imaging in cognitively normal individuals, at-risk populations and preclinical Alzheimer's disease. *NeuroImage Clin.* 2, 356–365. doi: 10.1016/j.nicl.2013.02.006

Chung, J., Yoo, K., Kim, E., Na, D. L., and Jeong, Y. (2016). Glucose metabolic brain networks in early-onset vs. late-onset Alzheimer's disease. *Front. Aging Neurosci.* 8:159. doi: 10.3389/fnagi.2016.00159

de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., et al. (2017). A comprehensive analysis of resting state fmri measures to classify individual patients with Alzheimer's disease. *NeuroImage* 167, 62–72. doi: 10.1016/j.neuroimage.2017.11.025

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di, X., and Biswal, B. B. (2012). Metabolic brain covariant networks as revealed by FDG-PET with reference to resting-state fMRI networks. *Brain Connect.* 2, 275–283. doi: 10.1089/brain.2012.0086

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Drzezga, A., Becker, J. A., van Dijk, K. R. A., Sreenivasan, A., Talukdar, T., Sullivan, C., et al. (2011). Neuronal dysfunction and disconnection of cortical hubs in non-demented subjects with elevated amyloid burden. *Brain* 134(Pt 6), 1635–1646. doi: 10.1093/brain/awr066

Dyrba, M., Grothe, M. J., Mohammadi, A., Binder, H., Kirste, T., and Teipel, S. J. (2017). Comparison of different hypotheses regarding the spread of Alzheimer's disease using Markov random fields and multimodal imaging. *J. Alzheimer's Dis.* 65, 731–746. doi: 10.3233/JAD-161197

Gong, G., He, Y., Chen, Z. J., and Evans, A. C. (2012). Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *NeuroImage* 59, 1239–1248. doi: 10.1016/j.neuroimage.2011.08.017

Gonzalez-Escamilla, G., Lange, C., Teipel, S., Buchert, R., and Grothe, M. J. (2017). PETPVE12: an SPM toolbox for partial volume effects correction in brain PET – application to amyloid imaging with AV45-PET. *NeuroImage* 147, 669–677. doi: 10.1016/j.neuroimage.2016.12.077

Grothe, M., Heinsen, H., and Teipel, S. (2013). Longitudinal measures of cholinergic forebrain atrophy in the transition from healthy aging to alzheimer's disease. *Neurobiol. Aging* 34, 1210–1220. doi: 10.1016/j.neurobiolaging.2012.10.018

Grothe, M. J., Barthel, H., Sepulcre, J., Dyrba, M., Sabri, O., and Teipel, S. J. (2017). *In vivo* staging of regional amyloid deposition. *Neurology* 89, 2031–2038. doi: 10.1212/WNL.0000000000004643

Grothe, M. J., Heinsen, H., Amaro, E., Grinberg, L. T., and Teipel, S. J. (2016). Cognitive correlates of basal forebrain atrophy and associated cortical hypometabolism in mild cognitive impairment. *Cereb. Cortex* 26, 2411–2426. doi: 10.1093/cercor/bhv062

Grothe, M. J., and Teipel, S. J. (2016). Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks. *Hum. Brain Mapp.* 37, 35–53. doi: 10.1002/hbm.23018

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn*. Springer Series in Statistics. New York, NY: Springer.

He, Y., Chen, Z., and Evans, A. (2008). Structural insights into aberrant topological patterns of large-scale cortical networks in alzheimer's disease. *J. Neurosci.* 28, 4756–4766. doi: 10.1523/JNEUROSCI.0141-08.2008

Hlinka, J., Hartman, D., Jajcay, N., Tomeček, D., Tintěra, J., and Paluš, M. (2017). Small-world bias of correlation networks: from brain to climate. *Chaos* 27:035812. doi: 10.1063/1.4977951

Hosseini, M. J., and Lee, S.-I. (2016). "Learning sparse gaussian graphical models with overlapping blocks," in *Advances in Neural Information Processing Systems 29*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 3808–3816.

Iturria-Medina, Y., Carbonell, F. M., Sotero, R. C., Chouinard-Decorte, F., and Evans, A. C. (2017). Multifactorial causal model of brain (dis)organization and therapeutic intervention: application to Alzheimer's disease. *NeuroImage* 152, 60–77. doi: 10.1016/j.neuroimage.2017.02.058

John, M., Ikuta, T., and Ferbinteanu, J. (2017). Graph analysis of structural brain networks in Alzheimer's disease: beyond small world properties. *Brain Struct. Funct.* 222, 923–942. doi: 10.1007/s00429-016-1255-4

Kljajevic, V., Grothe, M. J., Ewers, M., and Teipel, S. (2014). Distinct pattern of hypometabolism and atrophy in preclinical and predementia Alzheimer's disease. *Neurobiol. Aging* 35, 1973–1981. doi: 10.1016/j.neurobiolaging.2014.04.006

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

La Joie, R., Perrotin, A., Barré, L., Hommet, C., Mézenge, F., Ibazizene, M., et al. (2012). Region-specific hierarchy between atrophy, hypometabolism, and β-amyloid (aβ) load in Alzheimer's disease dementia. *J. Neurosci.* 32, 16265–16273. doi: 10.1523/JNEUROSCI.2170-12.2012

Landau, S. M., Breault, C., Joshi, A. D., Pontecorvo, M., Mathis, C. A., Jagust, W. J., et al. (2013). Amyloid-beta imaging with pittsburgh compound b and florbetapir: comparing radiotracers and quantification methods. *J. Nucl. Med.* 54, 70–77. doi: 10.2967/jnumed.112.109009

Lauritzen, S. L. (1996). *Graphical Models, Vol. 17*. Oxford Statistical Science Series. Oxford: Clarendon Press.

Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., and Shen, D. (2012). Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiol. Aging* 33, 427.e15–30. doi: 10.1016/j.neurobiolaging.2010.11.008

Mårtensson, G., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., et al. (2018). Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. *Sci. Rep.* 8:11592. doi: 10.1038/s41598-018-29927-0

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. (1996). "Bayesian model averaging," in *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models* (Portland, OR), 77–83.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* 34, 1436–1462. doi: 10.1214/009053606000000281

Mohammadi, A., and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* 10, 109–138. doi: 10.1214/14-BA889

Mohammadi, R., and Wit, E. C. (2019). BDgraph: an R package for Bayesian structure learning in graphical models. *J. Stat. Soft.* 89, 1–30. doi: 10.18637/jss.v089.i03

Montal, V., Vilaplana, E., Alcolea, D., Pegueroles, J., Pasternak, O., Gonz?lez-Ortiz, S., et al. (2018). Cortical microstructural changes along the Alzheimer's disease continuum. *Alzheimer's Dement.* 14, 340–351. doi: 10.1016/j.jalz.2017.09.013

Morbelli, S., Drzezga, A., Perneczky, R., Frisoni, G. B., Caroli, A., van Berckel, B. N. M., et al. (2012). Resting metabolic connectivity in prodromal Alzheimer's disease. A European Alzheimer disease consortium (EADC) project. *Neurobiol. Aging* 33, 2533–2550. doi: 10.1016/j.neurobiolaging.2012.01.005

Müller-Gärtner, H. W., Links, J. M., Prince, J. L., Bryan, R. N., McVeigh, E., Leal, J. P., et al. (1992). Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects. *J. Cereb. Blood Flow Metab.* 12, 571–583. doi: 10.1038/jcbfm.1992.81

Munilla, J., Ortiz, A., Górriz, J. M., and Ramírez, J. (2017). Construction and analysis of weighted brain networks from sice for the study of Alzheimer's disease. *Front. Neuroinform.* 11:19. doi: 10.3389/fninf.2017.00019

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* 41, 673–690. doi: 10.1007/s11135-006-9018-6

Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* 71:065103. doi: 10.1103/PhysRevE.71.065103

Pereira, J. B., Mijalkov, M., Kakaei, E., Mecocci, P., Vellas, B., Tsolaki, M., et al. (2016). Disrupted network topology in patients with stable and progressive mild cognitive impairment and Alzheimer's disease. *Cereb. Cortex* 26, 3476–3493. doi: 10.1093/cercor/bhw128

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electron. J. Stat.* 5, 935–980. doi: 10.1214/11-EJS631

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage* 59, 3852–3861. doi: 10.1016/j.neuroimage.2011.11.054

Sakr, F. A., Grothe, M. J., Cavedo, E., Jelistratova, I., Habert, M.-O., Dyrba, M., et al. (2019). Applicability of *in vivo* staging of regional amyloid burden in a cognitively normal cohort with subjective memory complaints: the INSIGHT-preAD study. *Alzheimer's Res. Ther.* 11:15. doi: 10.1186/s13195-019-0466-3

Savio, A., Fünger, S., Tahmasian, M., Rachakonda, S., Manoliu, A., Sorg, C., et al. (2017). Resting-state networks as simultaneously measured with functional MRI and PET. *J. Nucl. Med.* 58, 1314–1317. doi: 10.2967/jnumed.116.185835

Schultz, A. P., Chhatwal, J. P., Hedden, T., Mormino, E. C., Hanseeuw, B. J., Sepulcre, J., et al. (2017). Phases of hyperconnectivity and hypoconnectivity in the default mode and salience networks track with amyloid and tau in clinically normal individuals. *J. Neurosci.* 37, 4323–4331. doi: 10.1523/JNEUROSCI.3263-16.2017

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., and Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron* 62, 42–52. doi: 10.1016/j.neuron.2009.03.024

Sepulcre, J., Sabuncu, M. R., Becker, A., Sperling, R., and Johnson, K. A. (2013). *In vivo* characterization of the early states of the amyloid-beta network. *Brain* 136(Pt 7), 2239–2252. doi: 10.1093/brain/awt146

Sepulcre, J., Sabuncu, M. R., Li, Q., El Fakhri, G., Sperling, R., and Johnson, K. A. (2017). Tau and amyloid β proteins distinctively associate to functional network changes in the aging brain. *Alzheimer's Dement.* 13, 1261–1269. doi: 10.1016/j.jalz.2017.02.011

Spetsieris, P. G., Ko, J. H., Tang, C. C., Nazem, A., Sako, W., Peng, S., et al. (2015). Metabolic resting-state brain networks in health and disease. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2563–2568. doi: 10.1073/pnas.1411011112

Stam, C., Jones, B., Nolte, G., Breakspear, M., and Scheltens, P. (2006). Small-world networks and functional connectivity in Alzheimer's disease. *Cereb. Cortex* 17, 92–99. doi: 10.1093/cercor/bhj127

Sun, S., Zhu, Y., and Xu, J. (2014). " Adaptive variable clustering in Gaussian graphical models," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Vol. 33 of* Proceedings of Machine Learning Research, eds S. Kaski and J. Corander (Iceland: Reykjavik), 931–939.

Teipel, S., Drzezga, A., Grothe, M. J., Barthel, H., Chételat, G., Schuff, N., et al. (2015). Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. *Lancet Neurol.* 14, 1037–1053. doi: 10.1016/S1474-4422(15)00093-9

Teipel, S., and Grothe, M. J. (2016). Does posterior cingulate hypometabolism result from disconnection or local pathology across preclinical and clinical stages of Alzheimer's disease? *Eur. J. Nucl. Med. Mol. Imaging* 43, 526–536. doi: 10.1007/s00259-015-3222-3

Teipel, S., Grothe, M. J., Zhou, J., Sepulcre, J., Dyrba, M., Sorg, C., et al. (2016). Measuring cortical connectivity in Alzheimer's disease as a brain neural network pathology: toward clinical applications. *J. Int. Neuropsychol. Soc.* 22, 138–163. doi: 10.1017/S1355617715000995

Teipel, S. J., Kurth, J., Krause, B., and Grothe, M. J. (2015). The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - beyond classical regression. *NeuroImage Clin.* 8, 583–593. doi: 10.1016/j.nicl.2015.05.006

Tijms, B. M., Möller, C., Vrenken, H., Wink, A. M., de Haan, W., van der Flier, W. M., et al. (2013). Single-subject grey matter graphs in Alzheimer's disease. *PLoS ONE* 8:e58921. doi: 10.1371/journal.pone.0058921

Titov, D., Diehl-Schmid, J., Shi, K., Perneczky, R., Zou, N., Grimmer, T., et al. (2017). Metabolic connectivity for differential diagnosis of dementing disorders. *J. Cereb. Blood Flow Metab.* 37, 252–262. doi: 10.1177/0271678X15622465

Torok, J., Maia, P. D., Powell, F., Pandya, S., and Raj, A. (2018). A method for inferring regional origins of neurodegeneration. *Brain* 141, 863–876. doi: 10.1093/brain/awx371

Villain, N., Fouquet, M., Baron, J.-C., Mézenge, F., Landeau, B., de La Sayette, V., et al. (2010). Sequential relationships between grey matter and white matter atrophy and brain metabolic abnormalities in early Alzheimer's disease. *Brain* 133, 3301–3314. doi: 10.1093/brain/awq203

Voevodskaya, O., Pereira, J. B., Volpe, G., Lindberg, O., Stomrud, E., van Westen, D., et al. (2017). Altered structural network organization in cognitively normal individuals with amyloid pathology. *Neurobiol. Aging* 64, 15–24. doi: 10.1016/j.neurobiolaging.2017.11.014

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Front. Neurosci.* 10:123. doi: 10.3389/fnins.2016.00123

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Yao, Z., Zhang, Y., Lin, L., Zhou, Y., Xu, C., and Jiang, T. (2010). Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease. *PLoS Comput. Biol.* 6:e1001006. doi: 10.1371/journal.pcbi.1001006

Zhou, J., Gennatas, E. D., Kramer, J. H., Miller, B. L., and Seeley, W. W. (2012). Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron* 73, 1216–1227. doi: 10.1016/j.neuron.2012.03.004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material for Dyrba et al. (2020) Gaussian graphical models reveal inter-modal and inter-regional conditional dependencies of brain alterations in Alzheimer's disease.

Frontiers in Aging Neuroscience | doi: 10.3389/fnagi.2020.00099

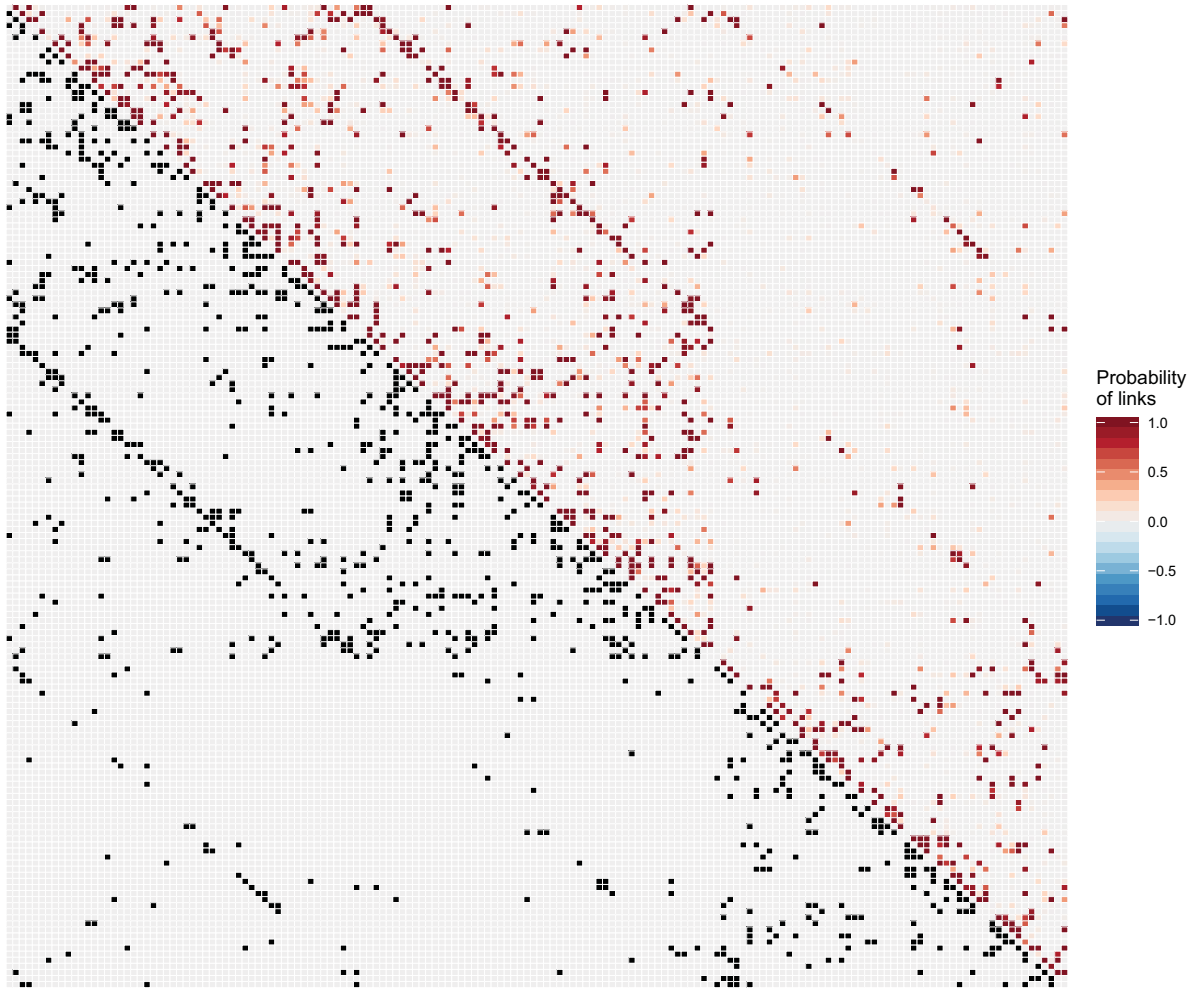


Figure S1. Probability of estimated edges for the left hemisphere. The upper right part provides the raw probability of each edge to exist. The lower left part indicates the selected edges exceeding the threshold of $P_{avg} > 0.5$.

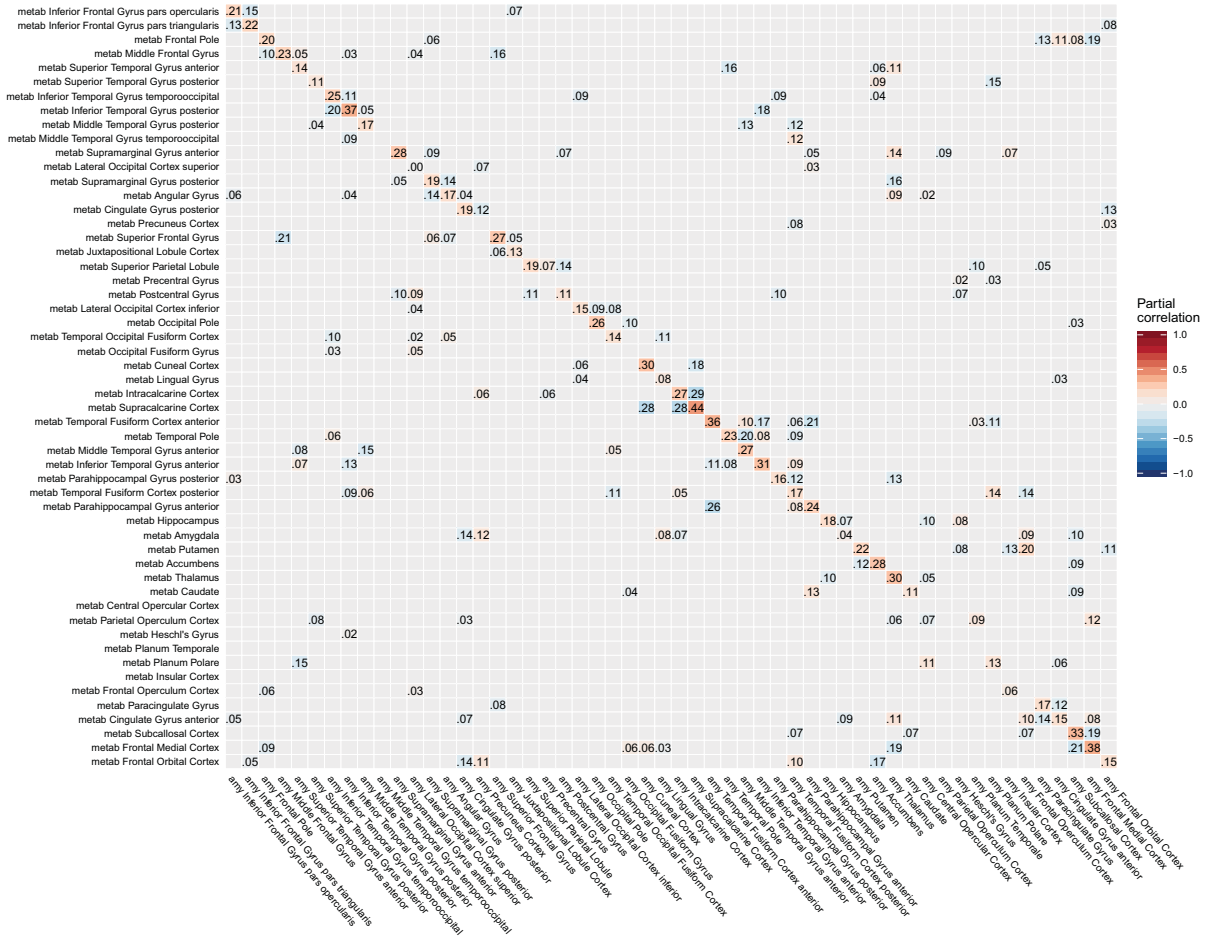


Figure S2. Partial correlation matrix for amyloid-β deposition and glucose metabolism in the left hemisphere estimated for the combined data of EMCI, LMCI and AD patients. Averaged over ten repetitions. Associations of lowest magnitude were not present in all iterations.
EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism.

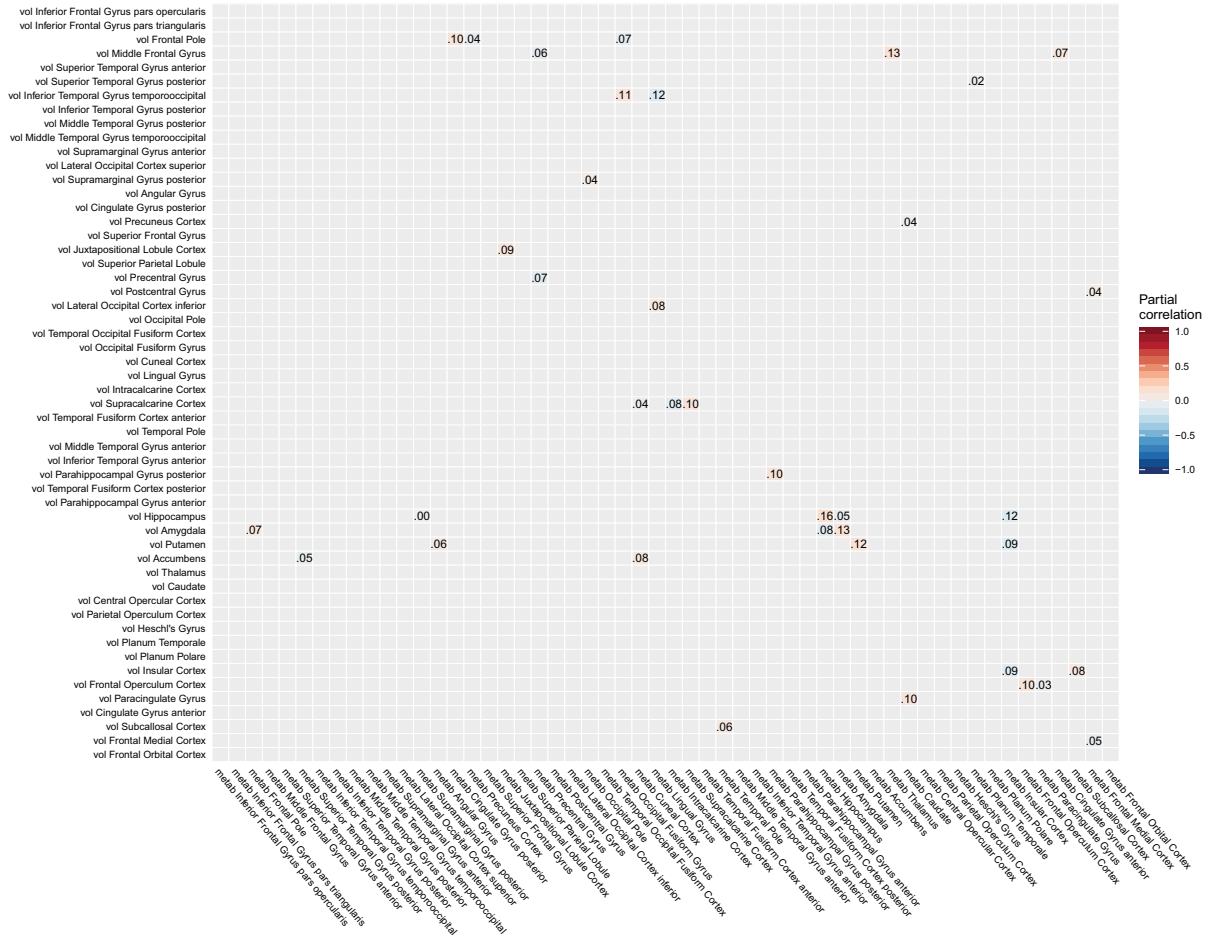


Figure S3. Partial correlation matrix for glucose metabolism and gray matter volume in the left hemisphere estimated for the combined data of EMCI, LMCI and AD patients. Averaged over ten repetitions. Associations of lowest magnitude were not present in all iterations.
EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, metab: glucose metabolism, vol: gray matter volume.

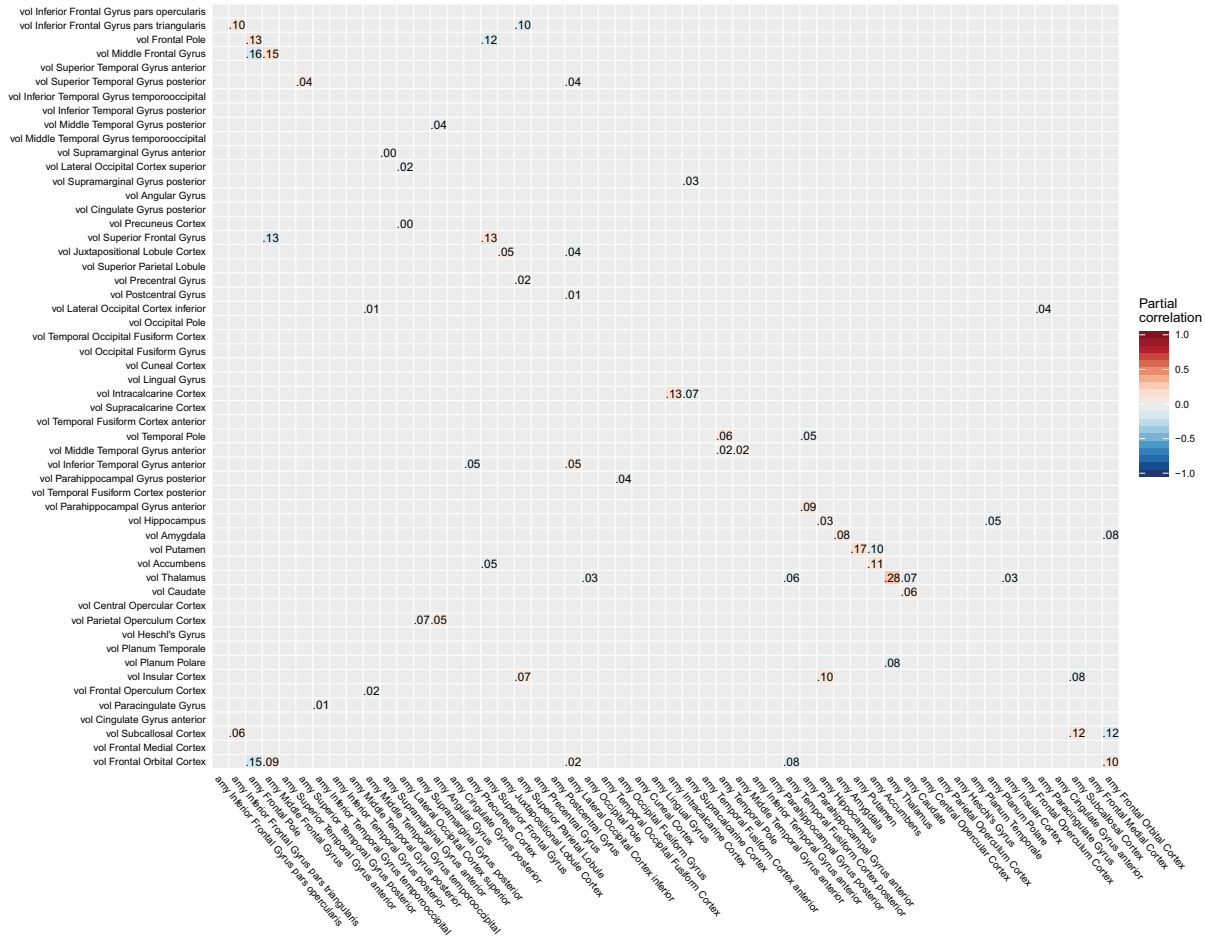


Figure S4. Partial correlation matrix for amyloid-β deposition and gray matter volume in the left hemisphere estimated for the combined data of EMCI, LMCI and AD patients. Averaged over ten repetitions. Associations of lowest magnitude were not present in all iterations.
EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, vol: gray matter volume.

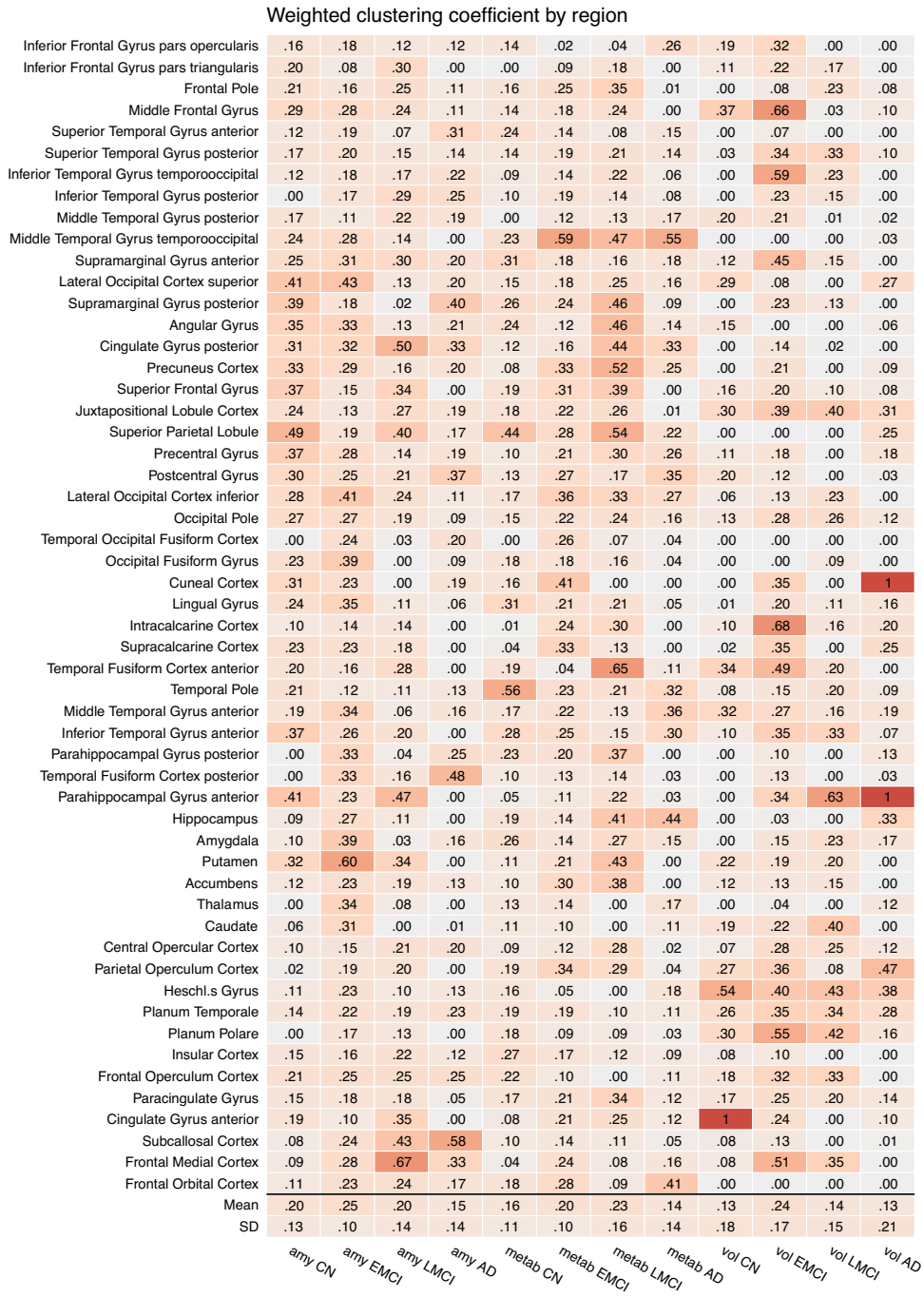


Figure S5. Comparison of weighted clustering coefficient stratified by brain region, diagnostic group and modality for the partial correlation matrices of the left hemisphere. Averaged over ten repetitions.
CN: cognitively healthy elderly controls, EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.

Characteristic path length by region

	amy CN	amy EMCI	amy LMCI	amy AD	metab CN	metab EMCI	metab LMCI	metab AD	vol CN	vol EMCI	vol LMCI	vol AD
Inferior Frontal Gyrus pars opercularis	12.24	12.64	10.92	11.78	13.22	12.84	10.34	11.72	13.36	14.22	14.75	12.87
Inferior Frontal Gyrus pars triangularis	12.33	12.69	10.86	12.16	12.72	12.22	10.97	12.26	13.01	12.53	12.86	13.25
Frontal Pole	11.71	13.01	10.05	12.32	11.15	11.20	11.77	12.22	12.75	13.87	13.19	11.74
Middle Frontal Gyrus	12.54	13.42	11.01	12.83	11.43	10.85	11.69	11.84	14.67	14.90	13.36	11.08
Superior Temporal Gyrus anterior	11.73	12.10	10.88	13.48	10.61	10.84	10.43	10.42	13.26	13.26	11.92	12.82
Superior Temporal Gyrus posterior	10.79	11.28	10.32	11.94	11.68	10.08	10.31	11.10	12.27	12.42	12.10	12.07
Inferior Temporal Gyrus temporooccipital	13.44	11.79	10.97	11.91	10.52	11.01	10.15	12.16	13.40	14.04	12.20	12.38
Inferior Temporal Gyrus posterior	12.78	12.19	10.72	11.66	11.56	11.33	10.18	11.78	12.94	13.44	11.69	12.11
Middle Temporal Gyrus posterior	11.35	11.53	9.96	11.86	12.31	10.99	10.56	11.11	12.74	12.10	11.99	11.44
Middle Temporal Gyrus temporooccipital	12.14	11.62	11.73	11.72	12.18	11.76	11.61	11.84	13.22	13.34	13.10	11.13
Supramarginal Gyrus anterior	12.27	12.00	12.61	12.21	10.71	10.14	9.92	11.87	13.11	13.60	12.92	12.70
Lateral Occipital Cortex superior	12.16	11.36	10.90	11.85	11.34	10.41	11.27	12.30	12.82	12.21	13.32	10.96
Supramarginal Gyrus posterior	11.73	11.65	11.96	12.32	10.97	11.05	10.03	11.71	13.19	13.12	12.31	12.86
Angular Gyrus	12.05	11.68	10.96	12.24	11.93	12.16	10.26	11.88	13.73	13.39	12.93	13.03
Cingulate Gyrus posterior	13.13	11.93	12.34	12.05	12.66	11.33	10.82	12.66	12.99	12.78	14.16	13.32
Precuneus Cortex	12.68	11.17	12.33	11.38	12.07	11.74	10.70	12.36	12.66	13.67	14.86	12.38
Superior Frontal Gyrus	14.09	13.05	12.21	12.57	11.68	11.66	11.46	12.67	12.94	13.56	11.22	12.47
Juxtapositional Lobule Cortex	14.42	12.35	11.33	11.44	11.88	12.01	11.81	12.31	13.35	14.76	13.51	11.65
Superior Parietal Lobule	12.30	12.89	13.41	11.58	12.22	11.79	11.02	13.11	13.56	13.69	15.65	11.87
Precentral Gyrus	13.03	12.17	10.51	10.45	11.75	10.79	11.39	11.65	12.09	12.70	11.86	10.74
Postcentral Gyrus	12.21	12.11	11.05	11.02	11.01	10.03	10.61	12.02	12.01	12.02	12.89	10.75
Lateral Occipital Cortex inferior	11.71	11.30	11.07	11.63	11.17	10.89	11.77	11.71	12.10	12.54	12.23	10.74
Occipital Pole	11.36	11.04	12.74	11.13	15.80	11.55	13.21	11.80	12.38	12.92	11.81	12.10
Temporal Occipital Fusiform Cortex	11.89	12.93	12.34	12.83	12.52	10.68	11.32	11.93	12.42	14.26	12.75	11.45
Occipital Fusiform Gyrus	12.13	13.02	12.22	12.80	12.22	10.14	11.40	12.23	13.12	14.47	13.11	11.02
Cuneal Cortex	13.72	12.17	16.60	13.46	12.94	13.28	14.00	14.17	14.42	15.10	15.74	13.63
Lingual Gyrus	12.72	13.60	12.64	13.79	12.15	9.68	9.51	12.34	14.89	15.47	13.36	11.32
Intracalcarine Cortex	12.75	13.95	14.35	14.57	12.73	11.95	12.19	14.23	15.52	15.59	12.69	13.29
Supracalcarine Cortex	13.72	13.49	15.37	14.25	12.71	12.76	12.75	14.46	14.61	14.74	14.36	13.87
Temporal Fusiform Cortex anterior	13.28	13.37	12.25	13.84	13.61	12.99	11.26	13.23	13.25	13.90	13.00	12.47
Temporal Pole	12.38	12.12	11.90	12.04	13.38	12.04	11.01	12.33	11.83	12.81	12.96	11.81
Middle Temporal Gyrus anterior	11.60	11.89	11.02	12.52	12.52	11.33	10.93	11.69	12.87	12.61	13.29	12.06
Inferior Temporal Gyrus anterior	13.32	13.89	13.32	13.05	12.45	12.08	12.17	12.27	13.02	13.90	13.08	12.08
Parahippocampal Gyrus posterior	14.26	12.56	10.91	13.04	11.78	10.21	11.06	13.63	14.89	15.55	13.77	11.71
Temporal Fusiform Cortex posterior	12.79	12.62	10.53	12.33	10.90	10.33	10.58	11.68	11.72	12.47	11.27	10.57
Parahippocampal Gyrus anterior	14.08	14.29	12.39	16.26	12.43	12.64	11.82	12.91	13.38	15.52	15.42	13.70
Hippocampus	14.20	12.81	11.95	15.23	12.01	10.46	10.19	12.14	14.32	15.28	16.35	13.00
Amygdala	14.49	12.78	14.22	15.28	11.42	10.72	10.13	11.76	13.08	14.89	16.24	12.70
Putamen	15.18	13.87	13.58	13.24	13.67	12.47	10.04	13.54	13.75	14.73	12.67	14.77
Accumbens	13.40	11.19	11.87	12.11	12.63	13.15	11.87	12.41	13.88	14.85	13.21	13.53
Thalamus	15.45	15.64	13.16	14.54	10.60	11.25	12.09	10.91	14.81	14.92	15.70	13.14
Caudate	14.51	11.98	12.80	13.40	11.92	10.79	13.52	12.04	13.78	14.20	14.90	13.35
Central Opercular Cortex	12.85	11.23	12.13	11.61	11.21	11.23	10.30	11.90	11.99	12.54	12.83	11.50
Parietal Operculum Cortex	12.70	12.00	11.19	12.66	11.57	11.07	10.19	13.61	12.36	13.12	13.66	13.70
Heschl.s Gyrus	12.46	11.89	11.21	12.16	10.66	10.92	11.76	13.61	13.25	12.14	12.61	12.65
Planum Temporale	11.94	11.28	11.01	12.41	11.63	10.96	11.61	13.08	12.71	12.70	12.79	12.92
Planum Polare	13.18	12.46	11.04	13.42	10.58	10.82	10.09	11.36	12.64	12.54	12.81	11.64
Insular Cortex	12.11	11.93	11.89	11.36	11.90	11.12	10.54	11.16	11.00	12.04	12.00	12.39
Frontal Operculum Cortex	12.27	11.32	11.38	10.43	12.96	11.81	12.81	11.36	11.81	12.23	12.73	13.40
Paracingulate Gyrus	11.72	12.26	10.79	11.27	10.48	10.90	12.21	12.05	12.57	14.06	14.96	12.04
Cingulate Gyrus anterior	12.66	12.19	11.55	11.55	10.33	11.14	11.92	11.99	15.65	14.07	17.87	15.94
Subcallosal Cortex	12.84	12.07	13.15	12.74	11.89	13.62	12.53	11.99	13.79	14.46	14.30	13.31
Frontal Medial Cortex	12.71	13.06	14.33	12.73	11.03	12.26	12.80	12.58	12.73	14.97	14.24	13.12
Frontal Orbital Cortex	12.38	11.59	11.04	11.37	13.94	12.15	11.50	12.48	12.56	12.97	13.21	13.05
Mean	12.78	12.38	11.94	12.52	11.95	11.40	11.27	12.25	13.17	13.67	13.42	12.44
SD	1.03	.93	1.35	1.23	1.05	.91	1.02	.89	.98	1.09	1.40	1.10

Figure S6. Comparison of characteristic path length stratified by brain region, diagnostic group and modality for the partial correlation matrices of the left hemisphere. Averaged over ten repetitions. CN: cognitively healthy elderly controls, EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.

Small−world coefficient by region

	amy CN	amy EMCI	amy LMCI	amy AD	metab CN	metab EMCI	metab LMCI	metab AD	vol CN	vol EMCI	vol LMCI	vol AD
Inferior Frontal Gyrus pars opercularis	13.01	13.90	10.92	10.34	10.58	1.85	3.79	22.33	14.32	22.39	.00	.00
Inferior Frontal Gyrus pars triangularis	16.63	6.07	27.30	.00	.00	7.62	16.72	.00	8.32	17.85	13.05	.00
Frontal Pole	17.98	11.95	24.99	8.95	14.55	22.66	29.67	1.18	.00	5.81	17.10	7.03
Middle Frontal Gyrus	23.03	20.91	21.42	8.67	11.90	16.99	20.75	.00	25.55	44.45	2.47	9.38
Superior Temporal Gyrus anterior	10.33	15.45	6.00	23.29	22.77	13.30	7.47	14.05	.00	5.31	.00	.00
Superior Temporal Gyrus posterior	15.87	17.50	14.66	11.30	11.56	18.39	20.65	12.61	2.47	27.21	27.00	8.68
Inferior Temporal Gyrus temporooccipital	8.90	14.97	15.30	18.67	8.38	12.51	21.35	4.87	.00	41.81	18.63	.00
Inferior Temporal Gyrus posterior	.00	13.80	26.82	21.82	8.25	16.66	13.92	7.22	.00	16.92	13.10	.00
Middle Temporal Gyrus posterior	15.40	9.74	21.70	15.71	.00	11.08	12.75	15.17	15.54	17.31	.56	1.67
Middle Temporal Gyrus temporooccipital	20.00	24.41	12.10	.00	19.03	50.11	40.54	46.49	.00	.00	.00	2.94
Supramarginal Gyrus anterior	20.61	26.06	23.93	16.36	29.13	18.10	16.26	15.43	9.43	33.21	11.66	.00
Lateral Occipital Cortex superior	33.82	37.66	11.74	16.51	13.38	17.23	22.36	12.92	22.67	6.83	.00	24.85
Supramarginal Gyrus posterior	33.38	15.30	1.63	32.67	23.89	21.46	46.56	7.77	.00	17.62	10.55	.00
Angular Gyrus	29.24	28.46	12.06	17.13	20.12	9.70	44.97	11.37	11.19	.00	.00	4.48
Cingulate Gyrus posterior	23.75	27.22	40.43	27.16	10.01	13.85	40.71	26.04	.00	10.93	1.64	.00
Precuneus Cortex	26.21	25.59	13.12	17.15	6.94	28.27	48.49	19.96	.00	15.26	.00	7.10
Superior Frontal Gyrus	25.96	11.82	27.87	.00	16.10	26.48	33.74	.00	12.07	15.05	8.98	6.06
Juxtapositional Lobule Cortex	16.36	10.46	23.74	16.81	15.38	18.32	21.97	.44	22.61	26.74	29.20	26.70
Superior Parietal Lobule	40.05	15.02	29.51	14.26	35.69	24.09	48.52	16.84	.00	.00	.00	20.81
Precentral Gyrus	28.29	22.98	13.23	18.53	8.81	19.81	26.35	22.44	9.05	13.90	.00	16.56
Postcentral Gyrus	24.82	20.99	19.38	33.24	11.49	26.93	15.86	29.16	16.28	10.29	.00	2.95
Lateral Occipital Cortex inferior	23.85	35.85	21.73	9.32	15.40	32.61	27.65	23.09	5.26	10.48	18.45	.00
Occipital Pole	23.40	24.43	14.90	8.21	9.22	19.41	17.96	13.83	10.74	21.84	22.26	10.16
Temporal Occipital Fusiform Cortex	.00	18.73	2.57	15.56	.00	24.04	6.33	3.12	.00	.00	.00	.00
Occipital Fusiform Gyrus	19.06	30.01	.00	6.84	14.67	17.39	13.72	3.35	.00	.00	7.23	.00
Cuneal Cortex	22.64	19.24	.00	13.90	12.17	31.01	.00	.00	.00	22.88	.00	73.39
Lingual Gyrus	18.72	25.40	8.86	4.62	25.51	21.52	21.73	4.32	.92	12.94	8.61	14.39
Intracalcarine Cortex	7.80	9.94	9.79	.00	.62	20.48	24.95	.00	6.07	43.61	12.27	15.86
Supracalcarine Cortex	17.03	16.83	11.61	.00	3.42	26.12	10.26	.00	1.08	23.42	.00	18.16
Temporal Fusiform Cortex anterior	14.79	12.34	22.90	.00	13.93	2.86	57.93	8.25	25.49	35.01	15.52	.00
Temporal Pole	16.90	9.52	9.32	10.48	41.77	19.17	19.40	25.82	6.86	11.50	15.29	7.52
Middle Temporal Gyrus anterior	16.16	28.74	5.86	12.56	13.38	19.20	11.60	30.47	24.60	21.25	12.26	15.86
Inferior Temporal Gyrus anterior	27.84	19.01	15.32	.00	22.70	20.79	12.10	24.69	7.59	25.52	25.49	6.15
Parahippocampal Gyrus posterior	.00	26.39	3.98	19.61	19.62	19.38	33.20	.00	.00	6.35	.00	11.49
Temporal Fusiform Cortex posterior	.00	25.93	15.49	39.24	8.97	12.62	12.98	2.30	.00	10.06	.00	2.80
Parahippocampal Gyrus anterior	29.46	15.71	38.01	.00	3.87	9.08	18.55	2.15	.00	22.05	40.86	72.98
Hippocampus	6.31	20.78	9.59	.00	15.75	13.83	40.37	36.53	.00	1.72	.00	25.31
Amygdala	7.12	30.87	2.22	10.48	22.83	12.80	26.71	13.16	.00	10.39	14.34	13.07
Putamen	21.14	43.58	24.85	.00	7.65	16.57	42.57	.00	15.71	12.85	15.58	.00
Accumbens	8.72	20.61	15.93	10.87	7.99	22.91	31.95	.00	8.76	8.69	11.07	.00
Thalamus	.00	21.58	6.08	.00	11.89	12.88	.00	15.92	.00	2.78	.00	9.68
Caudate	4.27	26.16	.00	.72	9.45	9.22	.00	9.13	14.04	15.41	26.82	.00
Central Opercular Cortex	7.86	13.55	17.70	17.04	7.75	10.95	26.75	1.86	5.59	22.02	19.19	10.71
Parietal Operculum Cortex	1.91	15.63	17.58	.00	16.11	30.69	28.09	3.25	22.10	27.11	5.94	33.75
Heschl.s Gyrus	8.74	19.49	8.71	10.32	14.78	4.75	.00	13.35	40.66	33.36	34.30	30.27
Planum Temporale	11.86	19.80	17.14	18.22	16.57	17.10	8.76	8.56	20.60	27.57	26.58	21.44
Planum Polare	.00	13.42	11.60	.00	17.46	8.25	8.82	3.03	23.64	44.12	32.25	14.13
Insular Cortex	12.64	13.23	18.69	10.63	22.59	15.44	11.83	8.20	6.98	8.51	.00	.00
Frontal Operculum Cortex	17.39	22.36	22.01	24.18	17.00	8.88	.00	9.35	15.28	26.27	25.79	.00
Paracingulate Gyrus	13.14	14.51	16.22	4.38	16.59	18.98	27.81	9.74	13.30	18.09	13.09	11.22
Cingulate Gyrus anterior	15.14	8.24	29.90	.00	7.55	18.82	21.13	9.74	63.92	17.42	.00	6.04
Subcallosal Cortex	6.27	19.96	32.96	45.66	8.22	9.95	8.85	3.75	5.58	8.66	.00	.48
Frontal Medial Cortex	7.07	21.11	47.05	25.73	3.79	19.58	5.88	13.04	6.35	34.22	24.70	.00
Frontal Orbital Cortex	8.93	19.82	21.47	15.28	13.24	23.41	7.71	33.16	.00	.00	.00	.00
Mean	15.55	19.87	16.63	12.27	13.71	17.89	21.09	11.47	9.64	17.32	10.77	10.45
SD	10.27	8.15	10.83	11.28	8.75	8.28	14.75	11.72	12.46	12.35	11.36	15.71

Figure S7. Comparison of small-world coefficient stratified by brain region, diagnostic group and modality for the partial correlation matrices of the left hemisphere. For better readability, individual values were upscaled by a factor of 1,000. Averaged over ten repetitions.
CN: cognitively healthy elderly controls, EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.

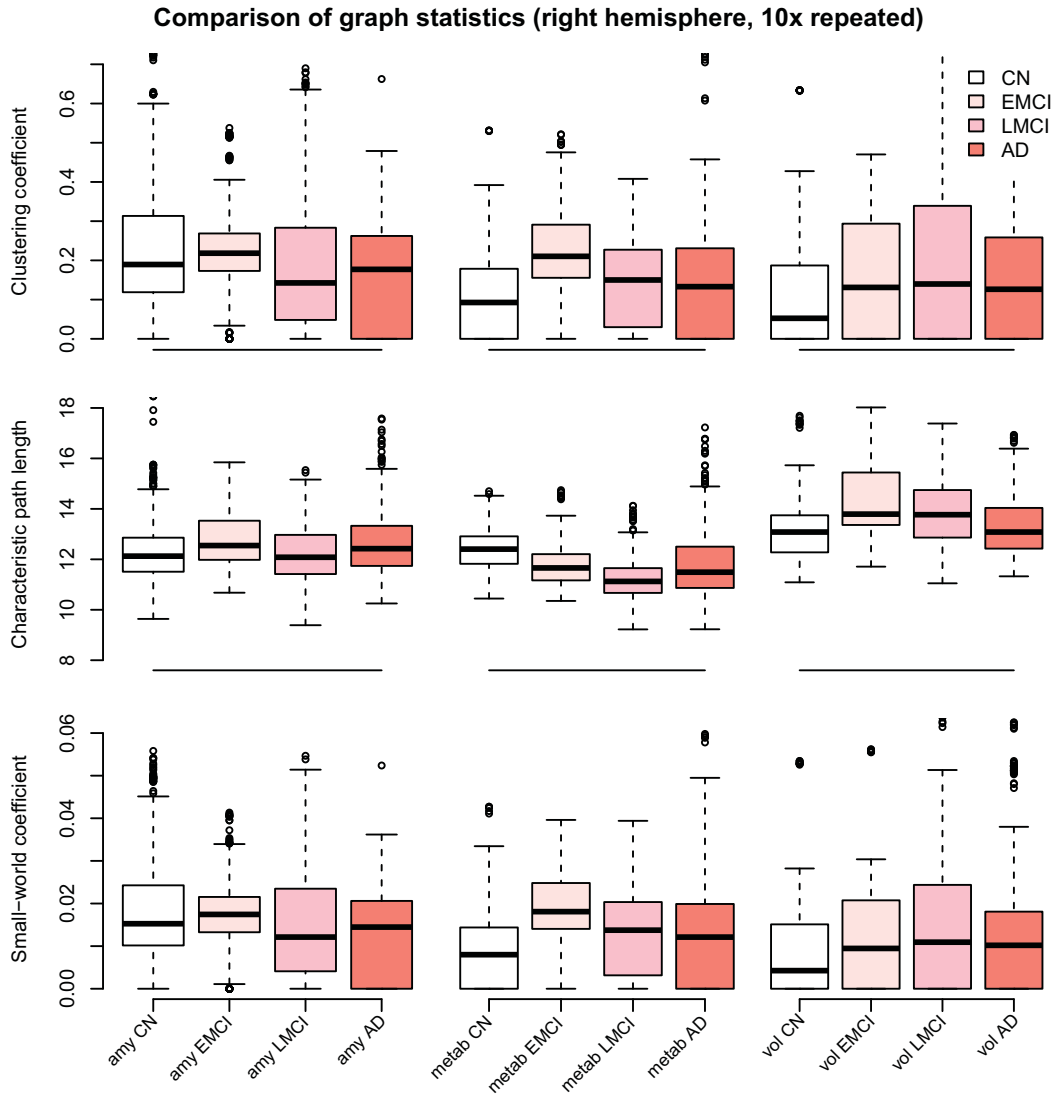


Figure S8. Comparison of graph statistics for the partial correlation matrices of the **right** hemisphere stratified by diagnostic group and image modality. Estimates based on Gaussian graphical models using multimodal neuroimaging data. The distribution of the weighted clustering coefficient, characteristic weighted path length, and small-world coefficient for individual brain regions is shown. Boxes display median, first and third quartile of the distributions, and whiskers indicate ±1.5×interquartile range.
CN: cognitively healthy elderly controls, EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.

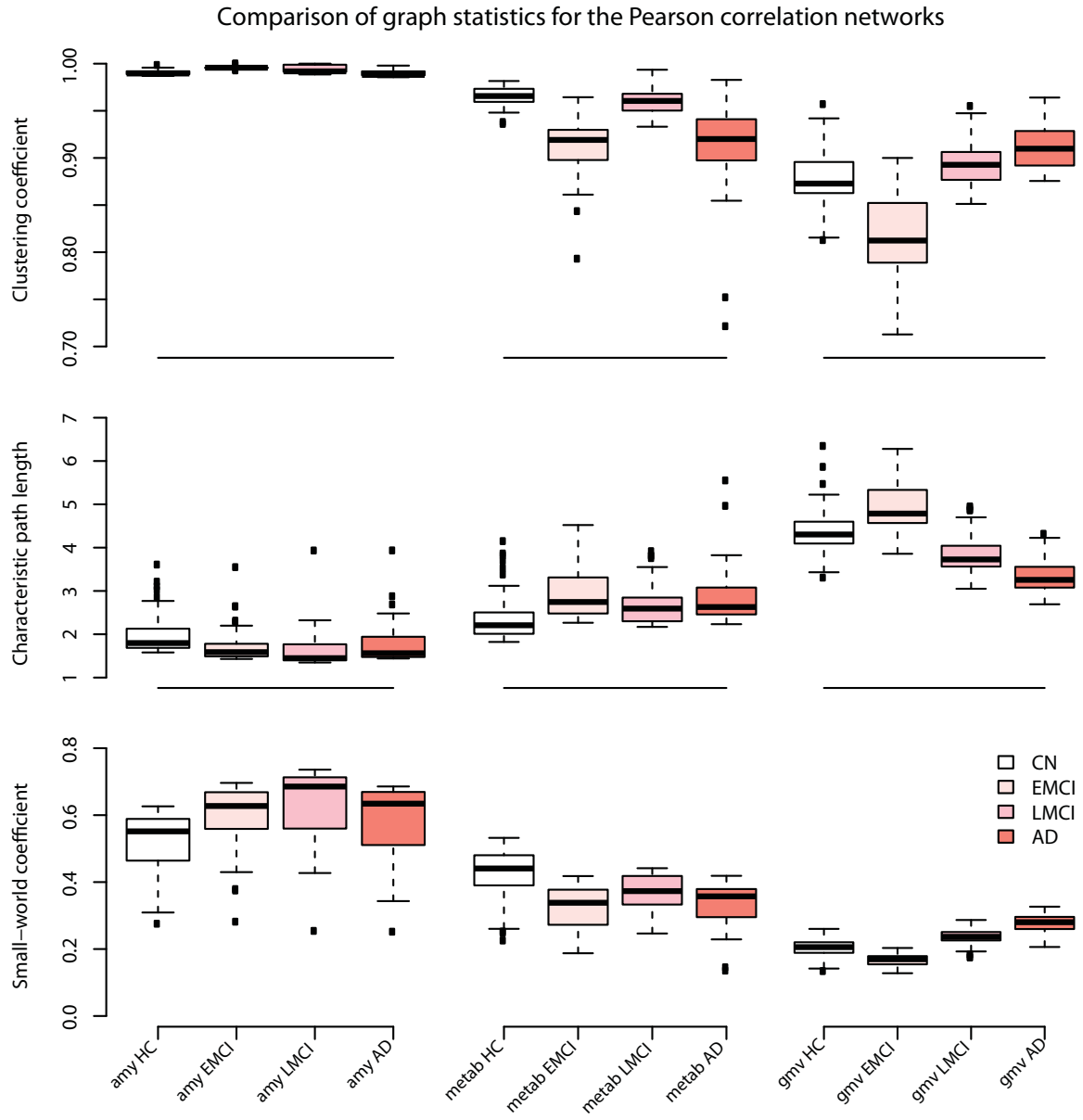


Figure S9. Comparison of graph statistics for the Pearson correlation matrices of the left hemisphere stratified by diagnostic group and image modality. The distribution of the weighted clustering coefficient, characteristic weighted path length, and small-world coefficient for individual brain regions is shown. Boxes display median, first and third quartile of the distributions, and whiskers indicate $\pm 1.5\times$interquartile range. Prior to calculating the graph measures, the correlation matrices were thresholded such that correlations with $p > 0.05$, i.e. approximately $r < 0.12$, were set to zero.
CN: cognitively healthy elderly controls, EMCI/LMCI: early and late amnestic mild cognitive impairment, AD: Alzheimer's dementia, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.

Table S1. P-values for the comparison of graph statistics based on Pearson correlation (Figure S9).

		Amyloid-β			Metabolism			Volume		
		EMCI	LMCI	AD	EMCI	LMCI	AD	EMCI	LMCI	AD
Clustering coefficient	CN	< 0.001	< 0.001	0.575	< 0.001	0.759	< 0.001	< 0.001	0.103	< 0.001
	EMCI		< 0.001	< 0.001		< 0.001	0.973		< 0.001	< 0.001
	LMCI			< 0.001			< 0.001			0.020
Path length	CN	0.013	< 0.001	0.026	< 0.001	0.051	< 0.001	< 0.001	< 0.001	< 0.001
	EMCI		0.437	0.996		0.113	0.857		< 0.001	< 0.001
	LMCI			0.315			0.464			< 0.001
Small-world coefficient	CN	< 0.001	< 0.001	0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	EMCI		0.085	0.991		0.004	0.797		< 0.001	< 0.001
	LMCI			0.040			0.055			< 0.001

Adjusted P-values from Tukey's honest significant difference tests, controlling for family-wise error rate within each comparison block. CN: cognitively normal controls, EMCI/LMCI: early/late amnestic mild cognitive impairment, AD: Alzheimer's dementia.

Table S2. Analysis of variance (ANOVA) results for the graph statistics for the partial correlation networks in Figure 9.

		F-statistic	P-value	Effect size η^2
Clustering coefficient	amy	5.6	0.001	0.07
	metab	6.2	< 0.001	0.08
	vol	4.6	0.004	0.06
Characteristic path length	amy	5.1	0.002	0.07
	metab	12.8	< 0.001	0.15
	vol	11.8	< 0.001	0.14
Small-world coefficient	amy	5.6	0.001	0.07
	metab	8.6	< 0.001	0.11
	vol	4.1	0.007	0.06

df=215 for all models, amy: amyloid-β, metab: glucose metabolism, vol: gray matter volume.