NeuroImage: Clinical 43 (2024) 103624

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Cortical thickness and grey-matter volume anomaly detection in individual MRI scans: Comparison of two methods

David Romascano [a,b], Michael Rebsamen [a], Piotr Radojewski [a,c], Timo Blattner [a], Richard McKinley [a], Roland Wiest [a,c], Christian Rummel [d,a,*]

[a] Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, University Hospital Bern, CH-3010 Bern, Switzerland

[b] Danish Research Center for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark

[c] Translational Imaging Center (TIC), Swiss Institute for Translational and Entrepreneurial Medicine, sitem-insel, Bern, Switzerland

[d] European Campus Rottal-Inn, Technische Hochschule Deggendorf, Max-Breiherr-Straße 32, D-84347 Pfarrkirchen, Germany

ARTICLE INFO

Keywords:
Normative modeling
MRI
Brain morphometry
Alzheimer's disease
Personalized medicine
Clinical decision support

ABSTRACT

Over the past decades, morphometric analysis of brain MRI has contributed substantially to the understanding of healthy brain structure, development and aging as well as to improved characterisation of disease related pathologies. Certified commercial tools based on normative modeling of these metrics are meanwhile available for diagnostic purposes, but they are cost intensive and their clinical evaluation is still in its infancy. Here we have compared the performance of "ScanOMetrics", an open-source research-level tool for detection of statistical anomalies in individual MRI scans, depending on whether it is operated on the output of FreeSurfer or of the deep learning based brain morphometry tool DL + DiReCT. When applied to the public OASIS3 dataset, containing patients with Alzheimer's disease (AD) and healthy controls (HC), cortical thickness anomalies in patient scans were mainly detected in regions that are known as predilection areas of cortical atrophy in AD, regardless of the software used for extraction of the metrics. By contrast, anomaly detections in HCs were up to twenty-fold reduced and spatially unspecific using both DL + DiReCT and FreeSurfer. Progression of the atrophy pattern with clinical dementia rating (CDR) was clearly observable with both methods. DL + DiReCT provided results in less than 25 min, more than 15 times faster than FreeSurfer. This difference in computation time might be relevant when considering application of this or similar methodology as diagnostic decision support for neuroradiologists.

1. Introduction

Many pathological processes affecting the central nervous system (CNS) have an impact on its structural organization. Various forms of brain morphometry have made it possible to describe brain shape mathematically, yielding variables for statistical evaluation, which have made important contributions towards a better understanding of healthy brain development and aging as well as to disease manifestation and mechanisms (see e.g. Mills et al., 2021; Statsenko et al., 2022; McCutcheon et al., 2023; Joy et al., 2023 for recent examples). Large group studies have demonstrated that metrics derived from routine structural MRI scans are sensitive to pathological brain changes (see e.g. Whelan et al., 2018; Laansma et al., 2021). For this reason, brain morphometric variables have been included as outcome measures in recent clinical trials (e.g. National Library of Medicine [NLM], NCT04860947 for the prediction of disease progression in multiple sclerosis, National Library of Medicine (U.S.), 2019, or NLM NCT06155942 for the use of morphometry as a biomarker for Parkinson's disease, National Library of Medicine (U.S.), 2024).

Surface based analysis (SBA) is a variant of brain morphometry, that attempts to represent the two-dimensional geometry of the cortex by tesselating the interface between white matter (WM) and gray matter (GM) with a mesh and estimating region specific metrics like the GM volume (GMV), cortical surface area (CSA) or cortical thickness (CTh). During the last two decades, substantial efforts have been invested into providing software to extract precise and accurate SBA metrics from MRI

* Corresponding author at: Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, University Hospital Bern, CH-3010 Bern, Switzerland.
E-mail address: crummel@web.de (C. Rummel).

https://doi.org/10.1016/j.nicl.2024.103624
Received 19 February 2024; Received in revised form 21 May 2024; Accepted 25 May 2024
Available online 28 May 2024

scans. For research purposes, FreeSurfer (Dale et al., 1999; Fischl et al., 1999a, 1999b; Fischl & Dale, 2000) has become the most widely used automated tool. Among its advantages are its free availability and extremely high acceptance and understanding by the community, which has led to more than 2′800 scientific publications (PubMed search on 2024/02/17).

Sensitivity of SBA metrics to pathological processes has been mostly established through cross-sectional and longitudinal *group studies* (see e.g. de Figueiredo et al., 2021; Alkan et al., 2021; Nkrumah et al., 2023; Fortea et al., 2023 for recent examples). In contrast, normative modeling aims at a quantitative evaluation of *single subject* scans by establishing healthy developmental trajectories and prediction intervals of SBA metrics in a reference population. It is a powerful tool to detect statistical anomalies at the individual level, making it much better suited to support personalized diagnostics and decision making (Marquand et al., 2016, 2019; Potvin et al., 2017; Ge et al., 2023; Potvin, 2021). In the meantime, CE-marked and FDA-approved commercial tools for clinical decision support by brain morphometry and normative modeling have become available for application in various forms of dementia (Pemberton et al., 2021) and in patients with MS (Mendelsohn et al., 2023).

To provide reliable predictions, the models should be derived from large normative databases (Rutherford et al., 2022). In the field of MRI, suitable datasets have recently become available as public resources and open doors towards the application of normative models in clinical settings. Since MRI acquisition settings like scanner type (Sinnecker et al., 2022) or scanning protocol (Rebsamen et al., 2023b) have been demonstrated to influence SBA estimates, control for these confounders by harmonization procedures is required (see e.g. Fortin et al., 2018). Our own work in the direction of normative modeling has demonstrated screening test characteristics of automated regional SBA metrics in patients with temporal lobe epilepsy (i.e. large negative predictive values, while positive predictive values were only moderate; Rummel et al., 2017) and provided markers for regional atrophy progression in patients with multiple sclerosis (Rummel et al., 2018).

One of the remaining obstacles hindering the use of SBA normative modeling as a decision support tool in the clinical routine is the long computation time required for tools like FreeSurfer to process a single MRI scan, which is in the order of ten hours on the central processing unit (CPU) of a current standard desktop computer. Indeed, to practically contribute information to clinical diagnostics, processing times should ideally be reduced to the order of minutes, to enable patient evaluation on demand or at least within the same shift. To overcome this limitation, new tools leveraging deep learning (DL) and convolutional neural networks (CNN) running on graphical processing units (GPU) have become available for SBA, like for example FastSurfer (Henschel et al., 2020). DL + DiReCT (Rebsamen et al., 2020, 2023a) and CortexMorph (McKinley & Rummel, 2023) are alternative approaches to DL-based estimation of CTh. A recent comparative study revealed that not only did DL + DiReCT substantially outperform FreeSurfer in terms of computation time required to estimate CTh, but it also provided comparable scan-rescan reproducibility and estimated atrophy rates (Rebsamen et al., 2020). Importantly, DL + DiReCT was shown superior to FreeSurfer (both cross-sectional and longitudinal) in terms of sensitivity to simulated cortical thinning, especially when the introduced atrophy was weak (Rusak et al., 2022).

The purpose of this work was to explore the performance of our normative modeling approach ("ScanOMetrics", Rummel et al., 2017, 2018) on metrics derived from DL + DiReCT and FreeSurfer, in the context of clinical evaluation. To achieve full reproducibility of our results, we focussed the analysis on OASIS3 (Open Access Series of Imaging Studies; LaMontagne et al., 2019), a large and freely available dataset containing clinical grade high-resolution isotropic T1-weighted MRI scans of patients with Alzheimer's disease (AD) and healthy controls (HC). This extends previous work on DL + DiReCT and normative modelling to a reference database of several thousand scans instead of hundreds, and to their use for the evaluation of AD scans. We restricted our software comparison to the jointly available SBA metrics of the Desikan-Killiany atlas (Desikan et al., 2006), namely regional cortical GMV as well as regional mean and standard deviation of the CTh. The ability to detect regional outliers was compared between the two processing tools and the accumulation of anomalies in brain regions that are known for atrophy in AD group studies was studied, effectively assessing whether normative modeling based on a faster morphometry tool (DL + DiReCT) provides evaluation metrics that are consistent with a slower but widely used and validated tool (FreeSurfer).

Our hypotheses were the following: Based on results by Rusak et al. (2022) and ourselves (Rebsamen et al., 2020), we expected normative models within ScanOMetrics to provide (1) more narrow distributions of fit residues, (2) higher scan-rescan reproducibility, as well as (3) more pronounced and more specific atrophy patterns in patients when using DL + DiReCT instead of FreeSurfer metrics. Based on previous work using PET and MRI imaging (Jansen et al., 2022; Verdi et al., 2023), we expected that (4) the AD group would yield a higher percentage of individual scans labeled as anomalous than a leave-one-out cross-validation (LOOCV) in the HC group. Finally, we hypothesized that (5) normative modeling at the level of individual scans/patients shows heterogeneous anomaly patterns. When averaging the individual anomaly maps over the whole group, the shared anomaly motifs should, however, be similar to the map obtained when testing for statistical differences between the entire AD and HC groups (i.e. effect sizes in a group analysis).

2. Materials and methods

All software tools used in this paper are open-source. The Python3 implementation of ScanOMetrics is available at https://github.com/SCAN-NRAD/ScanOMetrics. A code description is given in the Supplementary Materials and a more detailed documentation with tutorial is available at https://scanometrics.readthedocs.io. FreeSurfer can be downloaded from https://surfer.nmr.mgh.harvard.edu/ and DL + DiReCT is available at https://github.com/SCAN-NRAD/DL-DiReCT.

2.1. OASIS3 dataset

The OASIS3 dataset (Open Access Series of Imaging Studies, LaMontagne et al., 2019) is publicly available at www.oasis-brains.org/ and contains NIFTI files of 2′643 high-resolution (voxel sizes in the order of 1 mm x 1 mm x 1 mm) isotropic T1-weighted MRI scans from 1′038 participants. All scans were acquired at two field strengths using Siemens MRI scanners: Magnetom Sonata and Avanto (1.5 T, 42 scans) as well as Biograph mMR and Magnetom Trio (both 3 T, 2′601 scans). 2′014 scans are from subjects considered HCs with normal cognition (clinical dementia rating CDR = 0), 420 scans are from undetermined cases with CDR = 0.5, and 209 scans correspond to patients with established AD having CDR $\geq$ 1, leading to 629 scans with CDR > 0.

Of the 2′014 HC scans, 87 are from 41 subjects that had a mixture of scans with CDR = 0 and CDR $\geq$ 0.5 ("converters" between normal cognition and suspicion or established impairment). Those scans were excluded from building our normative model, which was therefore based on 1'927 scans from 696 non-converting subjects, see Table 1 for demographic information. The 41 'converter' subjects were instead used to investigate the change trajectories between the first and follow-up scans longitudinally.

2.2. SBA metric computation and normalization

All MRI scans were processed with Ubuntu Linux 22.04.3 LTS on a Dell Precision 7920 workstation with the following specifications. CPU: two Intel Xeon Gold 6148, each one equipped with 20 cores and 2.4 GHz processor base frequency, RAM: 256 GB, GPU: one NVIDIA GeForce GTX 1080 with 8 GB memory. SBA metrics were derived from FreeSurfer (Dale et al., 1999; Fischl et al., 1999a, 1999b; Fischl & Dale, 2000),

Table 1
Demographic characteristics of the used OASIS3 subgroups. To increase clarity, 60 scans (2.2 %) from 15 patients with CDR changing between different levels of CDR $\geq$ 0.5 are not included here.

	CDR = 0, 'non-converters' used for normative models	subjects with CDR = 0 and $\geq$ 0.5, 'converters' used for follow-up analysis	CDR = 0.5, 'undetermined' cases	CDR $\geq$ 1, 'established' dementia
participants (female)	696 (419, 60.2 %)	41 (19, 46.3 %)	174 (85, 48.9 %)	112 (56, 50.0 %)
scans (female)	1′927 (1′188, 61.7 %)	167 (84, 50.3 %)	309 (151, 48.9 %)	180 (91, 50.6 %)
age at scan (years, mean + -SD and range)	69.0 ± 9.3, (42.7—97.0)	74.0 ± 8.2 (54.0–94.4)	76.1 ± 7.2 (51.7–94.4)	74.2 ± 8.4 (50.3–95.6)

version 6.0.0 and DL + DiReCT (Rebsamen et al., 2020, 2023a) using default parameters. Results were exported in tabular form using the Desikan-Killiany atlas (Desikan et al., 2006). Because the current implementation of DL + DiReCT does not provide other SBA metrics, only the cortical GMV, mean and standard deviation of the CTh were included in our study. Structures with bilateral representations were used to compute an asymmetry index. In summary, for both processing pipelines, each scan yielded a total of 358 'raw' measurements:

- subcortical volumes: 8 structures (thalamus proper, caudate, putamen, pallidum, accumbens area, hippocampus, amygdala and ventral diencephalon) on 2 hemispheres plus 8 asymmetry indices
- 3 volumes of midline structures (brain stem, 3rd and 4th ventricles)
- cortical regions of the Desikan-Killiany atlas: 3 metrics for 34 regions on 2 hemispheres plus 3x34 asymmetry indices
- brain lobes: volumes for 6 lobes (frontal, parietal, occipital, temporal, cingulate and insula) on 2 hemispheres plus 6 asymmetry indices. Mean and standard deviation of CTh were not included here, since a size-weighted lobar aggregation requires an estimate of the CSA, which is currently not provided by DL + DiReCT.
- brain hemispheres: left/right cortex volume and mean CTh plus asymmetry indices
- whole brain: estimate for intracranial volume (ICV).

In addition to the 'raw' metrics, we used 'normalized' variants to account for the fact that most metrics vary with brain size (Potvin et al., 2017). All volumes were scaled to the mean ICV of the normative dataset. Mean and standard deviation of the CTh were instead scaled isometrically according to ICV^(1/3) to respect the geometry of the cortex as a thin two-dimensional sheet, which is folded into three-dimensional space, see (Rummel et al., 2017, 2018) for details. As estimates for ICV we used the Estimated Total Intracranial Volume (eTIV) for FreeSurfer and an exhaustive volume sum of all intracranial segmentations for DL + DiReCT. Since ICV normalized by itself has the same value for all scans and asymmetry indices do not change under the normalization procedure, we obtained 239 additional 'normalized' metrics.

2.3. Uniform age sampling

Deviating from the original procedures described in detail in (Rummel et al., 2017, 2018), each one of the 597 SBA metrics (raw plus normalized) extracted from all 1′927 scans with CDR = 0 was resampled 100 times by creating 10-bin-histograms of the participant age and drawing n_{min} random samples from each bin, where n_{min} was the smallest bin count. For uniform age distributions, this procedure has no effect, whereas non-uniform age distributions are rendered approximately uniform.

2.4. Normative modeling

Normative models were built for each software and metric independently according to the pipeline of (Rummel et al., 2017, 2018). In brief, low order polynomials were fitted to the 100 resamples of the SBA metrics of our HCs as a function of age. The degrees of the fit polynomials were adapted for each of the resamples separately by increasing from zero until the reduction of residual variance became insignificant (nested F tests). To exclude overfitting, the maximum degree was set as the odd number 2*floor(ln(n/10) + 1)-1, where n is the available number of samples (Rummel et al., 2010). For example, when using all 1′927 scans with CDR = 0, the maximal allowed degree was 11. The polynomial age trend and prediction intervals were finally computed from the average of all fits to the 100 resamples. Before each of these fits, outliers were removed based on whether they exceeded the 25th or 75th percentile of the distribution by more than 1.5 inter quartile ranges. This procedure was repeated for the fit residues, before a final age fit to the retained data points was generated in the same manner. Metric variability at a given age was computed by a combination of metric variance over subjects within 10 % of the age of interest, and measurement uncertainty derived from repeated scans in the reference dataset (for more details, see the Supplementary Materials, and Rummel et al. 2018).

2.5. Evaluating patient data against the normative models

With the normative age models available, we applied them to patient scans and compared their fit residues to the distribution in the HCs. Covariates other than age (i.e. sex, scanner and scanning protocol) were accounted for by selecting matched subgroups before computing statistics. Matching for scanning protocol allows to reduce variability in thickness estimation due to imaging parameters and corresponding differences in WM/GM contrast (Rebsamen et al., 2023b). Since this matching yielded variable group sizes, the probability P of finding a fit residue of the observed size was calculated accounting for the distribution in the matching HCs and the uncertainty of the measurement, see Rummel et al. (2017, 2018) for details. For individual scans, an initial z-score was computed to position the individual metric with respect to the matching and scans in the reference set that were not previously flagged as outliers. This z-score was divided by a variance estimate that takes into account both the standard deviation of metrics within the reference dataset, as well as the measurement uncertainty established from repeated scans. The resulting z-score was then converted to p-value using a standard z to p transformation, involving the cumulative distribution of a Gaussian. To account for metric and region specific measurement uncertainties, these were estimated based on repeated scans of the same HC within an age change of less than 10 %. Note that compared to same-session rescans under identical conditions, this estimate yields only an upper bound of the true uncertainty. Finally, log10(P), signed positive/negative for larger/smaller than expected fit residues, were used as the central objects to decide whether a regional metric was classified as statistically normal or abnormal. To detect a statistical anomaly, a significance threshold was set to q = 0.01, equivalent to $-\log10(P) > 2$.

2.6. Inspection of scans with extreme atrophy

DL based tools might underperform when evaluating scans that are too different from their training set. To assess whether one of the two used software tools systematically mislabeled scans with extreme atrophy, the three scans with the smallest mean log(p) score for right and left hemisphere mean thickness were identified, for both FreeSurfer and DL + DiReCT and the cortical log(p) maps of both softwares were

compared. We further inspected individual FreeSurfer and DL + DiReCT parcellations, focusing on regions where both tools strongly disagreed on log(p) values.

2.7. Comparison to other normative tools

Z-scores obtained when evaluating right and left mean CTh were compared to scores obtained using the PCN toolkit (Rutherford et al., 2022). Signed log(p) values for the average CTh were also compared to centiles obtained from Brain Chart (Bethlehem et al., 2022), another open-access tool for normative modelling. CTh for the HCs were submitted along the AD subjects for both PCN toolkit and BrainChart. Since both PCN toolkit and BrainChart rely on FreeSurfer metrics, and DL + DiReCT log(p) values correlated strongly with FreeSurfer, we refrained from including DL + DiReCT metrics in this comparison.

2.8. ROC curves

To explore the separation of the AD and HC groups, receiver operating characteristics (ROC) and areas under the curve (AUC) were estimated separately for DL + DiReCT and FreeSurfer. The percentage of abnormal metrics per scan (p-values below 0.01) was taken to assess scans as a whole. To focus on brain regions that are known to be affected in AD patients, a similar analysis was repeated for the signed log10(p) values of the CTh of the entorhinal cortex and the hippocampal GMV.

2.9. Comparing spatial patterns

Significance maps and anomaly maps of individuals or groups were compared using normalized L2-distances. L2 was used instead of the Pearson correlation coefficient, because the latter is invariant to shift and scale, which we want to account for when ranking individual maps relative to a template.

2.10. Evaluating and cleaning the normative dataset

The normative dataset was evaluated with a subject-wise leave-one-out cross-validation (LOOCV) study, building normative polynomial models under exclusion of a specific HC (all sessions and repeated scans) and testing all scans of the excluded subject against that model, similar to what was described above for patients. To test for normality of fit residues in our LOOCV, Shapiro-Wilk tests were performed on each of the 597 metrics separately. To test whether the number of detections during the LOOCV was abnormally high over all subjects and 'raw' metrics, we performed a binomial test with the number of positives given by the number of anomaly detections, the number of samples given by the number of metrics times the number of scans and the expected fraction of random outliers given by the significance threshold q = 0.01.

To clean our normative models from scans with artifacts or potential pathologies before final application, the LOOCV analysis was in addition used to identify anomalous scans separately for the DL + DiReCT and FreeSurfer pipelines and remove them from the normative datasets. We considered scans as not (entirely) normal if they yielded p-values lower than q = 0.01 for 18 or more out of the 358 raw metrics (5 % of metrics). As a final step, the LOOCV procedure was repeated after cleaning of the normative dataset. The patient evaluation described in the previous paragraph was done against the clean normative dataset.

3. Results

We first present results from the AD patient evaluation, followed by some more technical results regarding LOOCV evaluation of HC subjects and dataset cleaning required before patient evaluation.

3.1. Application to AD patients

The AD dataset consisted of 209 scans with CDR $\geq$ 1.0. When using the clean HC dataset to evaluate AD scans, both processing pipelines indicated increased proportions of anomalous scans (i.e. scans with more than 18 abnormal raw regional metrics out of 358, equivalent to 5 %) in the AD dataset compared to HC. DL + DiReCT resulted in 117 anomalous AD scans (56.0 % of all AD scans, compared to 1 % in the clean HC dataset), whereas more scans were classified as anomalous using FreeSurfer (129 CE scans, 61.7 %, compared to 0.8 % in the clean HC dataset). Details regarding anomaly detection rates in HC can be found in the section "Anomaly detection in healthy controls (LOOCV)" below.

Fig. 1 compares the regional percentage of statistical CTh anomalies detected by ScanOMetrics in individual scans (with significance P < q = 0.01, not corrected for multiple comparisons) in patients with AD as well as in the cleaned HC dataset. The patterns of preferred anomaly detection are remarkably similar between both processing tools and symmetric with respect to hemispheres. Comparison of the CTh reduction map in patients with AD (third row) with the effect size map of a direct statistical comparison between the AD and HC groups (Cohen's d, bottom row) displays remarkable agreement of the temporo-parietal atrophy patterns. In patients with AD, reduction of regional mean CTh is detectable in up to 28 % of individual scans with a strong regional preference for the bilateral entorhinal and fusiform cortex as well as in the precuneus and supramarginal gyrus. In the frontal lobe the CTh reduction is weakest. For HCs the peak percentage of detected CTh reductions is only in the order of ~ 1.3 %, i.e. twenty-fold reduced when compared to patients with AD. Increase of CTh is also observed in up to ~ 4.5 % of patients with AD, with peak in the bilateral medial orbito-frontal gyrus and cuneus. Supplementary Fig. S1 shows results equivalent to Fig. 1 when using raw thickness values (i.e. without scaling for brain size). In general, the percentage of subjects with increased thickness had similar spatial patterns, but atrophy in AD subjects was slightly stronger and more widespread. When using raw values, group effects were also slightly larger, including small regions with increased thickness in AD patients.

When stratifying scans from patients with AD by the clinical dementia rating (CDR), an apparent worsening of atrophy along the temporal, parietal and eventually frontal lobe regions is revealed by the CTh anomaly maps, see Fig. 2. Using DL + DiReCT, abnormal mean entorhinal CTh is detected already in about 24.8 % of patients with CDR = 0.5 (N = 416), which progresses to 47.2 % of patients with CDR = 1 (N = 159), and 46.0 % of patients with CDR $\geq$ 2 (N = 50). In contrast, using FreeSurfer, abnormal thickness is detected in 17.6 % of CDR = 0.5 patients, 30.2 % of CDR = 1 patients, and 32.0 % of CDR $\geq$ 2 patients. For CDR $\geq$ 1, CTh reduction becomes visible in the precuneus and supramarginal gyrus as well. For cases with CDR $\geq$ 2 also the fusiform gyrus (20 % of cases for DL + DiReCT and 34 % of cases for FreeSurfer) and the lateral temporal lobes are affected. In the OASIS3 dataset an increase of CTh in the bilateral medial orbito-frontal gyrus is observable and associated with increasing CDR, an effect which is clearer visible with FreeSurfer than with DL + DiReCT.

Fig. 3a compares normalized L2-distances between ScanOMetrics' individual significance maps of all scans with CDR $\geq$ 0.5 and the average significance map of all scans with CDR $\geq$ 2, which was used as a template for clear AD. Estimates from DL + DiReCT and FreeSurfer were found highly correlated (r = 0.87, p < 1e-16). Fig. 3b shows histograms of the L2-distances, separately for DL + DiReCT and FreeSurfer, grouped by increasing CDR and revealing a negative association for both tools. Fig. 3c presents examples of individual significance maps for five different scans. Selection was made based on quantiles of the normalized L2-distances shown in panels a and b. Interestingly, the scan closest to the CDR $\geq$ 2 template was the same one for DL + DiReCT and FreeSurfer (corresponding to the lowest left datapoint in Fig. 3a, a scan with CDR = 0.5). Fig. 3c illustrates at the same time how diverse significance maps

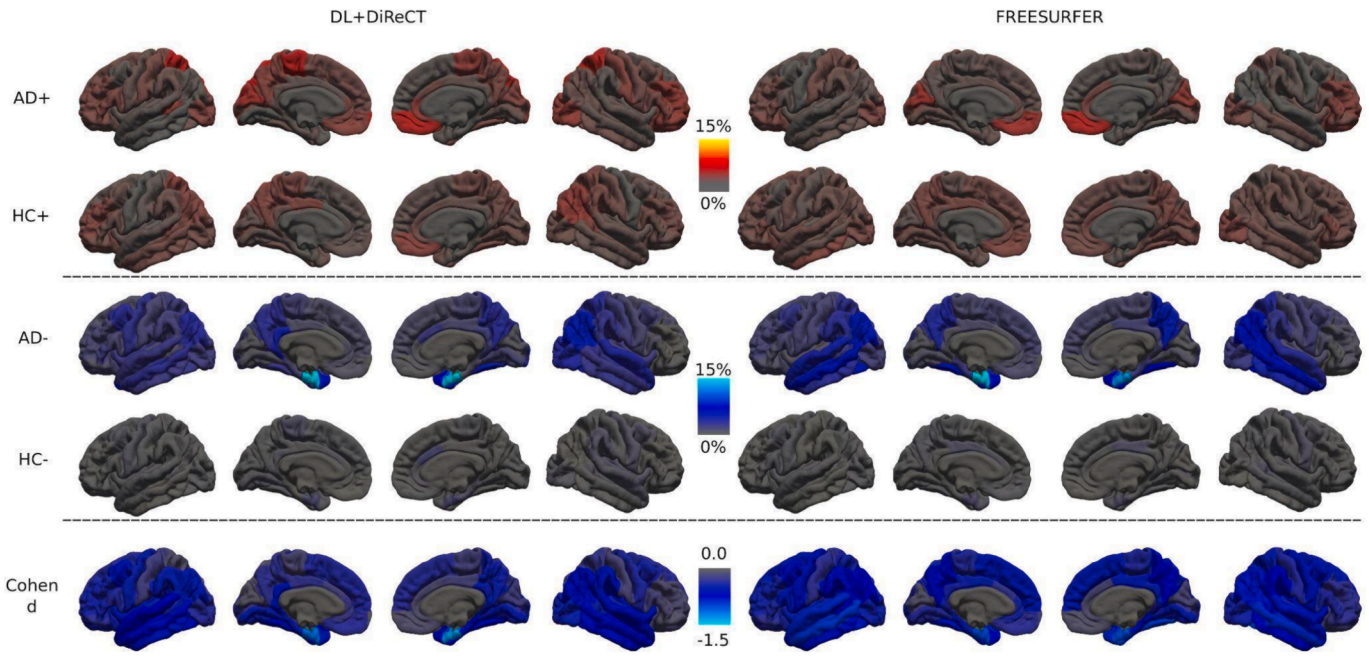

Fig. 1. Percentage of CTh anomalies in the AD and HC groups, detected with ScanOMetrics using both processing tools. AD patients with established dementia (CDR $\geq$ 1, 209 scans) are shown in rows 1 and 3, results of the LOOCV in cleaned non-converting HCs (CDR = 0, 1′828 scans) in rows 2 and 4. Deviations towards larger (rows 1 and 2, red-to-yellow colormap) and smaller (rows 3 and 4, blue-to-white colormap) than expected CTh are collected separately. The bottom row shows the effect size (Cohen's d) when contrasting the entire AD and HC groups. Positive effect sizes did not occur in this comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

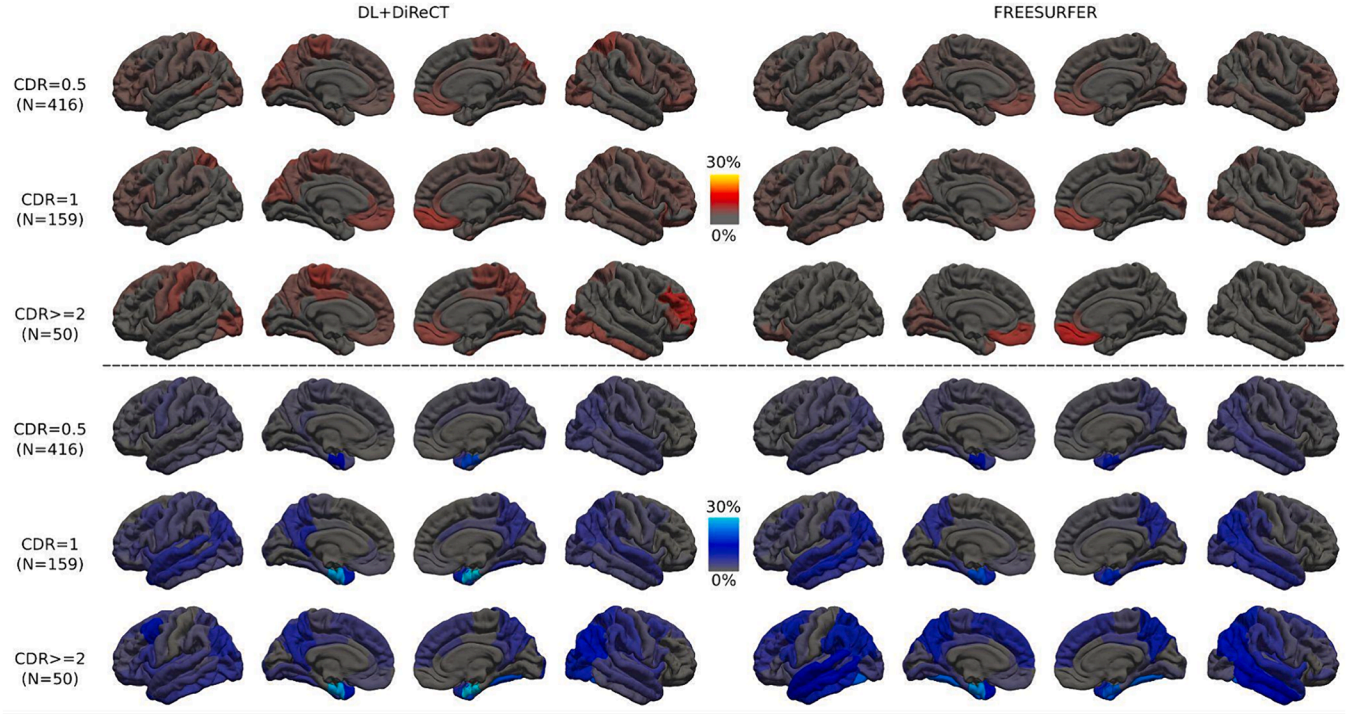

Fig. 2. Percentage of CTh anomalies detected by ScanOMetrics in patients with AD, stratified by cognitive impairment levels at scan time (CDR = 0.5: rows 1 and 4, CDR = 1: rows 2 and 5, and CDR $\geq$ 2: rows 3 and 6). The upper half depicts CTh increase, while the lower half shows progression of CTh reduction. Mind that the color scales are different from the ones used in Fig. 1.

can look like in different patients (top and middle row), how similar anomaly detection can be for both software tools (bottom row), and how loosely individual clinical scores and corresponding significance maps in structural MRI scans can be related (the scan closest to the AD template generated from CDR $\geq$ 2 cases has a CDR of only 0.5).

In agreement with published results (van Hoesen et al., 1991; Gómez-Isla et al., 1996; Juottonen et al., 1999; Du et al., 2001; Price et al., 2001; Mueller et al., 2010; Devanand et al., 2012; Igarashi, 2023), the bilateral entorhinal gyrus was identified as one of the earliest visible and most prominent deviations in patients with AD from the normative

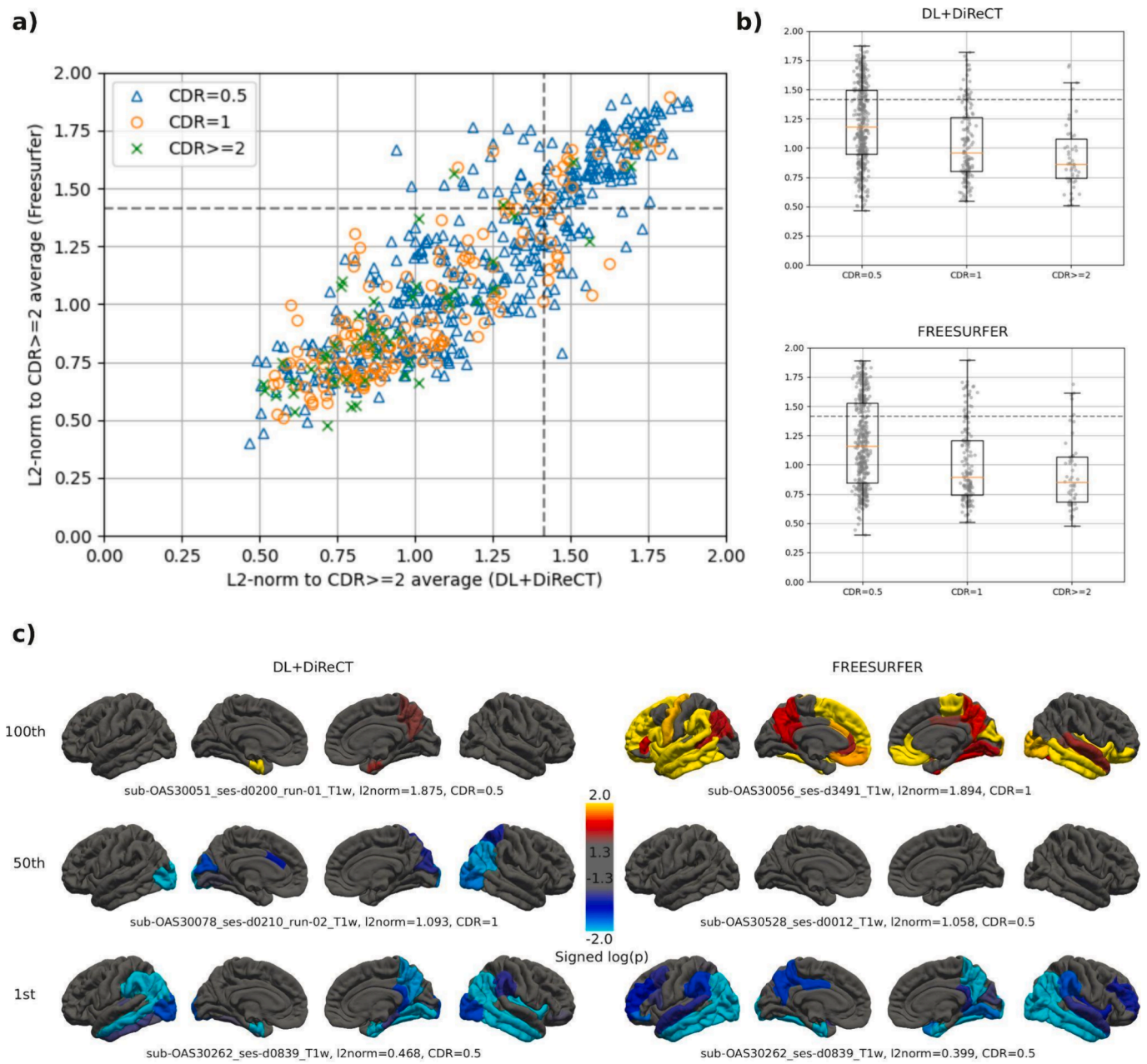

Fig. 3. Normalized L2-distances between individual significance maps and a template (i.e. the average map of all scans with CDR $\geq$ 2). Significance maps are log10 (p) maps with negative sign for CTh reduction and positive sign for CTh increase. Normalized L2-distances range from 0 for identical maps to 2 for antisymmetric maps, with sqrt(2) indicating orthogonal maps (marked by dotted lines in panels a and b). a) Correlation between DL + DiReCT and FreeSurfer, individual CDR scores are symbol/color coded. b) Grouping by CDR separately for both software tools. c) Significance maps in individual scans, selected according to their L2-distance. Scans on the 1st row are the 100th percentiles in the distributions (i.e. highest distance to the reference), while the lowest row are the most similar to the group average. In the lowest row the same scan was selected for both DL + DiReCT and FreeSurfer, and corresponds to the data point closest to the origin in panel a).

model, see Figs. 1 and 2. Hippocampal volume has also been reported to be prominently atrophic in AD (Juottonen et al., 1999; Du et al., 2001; Sluimer et al., 2008; Devanand et al., 2012). In Fig. 4 we display the mean normalized volume of the hippocampus (as provided as ScanO-Metrics output based on DL + DiReCT estimates) for the scan with the highest individual similarity with the AD group (i.e. the lowest row in Fig. 3c) and compare with the point cloud of the cleaned normative dataset. Hippocampal volume is in the order of only 2.5 ml on both hemispheres, much below the 95 % prediction interval [3.2 ml, 4.5 ml] estimated from our HCs at the same age. Furthermore, the hippocampal volume was found to decrease bilaterally from the first to the second scan available for this patient. Atrophy rate for the average volume of both hemispheres was 12.4 % over a period of 2.2 years. When using FreeSurfer (see Supplementary Fig. S2), the estimated atrophy rate was lower (6.7 % over 2.2 years).

We used the 87 MRI scans of the 41 subjects that were excluded from building the normative models (conversion from CDR = 0 to CDR $\geq$ 0.5) to investigate the change of thickness over time in more detail. To focus on the clinically relevant question of early atrophy detection, we restricted this analysis to participants where a scan with CDR = 0 was available, excluding any progression between higher CDR levels. Difference maps of mean regional CTh (ICV normalized, later scans minus baseline always, regardless the associated CDR values) were averaged over all scan pairs of the selected 41 subjects and are displayed in Fig. 5. Similar to Fig. 2, where progression is displayed by grouping according to CDR, the most prominent atrophy progression over time occurred in

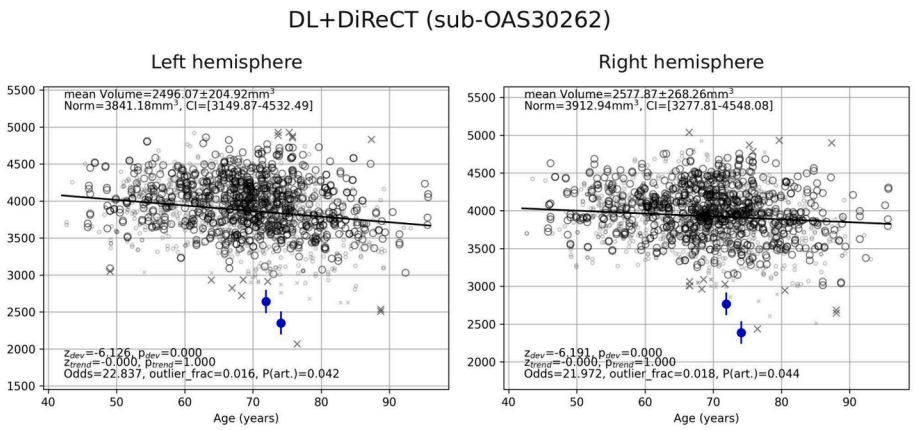

Fig. 4. Age dependence of the brain size normalized volume of the hippocampus, as displayed by ScanOMetrics (volume estimates by DL + DiReCT). The corresponding data derived from FreeSurfer is available in our Supplementary Fig. S2. Similar results were observed for the normalized CTh of the entorhinal cortex (not shown).Patient data (blue) are the two scans of the participant closest (i.e. had the smallest L2 norm) to the AD group's average significance map (log10(p) maps for the second scan are shown at the lower left section of Fig. 3c). Symbols in black represent the HCs used to build the cleaned normative dataset. Crosses are estimates flagged during outlier removal and did not contribute to statistics. Large symbols match the patient scans regarding sex, MRI scanner type and scanning protocol, whereas small symbols differ in at least one of these characteristics. Fully drawn lines indicate the fitted age trajectory of the normative models. Significance of statistical comparisons and the reliability of the measurement (see Rummel et al., 2017, 2018 for details) are reported in the lower left corners of the panels. Values reported in the upper left corner are the subject average across time points, along with the expected value from normative data and its prediction interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

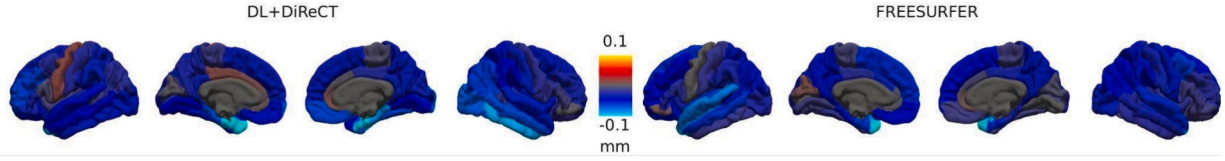

Fig. 5. Average change in mean normalized regional CTh in subjects converting between CDR = 0 and CDR ≥ 0.5. In contrast to Figs. 1, 2 and 3c changes are measured in millimeters here.

temporo-basal brain regions, like the entorhinal, parahippocampal, fusiform and inferior temporal gyrus, where mean CTh reduced up to 0.1 mm, a change equivalent to the expected reduction of whole brain mean CTh in 25 years of healthy aging (Lemaitre et al., 2012). Also remarkable is the relative sparing of the somato-sensory cortex from atrophy progression (Thompson et al., 2003; Lerch et al., 2005; Fennema-Notestine et al., 2009; Frisoni et al., 2010; Rebsamen et al., 2020), which becomes most transparent in the left precentral gyrus in Fig. 5 but can be identified in individual scans of Fig. 3 and in the percentage maps of Figs. 1 and 2 as well.

3.2. Scan classification

Classifying scans as AD/abnormal based on the percentage of metrics with p-value below 0.01 lead to AUCs of 0.76 for DL + DiReCT and 0.72 for FreeSurfer (Fig. 6). Sensitivity and specificity were the closest to the top-left corner when using a threshold of 1.04 % for DL + DiReCT (FPR = 0.29, TPR = 0.69) and 0.62 % for FreeSurfer (FPR = 0.35, TPR = 0.69). Instead, using a fixed threshold of 5 % abnormal metrics to label a scan as abnormal (i.e. the threshold used to clean the original dataset) lead to FPR = 0.03 and TPR = 0.28 for DL + DiReCT, while the rates were 0.04 and 0.20 for FreeSurfer. Similar results were obtained for the attempt to classify scans based on the signed log10(p) value of the CTh of the entorhinal cortex (DL + DiReCT slightly better, see Supplementary Fig. S6) or of the hippocampal GMV (FreeSurfer slightly better). Both tools had the same discriminant power when using the suitable metric (AUC = 0.75).

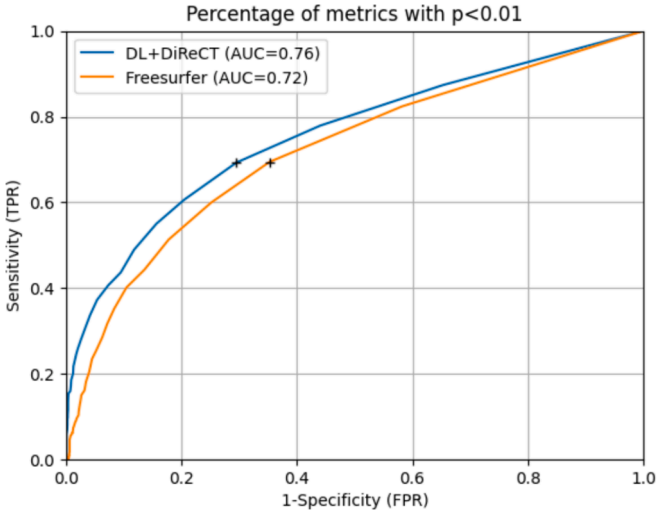

Fig. 6. Receiver operating characteristics (ROC) for classification of scans into AD and HC, based on their percentage of abnormal metrics. Patients with AD were evaluated against the normative model of the clean HC dataset and all HC scans against this subject's LOOCV model. Black crosses show thresholds for which the points on the ROC (sensitivity and 1-specificity) were closest to the top-left corner.

3.3. Within-subject reproducibility/homogeneity

Supplementary Fig. S3 shows an example of CTh deviations in the patient with CDR ≥ 2, who had the largest number of scans (OAS30902,

four rescans during the same session). The figure consistently shows atrophy patterns in the right parietal and temporal lobe, as well as the characteristic reduction in CTh in the entorhinal cortex, extended to the lingual gyrus. Interestingly, both FreeSurfer and DL + DiReCT indicate increased CTh in several regions in the first two rescans. Visual inspection of these scans showed reduced image contrast, presumably due to patient motion, explaining the need to acquire two additional scans, which had better image quality.

Reproducibility of CTh patterns across the whole OASIS3 dataset was assessed. Subject-wise distances of the CTh significance maps between rescans of the same participant were estimated by calculating the normalized L2-distance between signed log10(p) maps of mean CTh estimates (brain size normalized). HC maps were taken from the LOOCV analysis, whereas AD maps were taken from their evaluation against the clean normative dataset. When using DL + DiReCT, the distance between significance maps of repeated scans was lower in patients with AD ($L2 = 0.39 \pm 0.26$, median ± standard deviation) than in HCs ($L2 = 0.42 \pm 0.19$, $p = 0.05$ in a Wilcoxon rank sum test to account for the large skewness of both distributions). For FreeSurfer, there was no significant difference between AD and HC (AD: 0.56 ± 0.26; HC: 0.56 ± 0.20; $p = 0.66$). Repeated significance maps were significantly closer for DL + DiReCT than for FreeSurfer ($p = 9.2e\text{-}57$ in Wilcoxon signed rank test on AD maps, and $p = 2e\text{-}309$ on HC maps).

3.4. Inspection of scans with extreme atrophy

Fig. 7 shows the correspondence between FreeSurfer and DL + DiReCT regarding the average log(p) value for the mean thickness of the right and left hemispheres (Pearson correlation coefficient = 0.89, $p < 1e\text{-}12$). Cortical log(p) maps for the 3 scans with most atrophy for DL + DiReCT or FreeSurfer are shown in Fig. 8. Except for scans sub-OAS30373_ses-d1211 and sub-OAS31084_ses-d2319, log(p) maps of both tools had a correlation higher than 0.8.

Most differences in log(p) values were associated with differences in region labeling. FreeSurfer and DL + DiReCT parcellations for the selected scans are shown in Supplementary Fig. S8, where potential reasons for large log(p) in these scans are discussed. Both DL + DiReCT and FreeSurfer appeared to commit errors in challenging scans, which might have led to spuriously low log(p) of the mean CTh of the affected tool.

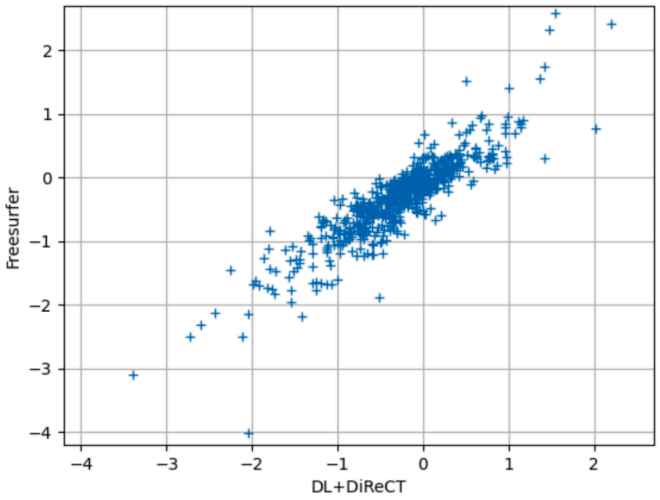

Fig. 7. Comparison of average log(p) values for the left and right hemisphere's mean thickness, obtained from DL + DiReCT (x-axis) and FreeSurfer (y-axis). Both methods showed strong agreement (pearson correlation coefficient = 0.89, $p < 1e\text{-}12$), except for a few subjects with extreme values. The 3 scans with most atrophy for both tools were visually inspected, and the corresponding maps and parcellations reported in Fig. 8 and Supplementary Fig. S8.

3.5. Comparison to other normative models

When evaluating cortical thickness of AD scans and using the OASIS3 reference dataset, ScanOMetrics provided z-scores and log(p) values that corresponded to larger models like warped Bayesian linear regression models (Rutherford et al., 2022). Fig. 9 shows the comparison of z-scores obtained using ScanOMetrics and PCN toolkit, which were found to follow a strong and almost linear correspondence (Pearson correlation coefficient = 0.98 and 0.99 for the right and left hemispheres respectively, $p < 1e\text{-}12$). Supplementary Fig. S7 compares log(p) values obtained with ScanOMetrics to centiles obtained using BrainChart. Here, a strong, non-linear correspondence was found (ranked Spearman coefficient 0.95, $p < 1e\text{-}12$).

3.6. Processing times

On our hardware the processing time for one MRI scan was 9h20m ± 2h50m (mean ± standard deviation) with FreeSurfer (running on CPU only), and 23m59s ± 4m30s with DL + DiReCT. This value was split into 1m55s ± 13 s for segmentation on the GPU and 22m04s ± 4m25s for CTh estimation with DiReCT (Das et al., 2009; Avants et al., 2014) on the CPU. Fitting the clean normative models on all subjects took 8m17s for FreeSurfer and 7m04s for DL + DiReCT. Time required for evaluation of a single scan against a normative model was 1.91 ± 0.57 s for FreeSurfer and 1.93 ± 0.58 s for DL + DiReCT.

3.7. Cleaning the normative models

Among the 1′927 HC scans, that were initially used for normative modeling, 99 scans were flagged as anomalous in the LOOCV analysis (i.e. more than 18 metrics with $P < q = 0.01$), using either DL + DiReCT or FreeSurfer. Fig. 10 shows the cumulative distributions (left) and probability densities (right) of the fraction of abnormal metrics per scan, and details the number of rejections for both pipelines. Regions that contributed to the large number of anomalies in these 99 scans were widely distributed over the entire cortex, see upper part of Supplementary Fig. S4. Furthermore, visual inspection of the scans with the largest number of deviant metrics (see Supplementary Fig. S4) revealed the following potential causes for outlier detection:

- low image quality, mainly due to susceptibility artifacts in the mouth region (rows 1 and 2); likely caused by dental implants (Chockattu et al., 2018)
- prominent lateral ventricles and/or enlarged CSF space suggestive of atrophy (e.g. subject OAS30662, row 3; about 20 % of DL + DiReCT's metrics were flagged as abnormal with respect to the expected values at the age of the subject (87 years), despite a reported CDR score of zero)
- blurring and ringing artifacts due to patient motion during the scan (rows 4 to 6)

The 99 HC scans with large number of anomalies (5.15 %) were removed before our final normative models were built from the of 1′828 remaining HC scans. These "clean HC datasets" were subjected to a final LOOCV and used for all subsequent analyses of scans from patients with AD. In these models the number of HC scans flagged as anomalous by either DL + DiReCT or FreeSurfer decreased to 27 (1.48 % of HC scans, $p_{binom} = 0.045$ for fraction 1 %). Especially for DL + DiReCT the anomalies in these scans were much more regionally specific than before (lower part of Supplementary Fig. S4). Shapiro-Wilk tests indicated that almost all metrics had non-normally distributed fit residues, with only minor improvements due to the cleaning procedures (decrease from 91 % to 85 % for DL + DiReCT and from 90 % to 84 % for FreeSurfer regarding the 358 raw metrics, and from 93 % to 90 % for DL + DiReCT and 98 % without change for FreeSurfer regarding the 240 brain size normalized metrics).

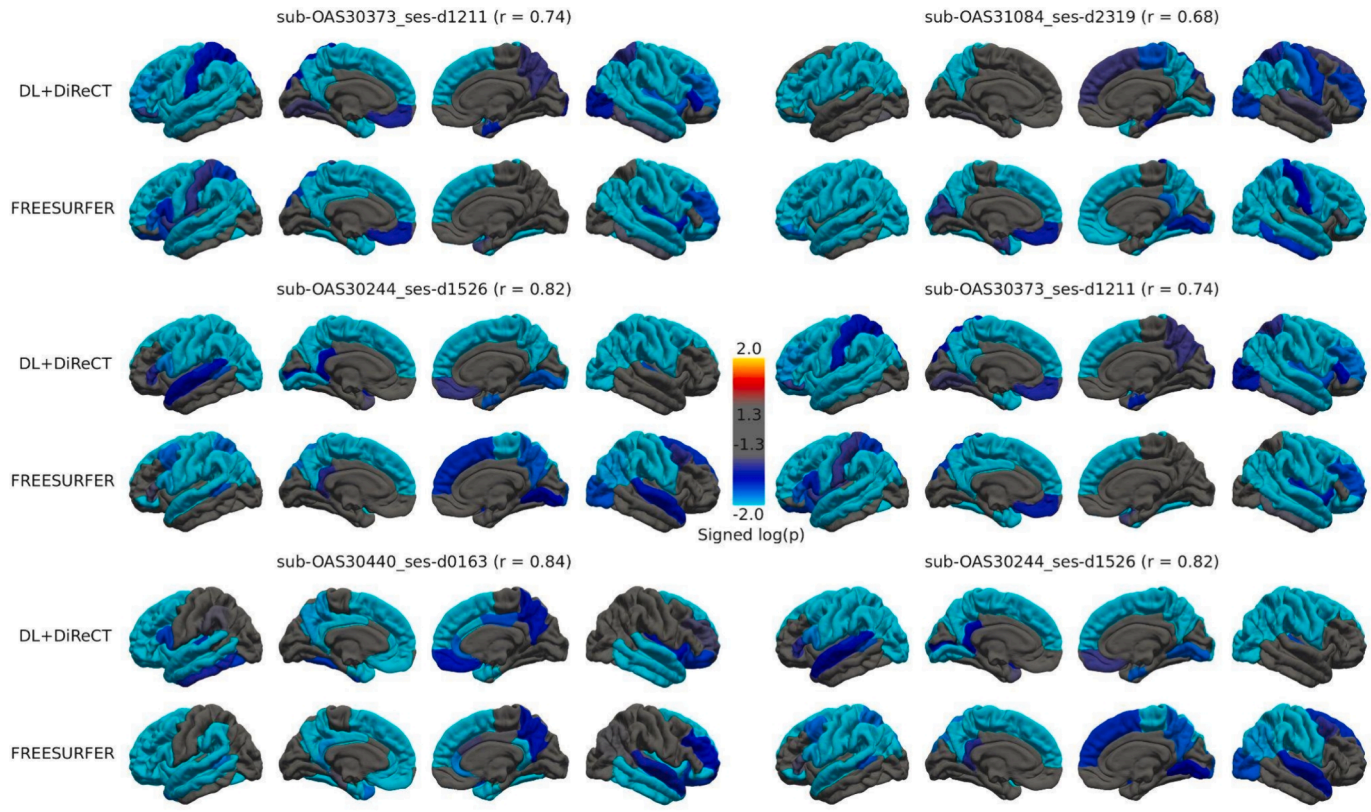

Fig. 8. Comparison of cortical log(p) maps for scans with most atrophy according to DL + DiReCT (left column) and FreeSurfer (right column). For each scan, the DL + DiReCT (upper row) and FreeSurfer (lower row) log(p) maps are shown.

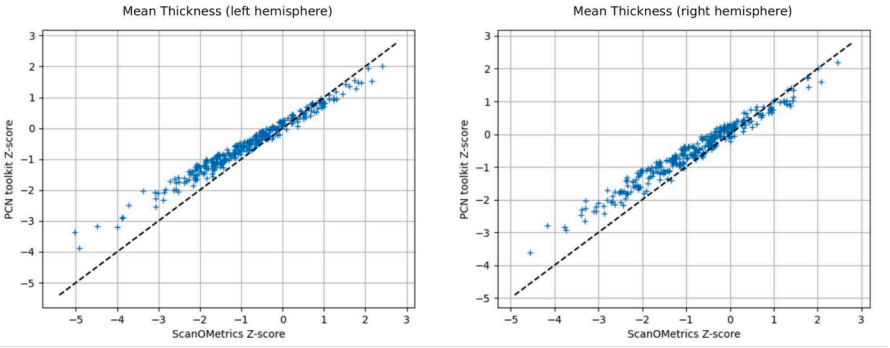

Fig. 9. Comparison of z-scores obtained for the left and right Mean Thickness, using ScanOMetrics (x-axis) and PCN toolkit (y-axis). The two normative tools provide z-scores that strongly correlate.

3.8. Anomaly detection in healthy controls (LOOCV before cleaning)

We assessed specificity of our approach to regional anomaly detection by running a subject-specific leave-one-out cross-validation (LOOCV) on the whole HC dataset. When considering the full set of tests made (358 raw metrics times 1′927 scans, yielding 689′866p-values), and using a significance threshold of q = 0.01, processing the HC dataset with DL + DiReCT resulted in 8′911 significant p-values (1.30 %, which is slightly but significantly higher than the expected 1 %, $p_{binom} < 1e\text{-}16$). When considering an alternative q = 0.05, only 4.44 % of p-values were significant, which was lower than the expected 5 %. Processing data with FreeSurfer, using q = 0.01 or q = 0.05 resulted in 1.16 % and 4.13 % of significant p-values, respectively, which more or less resembled the numbers reported by Rummel et al. (2018) for a completely different dataset. Similar to the suspicion raised there, the reason for this observation might be due to the residues of the polynomial age fits not being normally distributed (Shapiro-Wilk tests) for the vast majority of metrics in our LOOCV, regardless of whether using DL + DiReCT (305 metrics out of 358 were not normally distributed) or FreeSurfer (299 of 358 non-normally distributed sets of residuals), and independently of using uniform subsampling or not.

3.9. Features of the cleaned normative models

After cleaning, using a significance threshold of q = 0.01 lead to 1.00 % of anomalous raw metrics when using FreeSurfer, and 1.06 % when using DL + DiReCT. A threshold of 0.001 led to 0.23 % and 0.25 % anomalous metrics, respectively. And a threshold of 0.05 led to 3.9 % and 4.2 % of anomalous metrics. A threshold of q = 0.01 was fixed throughout our experiments. In terms of anomalous scans, DL + DiReCT yielded 18 anomalous scans with more than 5 % of regions detected as anomalies (0.98 % of the 1′828 scans), while FreeSurfer detected 14

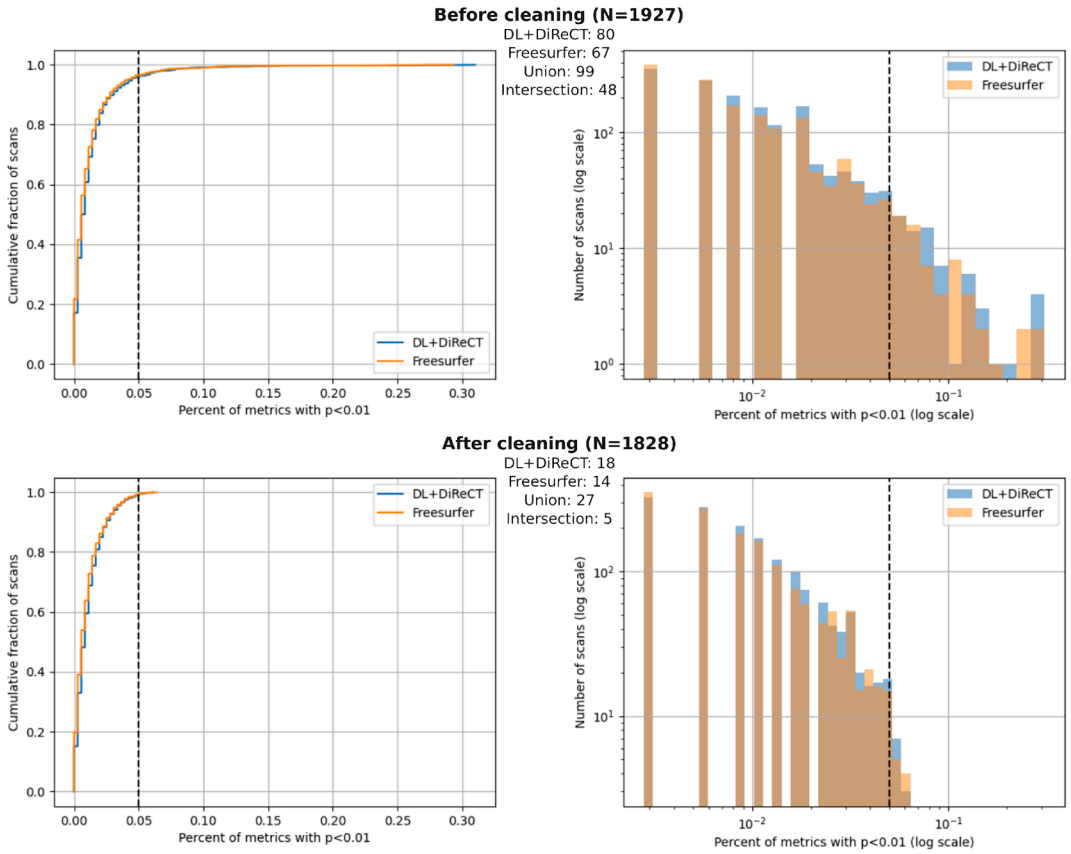

Fig. 10. Distribution of the abnormal fraction of metrics per scan, before (top) and after (bottom) cleaning the normative dataset. Numbers in bold correspond to the total number of scans before and after cleaning. Numbers in smaller font correspond to anomalous scans in the dataset, as detected when using either DL + DiReCT or FreeSurfer. Cleaning consisted in removing the 99 scans detected as anomalous by either one of the two software tools.

scans (0.77 % of scans). Union of both sets yielded 27 scans (1.48 %), and the intersection 5 (0.27 %). Values are reported in the lower part of Fig. 7, and correspond to scans on the right of the dotted lines in the cumulative distribution and histogram.

Since mean age was different between our participants with CDR = 0 and CDR $\geq$ 1 (see Table 1, p = 3.3e-13 in a *t*-test), fitting age models and working with residues rather than with the original metrics was appropriate. Normative models for metrics estimated with DL + DiReCT had degree d = 0 (constant) in 55 % of the fits, d = 1 (linear) in 42 % of the fits, and d = 2 (quadratic) in 3 % of the fits. No higher degree was selected. When using FreeSurfer, the degree was d = 0 in 25.7 % of the fits, d = 1 in 70.3 % of the fits, d = 2 in 3.9 % of the fits, and d = 3 (cubic) in 0.1 % of the fits.

We performed 240 one-tailed F-tests (once for either direction) for different residual variance when fitting normative models to brain size normalized metrics calculated with DL + DiReCT or FreeSurfer. Testing smaller variance for DL + DiReCT than for FreeSurfer and performing FDR correction to account for the many comparisons, 56 of the 240 metrics were significant on level q < 0.01. Most of these were either mean normalized CTh (16 of 34 on the left hemisphere and 15 on the right) or its standard deviation (8 on the left and 11 on the right). In contrast, only seven residual variances of normalized GMV were smaller for DL + DiReCT than for FreeSurfer, among which five were subcortical regions. Testing in the other direction, 97 metrics showed smaller residual variance for FreeSurfer than for DL + DiReCT. Among these, 78 were cortical and subcortical volumes, whereas only 2 (4) were mean and 6 (7) were standard deviations of normalized CTh on the left (right) hemisphere.

4. Summary and discussion

In this paper we have compared identical metrics derived from two brain morphometry software tools, i.e. DL + DiReCT (Rebsamen et al., 2020, 2023) and FreeSurfer (Dale et al., 1999; Fischl et al., 1999a, 1999b; Fischl & Dale, 2000), regarding their use in the context of ScanOMetrics, an open-source pipeline for normative modeling and detection of statistical anomalies (Rummel et al., 2017, 2018). ScanOMetrics processing is supposed to detect abnormal regions in individual MRI scans, which may support neuroradiological assessment of the cases with respect to many clinical questions. An implementation of ScanOMetrics in Python3 has been made publicly available to the community as open source software. Together with the public availability of the used OASIS3 dataset, ScanOMetrics tutorials available online and the normative models used in the present work (specific for OASIS3), this makes our results completely reproducible.

Our main findings are that regardless of the software used for extraction of the metrics, in patients with Alzheimer's disease (AD) anomaly detections were up to twenty-fold more frequent than in healthy controls (HC). Cortical thickness (CTh) anomalies were mainly detected in regions that are known as predilection areas of cortical atrophy in AD. Atrophy patterns extend to larger regions when stratifying by clinical dementia rating (CDR) was clearly observable with both methods. DL + DiReCT provided CTh results more than 15 times faster than FreeSurfer.

4.1. Origin of statistical brain anomalies

Detected statistical anomalies may have at least three origins, which influence their differential statistical properties. First, regional metrics

of brain shape can artifactually be detected as abnormal. These detections depend on the measurement uncertainty of the metric and the image quality of the scans. They have spatial predilection regions, which can be identified by studying scan-rescan variabilities (see e.g. Rummel et al., 2018; Rebsamen et al., 2020). In ScanOMetrics the artifact probability and odds for valid vs. artifactual detections are reported (see text elements in Fig. 5) to guide the user's judgment of reliability. Second, different individuals have different brain shapes, which yields highly reproducible deviations from the expectation, see Section "Within-subject reproducibility/homogeneity" and Rummel et al. (2017, 2018). Where these deviations are strong enough, they can trigger subject specific anomaly detections. Finally, brain pathologies yield anomaly detections as well, which are often concentrated in brain regions that have been revealed as disease specific alterations of brain shape in large morphometric group studies in the past.

Importantly, when investigating an individual MRI scan (as is often the case for diagnostic purposes), all three sources contribute to detections of statistical anomalies, but only the last category is relevant to answer clinical questions. In consequence, the pattern of detections (e.g. spatial extent of alterations of brain structure, like e.g. atrophy) in an individual almost never matches disease specific patterns as described in the literature or derived by group assessment exactly. While often centered in these predilection areas, detections depend on image quality and usually reach beyond these regions, see Fig. 3c for an illustration of the variability in patients with AD of the OASIS3 dataset. However, when pooling the detections made in many individuals over groups representing the same clinical condition like in our Figs. 1 and 2, the first two causes for anomalies have a chance to level out and the expected group patterns usually become visible more clearly. Similarly, when pooling several scans of the same subject, primarily the subject-individual anomalies would become more clearly visible, whereas the same is true for disease-specific patterns only if they remain stable over the observation period.

4.2. Atrophy patterns in individual patients with Alzheimer's disease

Our hypothesis (4) was that ScanOMetrics yields a much higher rate of detected anomalies in patients with AD than in HCs when using a leave-one-out cross-validation (LOOCV). Indeed, independently from the used software, about 60 % of scans were rated as abnormal in the AD group, compared with only ~ 1 % in the cleaned HC dataset. Comparison of rows 3 and 4 in Fig. 1 shows the associated rates of atrophy detection in individuals. Particularly in the bilateral entorhinal gyrus (see e.g. Gómez-Isla et al., 1996, or Mueller et al., 2010), but also in parieto-temporal brain regions, the rate of detected significant reductions in CTh was elevated in patients with AD up to twenty-fold. Regarding detected regional increase in CTh, the difference between AD and HC scans was much less pronounced (rows 1 and 2).

The spatial pattern of detected CTh reductions was consistent with the temporo-parietal predilection areas of atrophy in patients with AD obtained from a group comparison in the same data (see Fig. 1, rows 3 and 5) or existing literature like (Whitwell et al., 2011, Harper et al., 2017, Ferreira et al., 2017), confirming the second part of our hypothesis (5). A similar correspondence between pooled detections in individuals and results of a group study was recently found by Verdi et al. (2023).

Fig. 2 reveals that the rate of detected brain atrophy (rows 4 to 6) increased with clinical dementia rating (CDR). Using DL + DiReCT, the mean CTh of the bilateral entorhinal gyrus was detected as significantly reduced already in one fourth of patients with CDR = 0.5, a value that almost reached half of the scans in cases with CDR $\geq$ 1. Using FreeSurfer the progression with CDR was observable as well, but detection rates were lower (i.e. only about one sixth for CDR = 0.5 and in one third in CDR $\geq$ 1). A similar association between atrophy detection in scans of patients with AD and their total scores from the Mini-Mental State Examination (MMSE) has been observed recently by Verdi et al. (2023). Furthermore, our Fig. 2 showcases what would appear to be a cross-sectional snapshot of the posterior-to-anterior atrophy progression reported by Contador et al. (2021).

Regional CTh increase for higher CDR (rows 1 to 3 of Fig. 2) was observed only for the medial orbito-frontal gyrus (predominantly on the right hemisphere, more pronounced for FreeSurfer than for DL + DiReCT). Interestingly, this is exactly the region in which increased CTh was detected in individual scans with reduced image quality due to patient motion (see rows 1 and 2 of Supplementary Fig. S3 for an example). Regions with increased thickness in these motion-contaminated scans were not highlighted anymore in the two following scans (acquired during the same session) which had a better image quality. In addition, Supplementary Fig. S3 reveals that the medial orbito-frontal gyrus was one of the more frequently made false positive detections in HCs after cleaning of the normative dataset. Given these observations, and since we do not have any plausible biological interpretation underlying the observed CTh increase, we hypothesize that increased thickness might be related to motion artifacts (e.g. ringing). However, additional and thorough tests should be conducted to verify this hypothesis..

Compared to the large fraction of entire scans rated as anomalous in patients with AD (~60 % for both software tools), the peak effect of CTh reduction (detectable in the entorhinal gyrus in "only" about half of scans of the general AD group, see Fig. 1, row 3) was relatively small, indicating that the anomaly patterns detectable in the individual patient with AD are largely non-overlapping. This observation confirms the first part of our hypothesis (5) and is consistent with recent observations by Verdi et al. (2023), who have also reported widespread detection patterns with only moderate peak proportions of detections in the basal temporal lobes.

4.3. Value for clinical decision support

Normative modeling of healthy brain shape, its development and aging have great potential to support clinical routine assessment of suspected pathologies in neuroradiological MRI exams. It is important to stress that we envision the automated detection of statistical anomalies in individuals (like shown for example in Fig. 3 or S3) as a trigger for secondary inspection by the human expert, rather than as an automated disease classification tool. Used as a screening tool for further regional image analysis (Rummel et al., 2017), normative modeling could indeed provide valuable decision support to the neuroradiologist.

In contrast, classifying entire scans as anomalous based on a threshold on the accepted rate of abnormal metrics is not sufficiently reliable. The same is true for classifying scans into AD or HC based on the accepted degree of anomaly in selected regional SBA metrics, which are known as frequently compromised in dementia (like e.g. the hippocampal GMV or the entorhinal CTh). In our study, both approaches lead to areas under the ROC curve below 0.8, without greater difference between the evaluated software tools.

CE-marked and FDA-approved commercial tools for clinical decision support by brain morphometry have meanwhile become available for application in patients with multiple sclerosis and various forms of dementia. Despite formal approval for diagnostic purposes, a deficiency of these tools is that validation, especially in clinical terms, in many cases still is an open topic of research (Pemberton et al., 2021; Mendelsohn et al., 2023) due to a multitude of factors (Haller et al., 2022; Leming et al., 2023; Hedderich et al., 2023). This is remarkable, since an international survey among practitioners investigating their application of (commercial or scientific) brain morphometry tools has clearly shown that user acceptance is associated with the availability of technical and clinical validation studies (Vernooij et al., 2019).

4.4. DL + DiReCT vs. FreeSurfer

Our comparison between using DL + DiReCT and FreeSurfer for

metrics estimation was motivated by the question, whether one of the two methods yielded more stable or more plausible spatial patterns of statistical anomaly detections than the other, see our hypotheses (1), (2) and (3). Our findings show that this question cannot be answered so clearly. In general, the group aggregations in Figs. 1, 2 and 5 reveal very similar patterns for both software tools. The lowest row in Fig. 3c and Supplementary Fig. S3 show that the same can be true for the degree of regional CTh anomaly detected in rescans of an individual patient.

Our results showed that at the level of hemispheres, log(p) values obtained from FreeSurfer and DL + DiReCT correlated strongly. ScanOMetrics also provided evaluation metrics that highly agreed with other more complex normative tools (e.g. BrainChart and PCN toolkit). Since for the given age range of the OASIS3 dataset most polynomials fitted by ScanOMetrics had a low degree, the strong correspondence with other normative models might be due to a smooth and gradual thickness decrease. Datasets with a larger and less linear age dependence would be required to investigate the relation between different approaches to normative modelling under more general circumstances.

Our hypothesis (1) that fit residues are in general more narrowly distributed for DL + DiReCT than for FreeSurfer could not be confirmed by our study. Rather, this was true only for one fourth of the normalized metrics (56 of 240), whereas 97 showed the opposite behavior. Remarkably, we observed that metrics with smaller residual variance for DL + DiReCT were predominantly thickness measures, whereas volume metrics dominated the group where residual variance was smaller for FreeSurfer. More narrow distribution of CTh fit residues when using DL + DiReCT is in line with recent observations by Rusak et al. (2022), who have found that the DL-based tool is more sensitive and more reproducible at weak synthetic reduction of CTh than FreeSurfer's cross-sectional or longitudinal pipelines. For GMV the situation is different: DL + DiReCT counts voxels and thus is prone to uncertainties introduced by its voxel-wise hard classification into one of several brain regions or tissue classes. By contrast, FreeSurfer's GMV estimates are based on the volumes enclosed inside its much smoother surface meshes, which likely explains the more narrow distribution of volume fit residues.

Similarity of deviations from the normative models between rescans of the same participant (assessed by normalized L2-distance of thickness-based signed log10(p) maps) was large in general. This is in line with the observation that ScanOMetrics' deviations from the expectation are subject specific (Rummel et al., 2017, 2018) and remarkably reproducible, also see Supplementary Fig. S2 for an example. Confirming our hypothesis (2), L2-distances of CTh significance maps between rescans were significantly smaller for DL + DiReCT than for FreeSurfer and a similar effect was observed for cortical GMV (not shown). This is consistent with the interpretation that rescan errors and disturbance by artifacts depending on image quality are smaller for DL + DiReCT than for FreeSurfer. The L2-distances were in addition smaller in patients with AD than in HCs when using DL + DiReCT. We interpret this finding as a sign that subject-specific signed log10(p) values derived using DL + DiReCT are small and spatially unspecific for HC subjects, whereas those of patients with AD have additional disease related and spatially specific deviations from the normative model that are larger in size and thus determine the L2-distance.

In Fig. 4 we have detected an annual hippocampal atrophy rate of almost 6 % in an individual patient using DL + DiReCT, which is in agreement with group estimates found in the literature (Sluimer et al., 2008). Using FreeSurfer the annual atrophy rate was only half as large (see Supplementary Fig. S2). This might indicate a higher sensitivity of DL + DiReCT to atrophy progression in the individual, supporting our hypothesis (3). Since sensitivity to atrophy and reproducibility of patterns has mainly been compared for CTh and not for GMV so far (Rebsamen et al., 2020, 2023; Rusak et al., 2022), this hypothesis requires additional investigation in subsequent work.

Processing with DL + DiReCT (<25 min) yielded comparable results for mean CTh more than 15 times faster than the full FreeSurfer pipeline. However, DL + DiReCT's output is drastically reduced, currently focusing on some of the most frequently used SBA metrics of brain morphometry, namely mean and standard deviation of regional CTh, GMV and volumes of some subcortical segmentations. Using the OASIS3 dataset with 1′828 HCs after cleaning, fitting a normative model to the set of regional brain metrics used in this paper took less than 10 min. Importantly, this procedure has to be performed only once for each normative dataset. Despite the expectation of statistical post-processing with ScanOMetrics to only depend on the number of metrics and scans and not on the software used for metrics estimation, we observed a minimally smaller processing time for DL + DiReCT than for FreeSurfer. We explain this minor discrepancy by a different number of outliers rejected during the fitting procedures. Application of the normative models to a new case required less than two seconds computation time for both tools, almost three orders of magnitude quicker than the calculation of the metrics with DL + DiReCT and practically not contributing to the entire computation time.

Inspecting individual scans, in particular those with highest atrophy, did not reveal a systematic bias in any of the two methods. However, such inspection of outlier scans sporadically highlighted segmentation and parcellation errors when the two methods disagreed on the corresponding log(p) value. In Supplementary Fig. S8, FreeSurfer showed what one could call "indexing" or "offset" errors and which might be related to the reliance of FreeSurfer on an atlas to identify different gyri. In the figure, scan sub-OAS30373_ses-d1211 has several mislabeled gyri in the right temporal lobe. Mislabeling part of the inferior-temporal gyrus as fusiform could have driven FreeSurfer into labeling the fusiform gyrus as entorhinal gyrus, and the entorhinal gyrus as temporal pole. This effect can lead to the evaluation of CTh against the wrong reference. DL + DiReCT on the other hand seemed more prone to segmentation errors, failing at capturing the whole extent of a region, likely due to challenging WM/GM contrast.

Since DL + DiReCT provided similar results to the widely used and validated FreeSurfer software (both in terms of evaluation and classification), but was about 15 times faster, we tend to recommend using DL + DiReCT for clinically oriented morphometry evaluation and normative modelling, when time and computing power are limiting factors. If the aim is to compare features to normative models developed based on FreeSurfer, we recommend continuing with this software, for the sake of reproducibility.

4.5. Outlook

Future work should combine residues from normative modeling with proportional hazard models for AD conversion (e.g. Devanand et al., 2012), AD classifiers (e.g. logistic regression as used in Bobinski et al., 1999, linear discriminant analysis as in Juottonen et al., 1999, support vector classifiers as in Schmitter et al., 2014; Gupta et al., 2019, random forests or K-nearest neighbor classifiers as in Gupta et al., 2019, probabilistic multi-kernel classifiers as in Popuri et al., 2020) or disease progression models (Fonteijn et al., 2012; Sivera et al., 2019; Planche et al., 2022; Saint-Jalmes et al., 2023) to thoroughly investigate if improved diagnostic accuracy can be obtained at the subject level. Special care should be devoted to avoid data leakage (Kapoor and Narayanan, 2023), and into addressing the heterogeneity/similarity of atrophy patterns across dementia subtypes.

Atrophy patterns have been shown to differ between early and late onset dementia (Harper et al., 2017), to be similar between AD subjects with and without amnestic clinical syndromes (Whitwell et al., 2011), or even to be undetectable with current methodology in some AD patients (Ferreira et al., 2017). Grouping subtypes with different atrophy patterns might impair the accuracy of clinical decision support models, while splitting datasets in too many groups will reduce statistical power. Further work should explore if individual normative metrics could be of interest for certain dementia subtypes, or if multivariate and disease progression models are required in order to properly classify subtypes of dementia.

Recently, first clinical evaluation studies have become available for non-commercial, research-level and open-source tools for brain morphometry. In small case-control studies focusing on hippocampal sclerosis in temporal lobe epilepsy, Goodkin et al. (2021) and Rebsamen et al., (2023c) have compared expert ratings without and with availability of quantitative reports (QReports). Both found that with QReports available the accuracy and rater confidence for presence of hippocampal sclerosis increased, whereas disagreement among experts reduced. An obvious next step of our research will be to conduct similar studies with our open-source tool ScanOMetrics. Depending on the clinical question and suspected disease, different quantitative findings are in the center of the user's interest. To ease ScanOMetrics usage, we will develop a graphical user interface (GUI) and design disease specific QReports.

Code availability

All software tools used in this paper are open source. The Python3 implementation of ScanOMetrics is available at https://github.com/SCAN-NRAD/ScanOMetrics. A code description is given in the Supplementary Materials and a more detailed documentation with tutorial is available at https://scanometrics.readthedocs.io. FreeSurfer can be downloaded from https://surfer.nmr.mgh.harvard.edu/ and DL + DiReCT is available at https://github.com/SCAN-NRAD/DL-DiReCT.

CRediT authorship contribution statement

David Romascano: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Michael Rebsamen:** Writing – review & editing, Software, Methodology, Data curation. **Piotr Radojewski:** Writing – review & editing, Methodology. **Timo Blattner:** Writing – review & editing. **Richard McKinley:** Writing – review & editing, Supervision. **Roland Wiest:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Christian Rummel:** Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Data availability

The openly available OASIS3 dataset was downloaded from https://www.oasis-brains.org/#data.

Acknowledgements

This work was supported by Swiss National Science Foundation (SNSF) under Grant/Award Numbers: 204593 (ScanOMetrics) and CRSII5_180365 (The Swiss-First Study).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2024.103624.

References

Alkan, E., Davies, G., Evans, S.L., 2021. Cognitive impairment in schizophrenia: relationships with cortical thickness in fronto-temporal regions, and dissociability from symptom severity. NPJ Schizophr. 7 (1), 20. https://doi.org/10.1038/s41537-021-00149-0.

Bethlehem, R.A., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., et al., 2022. Brain charts for the human lifespan. Nature 604 (7906), 525–533. https://doi.org/10.1038/s41586-022-04554-y.

Bobinski, M., de Leon, M.J., Convit, A., De Santi, S., Wegiel, J., Tarshish, C.Y., Saint Louis, L.A., Wisniewski, H.M., 1999. MRI of entorhinal cortex in mild Alzheimer's disease. Lancet 353 (9146), 38–40. https://doi.org/10.1016/s0140-6736(05)74869-8.

Chockattu, S.J., Suryakant, D.B., Thakur, S., 2018. Unwanted effects due to interactions between dental materials and magnetic resonance imaging: a review of the literature. Restorat. Dentis. Endodont. 43 (4), e39.

Contador, J., Pérez-Millán, A., Tort-Merino, A., Balasa, M., Falgàs, N., Olives, J., Castellví, M., Borrego-Écija, S., Bosch, B., Fernández-Villullas, G., Ramos-Campoy, O., Antonell, A., Bargalló, N., Sanchez-Valle, R., Sala-Llonch, R., Lladó, A., Initiative, A.D.N., 2021. Longitudinal brain atrophy and CSF biomarkers in early-onset Alzheimer's disease. NeuroImage. Clinical 32, 102804. https://doi.org/10.1016/j.nicl.2021.102804.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis I. Segmentation and Surface Reconstruction. Neuroimage 9 (2), 179–194. https://doi.org/10.1006/nimg.1998.0395.

de Figueiredo, N.S.V., Gaça, L.B., Assunção-Leme, I.B., Mazetto, L., Garcia, M.T.F.C., Sandim, G.B., Alonso, N.B., Centeno, R.S., Filho, G.M.A., Jackowski, A.P., Júnior, H. C., Yacubian, E.M.T., 2021. A pioneering FreeSurfer volumetric study of a series of patients with mesial temporal lobe epilepsy and hippocampal sclerosis with comorbid depression. Psychiatry Res. Neuroimaging 311, 111281. https://doi.org/10.1016/j.pscychresns.2021.111281.

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., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Devanand, D.P., Bansal, R., Liu, J., Hao, X., Pradhaban, G., Peterson, B.S., 2012. MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease. Neuroimage 60 (3), 1622–1629. https://doi.org/10.1016/j.neuroimage.2012.01.075.

Du, A.T., Schuff, N., Amend, D., Laakso, M.P., Hsu, Y.Y., Jagust, W.J., Yaffe, K., Kramer, J.H., Reed, B., Norman, D., Chui, H.C., Weiner, M.W., 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 71 (4), 441–447. https://doi.org/10.1136/jnnp.71.4.441.

Fennema-Notestine, C., Hagler Jr, D.J., McEvoy, L.K., Fleisher, A.S., Wu, E.H., Karow, D. S., Dale, A.M., Initiative, A.D.N., 2009. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. Hum. Brain Mapp. 30 (10), 3238–3253. https://doi.org/10.1002/hbm.20744.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J.A., Cavallin, L., Guo, C.J., Ekman, U., Muehlboeck, J.S., Simmons, A., Barroso, J., Wahlund, L.O., Westman, E., 2017. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. Sci. Rep. 7, 46263. https://doi.org/10.1038/srep46263.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. PNAS 97 (20), 11050–11055. https://doi.org/10.1073/pnas.200033797.

Fischl, B., Sereno, M.I., Dale, A.M., 1999a. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207. https://doi.org/10.1006/nimg.1998.0396.

Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999b. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8 (4), 272–284. https://doi.org/10.1002/(sici)1097-0193(1999)8:4<272::aid-hbm10>3.0.co;2-4.

Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., Alexander, D.C., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. Neuroimage 60 (3), 1880–1889. https://doi.org/10.1016/j.neuroimage.2012.01.062.

Fortea, A., van Eijndhoven, P., Calvet-Mirabent, A., Ilzarbe, D., Batalla, A., de la Serna, E., Puig, O., Castro-Fornieles, J., Dolz, M., Tor, J., Parrilla, S., Via, E., Stephan-Otto, C., Baeza, I., Sugranyes, G., 2023. Age-related change in cortical thickness in adolescents at clinical high risk for psychosis: a longitudinal study. Eur. Child Adolesc. Psychiatry. https://doi.org/10.1007/s00787-023-02278-6.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024.

Frisoni, G.B., Fox, N.C., Jack Jr, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6 (2), 67–77. https://doi.org/10.1038/nrneurol.2009.215.

Ge, R., Yu, Y., Qi, Y. X., Fan, Y. V., Chen, S., Gao, C., Haas, S. S., Modabbernia, A., New, F., Agartz, I., Asherson, P., Ayesa-Arriola, R., Banaj, N., Banaschewski, T., Baumeister, S., Bertolino, A., Boomsma, D. I., Borgwardt, S., Bourque, J., Brandeis, D., … Frangou, S. (2023). Normative Modeling of Brain Morphometry Across the Lifespan Using CentileBrain: Algorithm Benchmarking and Model Optimization. *bioRxiv*, 2023.01.30.523509. https://doi.org/10.1101/2023.01.30.523509.

Gómez-Isla, T., Price, J.L., McKeel Jr, D.W., Morris, J.C., Growdon, J.H., Hyman, B.T., 1996. Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. J. Neurosci. 16 (14), 4491–4500. https://doi.org/10.1523/JNEUROSCI.16-14-04491.1996.

Goodkin, O., Pemberton, H.G., Vos, S.B., Prados, F., Das, R.K., Moggridge, J., De Blasi, B., Bartlett, P., Williams, E., Campion, T., Haider, L., Pearce, K., Bargalló, N., Sanchez, E., Bisdas, S., White, M., Ourselin, S., Winston, G.P., Duncan, J.S., Cardoso, J., Barkhof, F., 2021. Clinical evaluation of automated quantitative MRI reports for assessment of hippocampal sclerosis. Eur. Radiol. 31 (1), 34–44. https://doi.org/10.1007/s00330-020-07075-2.

Gupta, Y., Lee, K.H., Choi, K.Y., Lee, J.J., Kim, B.C., Kwon, G.R., National Research Center for Dementia, Alzheimer's Disease Neuroimaging Initiative, 2019. Early diagnosis of Alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of MRI T1 brain images. PLoS One 14 (10), e0222446.

Haller, S., Van Cauter, S., Federau, C., Hedderich, D.M., Edjlali, M., 2022 May. The R-AI-DIOLOGY checklist: a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology. Neuroradiology 64 (5), 851–864. https://doi.org/10.1007/s00234-021-02890-w. Epub 2022 Jan 31 PMID: 35098343.

Harper, L., Bouwman, F., Burton, E.J., Barkhof, F., Scheltens, P., O'Brien, J.T., Fox, N.C., Ridgway, G.R., Schott, J.M., 2017. Patterns of atrophy in pathologically confirmed dementias: a voxelwise analysis. J. Neurol. Neurosurg. Psychiatry 88 (11), 908–916. https://doi.org/10.1136/jnnp-2016-314978.

Hedderich DM, Weisstanner C, Van Cauter S, Federau C, Edjlali M, Radbruch A, Gerke S, Haller S. Artificial intelligence tools in clinical neuroradiology: essential medico-legal aspects. Neuroradiology. 2023 Jul;65(7):1091-1099. doi: https://doi.org/10.1007/s00234-023-03152-7. Epub 2023 May 9. PMID: 37160454; PMCID: PMC10272241.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. Neuroimage 219. https://doi.org/10.1016/j.neuroimage.2020.117012.

Igarashi, K.M., 2023. Entorhinal cortex dysfunction in Alzheimer's disease. Trends Neurosci. 46 (2), 124–136. https://doi.org/10.1016/j.tins.2022.11.006.

Jansen, W.J., Janssen, O., Tijms, B.M., Vos, S.J.B., Ossenkoppele, R., Visser, P.J., Amyloid Biomarker Study Group, Aarsland, D., Alcolea, D., von Altomare, D., Arnim, C., Baiardi, S., Baldeiras, I., Barthel, H., Bateman, R.J., Van Berckel, B., Binette, A.P., Blennow, K., Boada, M., Boecker, H., Zetterberg, H., 2022. Prevalence Estimates of Amyloid Abnormality Across the Alzheimer Disease Clinical Spectrum. JAMA Neurol. 79 (3), 228–243. https://doi.org/10.1001/jamaneurol.2021.5216.

Joy, A., Nagarajan, R., Daar, E.S., Paul, J., Saucedo, A., Yadav, S.K., Guerrero, M., Haroon, E., Macey, P., Thomas, M.A., 2023 Jan. Alterations of gray and white matter volumes and cortical thickness in treated HIV-positive patients. Magn. Reson. Imaging 95, 27–38. https://doi.org/10.1016/j.mri.2022.10.006. Epub 2022 Oct 17 PMID: 36265696.

Juottonen, K., Laakso, M.P., Partanen, K., Soininen, H., 1999. Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. AJNR Am. J. Neuroradiol. 20 (1), 139–144.

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 4 (9), 100804 https://doi.org/10.1016/j.patter.2023.100804.

Laansma, M.A., Bright, J.K., Al-Bachari, S., Anderson, T.J., Ard, T., Assogna, F., Baquero, K.A., Berendse, H.W., Blair, J., Cendes, F., Dalrymple-Alford, J.C., de Bie, R.M.A., Debove, I., Dirkx, M.F., Druzgal, J., Emsley, H.C.A., Garraux, G., Guimarães, R.P., Gutman, B.A., Helmich, R.C., ENIGMA-Parkinson's Study, 2021. International Multicenter Analysis of Brain Structure Across Clinical Stages of Parkinson's Disease. Movement Dis.: Off. J. Movement Disorder Soc. 36 (11), 2583–2594. https://doi.org/10.1002/mds.28706.

Pamela J. LaMontagne, Tammie LS. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, Daniel Marcus. medRxiv 2019.12.13.19014902; doi: https://doi.org/10.1101/2019.12.13.19014902 OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease.

Lemaitre, H., Goldman, A.L., Sambataro, F., Verchinski, B.A., Meyer-Lindenberg, A., Weinberger, D.R., et al., 2012. Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? Neurobiol. Aging 33, 617.e1. https://doi.org/10.1016/j.neurobiolaging.2010.07.013.

Leming, M.J., Bron, E.E., Bruffaerts, R., Ou, Y., Iglesias, J.E., Gollub, R.L., Im, H., 2023. Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. NPJ Digit Med. 6 (1), 129. https://doi.org/10.1038/s41746-023-00868-x.

Lerch, J.P., Pruessner, J.C., Zijdenbos, A., Hampel, H., Teipel, S.J., Evans, A.C., 2005. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb. Cortex 15 (7), 995–1001. https://doi.org/10.1093/cercor/bhh200.

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biol. Psychiatry 80 (7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023.

Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F., 2019. Conceptualizing mental disorders as deviations from normative functioning. Mol. Psychiatry 24, 1415–1424. https://doi.org/10.1038/s41380-019-0441-1.

McCutcheon, R.A., Pillinger, T., Guo, X., Rogdaki, M., Welby, G., Vano, L., Howes, O.D., 2023. Shared and separate patterns in brain morphometry across transdiagnostic dimensions. Nature Mental Health 1 (1), 55–65.

McKinley, R., Rummel, C. (2023). CortexMorph: Fast Cortical Thickness Estimation via Diffeomorphic Registration Using VoxelMorph. MICCAI 2023. Lecture Notes in Computer Science, vol 14229. https://doi.org/10.1007/978-3-031-43999-5_69.

Mendelsohn, Z., Pemberton, H.G., Gray, J., Goodkin, O., Carrasco, F.P., Scheel, M., Nawabi, J., Barkhof, F., 2023. Commercial volumetric MRI reporting tools in multiple sclerosis: a systematic review of the evidence. Neuroradiology 65 (1), 5–24. https://doi.org/10.1007/s00234-022-03074-w. Epub 2022 Nov 4. PMID: 36331588; PMCID: PMC9816195.

Mills, K.L., Siegmund, K.D., Tamnes, C.K., Ferschmann, L., Wierenga, L.M., Bos, M.G., Luna, B., Li, C., Herting, M.M., 2021. Inter-individual variability in structural brain development from late childhood to young adulthood. Neuroimage 242, 118450.

Mueller, S.G., Schuff, N., Yaffe, K., Madison, C., Miller, B., Weiner, M.W., 2010 Sep. Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. Hum. Brain Mapp. 31 (9), 1339–1347. https://doi.org/10.1002/hbm.20934. PMID: 20839293; PMCID: PMC2943433.

National Library of Medicine (U.S.). (2019). Assessment of NfL and GFAP Levels, Atrophy of the Macula GCC by OCT and Whole Brain Atrophy by MRI to Predict Evolution of Neurological Disability in MS Patients. Identifier NCT04860947. https://clinicaltrials.gov/study/NCT04860947.

National Library of Medicine (U.S.). (2024). Early Biomarkers of Neurodegeneration in Parkinsonian Syndromes. Identifier NCT06155942. https://clinicaltrials.gov/study/NCT06155942.

Pemberton HG, Zaki LAM, Goodkin O, Das RK, Steketee RME, Barkhof F, Vernooij MW (2021). Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis-a systematic review. Neuroradiology. 63(11):1773-1789. doi: https://doi.org/10.1007/s00234-021-02746-3. Epub 2021 Sep 3. Erratum in: Neuroradiology. 2021 Sep 24: PMID: 34476511; PMCID: PMC8528755.

Planche, V., Manjon, J.V., Mansencal, B., Lanuza, E., Tourdias, T., Catheline, G., Coupé, P., 2022. Structural progression of Alzheimer's disease over decades: the MRI staging scheme. Brain Commun. 4 (3), fcac109. https://doi.org/10.1093/braincomms/fcac109.

Popuri, K., Ma, D., Wang, L., Beg, M.F., 2020. Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. Hum. Brain Mapp. 41 (14), 4127–4147. https://doi.org/10.1002/hbm.25115.

Potvin, O., Dieumegarde, L., Duchesne, S., 2017. Normative morphometric data for cerebral cortical areas over the lifetime of the adult human brain. Neuroimage 56, 315–339. https://doi.org/10.1016/j.neuroimage.2017.05.019.

Olivier Potvin, Louis Dieumegarde, Simon Duchesne, the Alzheimer's Disease Neuroimaging Initiative, the CIMA-Q, the CCNA groups (2021). NOMIS: Quantifying morphometric deviations from normality over the lifetime of the adult human brain. bioRxiv 2021.01.25.428063; doi: https://doi.org/10.1101/2021.01.25.428063.

Price, J.L., Ko, A.I., Wade, M.J., Tsou, S.K., McKeel, D.W., Morris, J.C., 2001. Neuron number in the entorhinal cortex and CA1 in preclinical Alzheimer disease. Arch. Neurol. 58 (9), 1395–1402. https://doi.org/10.1001/archneur.58.9.1395.

Rebsamen, M., Capiglioni, M., Hoepner, R., Salmen, A., Wiest, R., Radojewski, P., Rummel, C., 2023b. Growing importance of brain morphometry analysis in the clinical routine: The hidden impact of MR sequence parameters. J. Neuroradiol. https://doi.org/10.1016/j.neurad.2023.04.003. Epub ahead of print. PMID: 37116782.

Rebsamen M, Rummel C, Reyes M, Wiest R, McKinley R. Direct cortical thickness estimation using deep learning-based anatomy segmentation and cortex parcellation. Hum Brain Mapp. 2020 Dec;41(17):4804-4814. doi: https://doi.org/10.1002/hbm.25159. Epub 2020 Aug 12. PMID: 32786059; PMCID: PMC7643371.

Rebsamen M, McKinley R, Radojewski P, Pistor M, Friedli C, Hoepner R, Salmen A, Chan A, Reyes M, Wagner F, Wiest R, Rummel C. Reliable brain morphometry from contrast-enhanced T1w-MRI in patients with multiple sclerosis. Hum Brain Mapp. 2023a Feb 15;44(3):970-979. doi: https://doi.org/10.1002/hbm.26117. Epub 2022 Oct 17. PMID: 36250711; PMCID: PMC9875932.

Rebsamen M, Jin BZ, Klail T, De Beukelaer S, Barth R, Rezny-Kasprzak B, Ahmadli U, Vulliemoz S, Seeck M, Schindler K, Wiest R, Radojewski P, Rummel C. Clinical Evaluation of a Quantitative Imaging Biomarker Supporting Radiological Assessment of Hippocampal Sclerosis. Clin Neuroradiol. 2023c Dec;33(4):1045-1053. doi: https://doi.org/10.1007/s00062-023-01308-9. Epub 2023 Jun 26. PMID: 37358608; PMCID: PMC10654177.

Rummel, C., Gast, H., Schindler, K., Müller, M., Amor, F., Hess, C.W., Mathis, J., 2010 Sep. Assessing periodicity of periodic leg movements during sleep. Front. Neurosci. 22 (4), 58. https://doi.org/10.3389/fnins.2010.00058. PMID: 20948585; PMCID: PMC2953451.

Rummel, C., Slavova, N., Seiler, A., Abela, E., Hauf, M., Burren, Y., Weisstanner, C., Vulliemoz, S., Seeck, M., Schindler, K., Wiest, R., 2017 Sep 7. Personalized structural image analysis in patients with temporal lobe epilepsy. Sci. Rep. 7 (1), 10883. https://doi.org/10.1038/s41598-017-10707-1. Erratum. In: Sci Rep. 2018 Jan 9;8 (1):681. PMID: 28883420; PMCID: PMC5589799.

Rummel, C., Aschwanden, F., McKinley, R., Wagner, F., Salmen, A., Chan, A., Wiest, R., 2018 Jan. A Fully Automated Pipeline for Normative Atrophy in Patients with Neurodegenerative Disease. Front. Neurol. 24 (8), 727. https://doi.org/10.3389/fneur.2017.00727. PMID: 29416523; PMCID: PMC5787548.

Rusak, F., Santa Cruz, R., Lebrat, L., Hlinka, O., Fripp, J., Smith, E., Fookes, C., Bradley, A.P., Bourgeat, P., Alzheimer's Disease Neuroimaging Initiative, 2022 Nov. Quantifiable brain atrophy synthesis for benchmarking of cortical thickness estimation methods. Med. Image Anal. 82, 102576 https://doi.org/10.1016/j.media.2022.102576. Epub 2022 Aug 24. PMID: 36126404.

Rutherford, S., Fraza, C., Dinga, R., Kia, S.M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L.K., Bayer, J.M., Dazzan, P., McGuire, P., Mocking, R.T., Schene, A., Sripada, C., Tso, I.F., Duval, E.R., Chang, S.E., Penninx, B. W., Heitzeg, M.M., Burt, S.A., Hyde, L.W., Amaral, D., Wu Nordahl, C., Andreasssen, O.A., Westlye, L.T., Zahn, R., Ruhe, H.G., Beckmann, C., Marquand, A. F., 2022. Charting brain growth and aging at high spatial precision. Elife 11, e72904.

Saint-Jalmes, M., Fedyashov, V., Beck, D., Baldwin, T., Faux, N.G., Bourgeat, P., Fripp, J., Masters, C.L., Goudey, B., Alzheimer's Disease Neuroimaging Initiative, 2023. Disease progression modelling of Alzheimer's disease using probabilistic principal

components analysis. Neuroimage 278, 120279. https://doi.org/10.1016/j.neuroimage.2023.120279.

Schmitter, D., Roche, A., Maréchal, B., Ribes, D., Abdulkadir, A., Bach-Cuadra, M., Daducci, A., Granziera, C., Klöppel, S., Maeder, P., Meuli, R., Krueger, G., Initiative, A.D.N., 2014. An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease. NeuroImage. Clinical 7, 7–17. https://doi.org/10.1016/j.nicl.2014.11.001.

Sinnecker T, Schädelin S, Benkert P, Ruberte E, Amann M, Lieb JM, Naegelin Y, Müller J, Kuhle J, Derfuss T, Kappos L, Wuerfel J, Granziera C, Yaldizli Ö (2022). Brain atrophy measurement over a MRI scanner change in multiple sclerosis. Neuroimage Clin. 36:103148. doi: https://doi.org/10.1016/j.nicl.2022.103148. Epub 2022 Aug 10. PMID: 36007437; PMCID: PMC9424626.

Sivera, R., Delingette, H., Lorenzi, M., Pennec, X., Ayache, N., Initiative, A.D.N., 2019. A model of brain morphological changes related to aging and Alzheimer's disease from cross-sectional assessments. Neuroimage 198, 255–270. https://doi.org/10.1016/j.neuroimage.2019.05.040.

Sluimer, J.D., Vrenken, H., Blankenstein, M.A., Fox, N.C., Scheltens, P., Barkhof, F., van der Flier, W.M., 2008 May 6. Whole-brain atrophy rate in Alzheimer disease: identifying fast progressors. Neurology 70 (19 Pt 2), 1836–1841. https://doi.org/10.1212/01.wnl.0000311446.61861.e3. PMID: 18458218.

Statsenko, Y., Habuza, T., Smetanina, D., Simiyu, G.L., Uzianbaeva, L., Neidl-Van Gorkom, K., Zaki, N., Charykova, I., Al Koteesh, J., Almansoori, T.M., Belghali, M., Ljubisavljevic, M., 2022. Brain Morphometry and Cognitive Performance in Normal Brain Aging: Age- and Sex-Related Structural and Functional Changes. Front. Aging Neurosci. 13, 713680 https://doi.org/10.3389/fnagi.2021.713680.

Thompson, P.M., Hayashi, K.M., De Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., et al., 2003. Dynamics of gray matter loss in Alzheimer's disease. J. Neurosci. 23 (3), 994–1005. https://doi.org/10.1523/JNEUROSCI.23-03-00994.2003.

van Hoesen, G.W., Hyman, B.T., Damasio, A.R., 1991. Entorhinal cortex pathology in Alzheimer's disease. Hippocampus 1 (1), 1–8. https://doi.org/10.1002/hipo.450010102.

Verdi, S., Kia, S.M., Yong, K.X.X., Tosun, D., Schott, J.M., Marquand, A.F., Cole, J.H., 2023. Revealing Individual Neuroanatomical Heterogeneity in Alzheimer Disease Using Neuroanatomical Normative Modeling. Neurology 100 (24), e2442–e2453. https://doi.org/10.1212/WNL.0000000000207298.

Vernooij, M.W., Pizzini, F.B., Schmidt, R., Smits, M., Yousry, T.A., Bargallo, N., Frisoni, G.B., Haller, S., Barkhof, F., 2019. Dementia imaging in clinical practice: a European-wide survey of 193 centres and conclusions by the ESNR working group. Neuroradiology 61 (6), 633–642. https://doi.org/10.1007/s00234-019-02188-y.

Whelan, C.D., Altmann, A., Botía, J.A., Jahanshad, N., Hibar, D.P., Absil, J., Alhusaini, S., Alvim, M.K.M., Auvinen, P., Bartolini, E., Bergo, F.P.G., Bernardes, T., Blackmon, K., Braga, B., Caligiuri, M.E., Calvo, A., Carr, S.J., Chen, J., Chen, S., Cherubini, A., David, P., Domin, M., Foley, S., França, W., Haaker, G., Isaev, D., Keller, S.S., Kotikalapudi, R., Kowalczyk, M.A., Kuzniecky, R., Langner, S., Lenge, M., Leyden, K. M., Liu, M., Loi, R.Q., Martin, P., Mascalchi, M., Morita, M.E., Pariente, J.C., Rodríguez-Cruces, R., Rummel, C., Saavalainen, T., Semmelroch, M.K., Severino, M., Thomas, R.H., Tondelli, M., Tortora, D., Vaudano, A.E., Vivash, L., von Podewils, F., Wagner, J., Weber, B., Yao, Y., Yasuda, C.L., Zhang, G., Bargalló, N., Bender, B., Bernasconi, N., Bernasconi, A., Bernhardt, B.C., Blümcke, I., Carlson, C., Cavalleri, G. L., Cendes, F., Concha, L., Delanty, N., Depondt, C., Devinsky, O., Doherty, C.P., Focke, N.K., Gambardella, A., Guerrini, R., Hamandi, K., Jackson, G.D., Kälviäinen, R., Kochunov, P., Kwan, P., Labate, A., McDonald, C.R., Meletti, S., O'Brien, T.J., Ourselin, S., Richardson, M.P., Striano, P., Thesen, T., Wiest, R., Zhang, J., Vezzani, A., Ryten, M., Thompson, P.M., Sisodiya, S.M., 2018 Feb 1. Structural brain abnormalities in the common epilepsies assessed in a worldwide ENIGMA study. Brain 141 (2), 391–408. https://doi.org/10.1093/brain/awx341. PMID: 29365066; PMCID: PMC5837616.

Whitwell, J.L., Jack Jr, C.R., Przybelski, S.A., Parisi, J.E., Senjem, M.L., Boeve, B.F., Knopman, D.S., Petersen, R.C., Dickson, D.W., Josephs, K.A., 2011. Temporoparietal atrophy: a marker of AD pathology independent of clinical diagnosis. Neurobiol. Aging 32 (9), 1531–1541. https://doi.org/10.1016/j.neurobiolaging.2009.10.012.