Author/s:
Quek, YE;Fung, YL;Cheung, MWL;Vogrin, SJ;Collins, SJ;Bowden, SC

Title:
Agreement Between Automated and Manual MRI Volumetry in Alzheimer's Disease: A Systematic Review and Meta-Analysis

Date:
2022-08-01

Citation:
Quek, Y. E., Fung, Y. L., Cheung, M. W. L., Vogrin, S. J., Collins, S. J. & Bowden, S. C. (2022). Agreement Between Automated and Manual MRI Volumetry in Alzheimer's Disease: A Systematic Review and Meta-Analysis. Journal of Magnetic Resonance Imaging, 56 (2), pp.490-507. https://doi.org/10.1002/jmri.28037.

Quek Yi-En (Orcid ID: 0000-0003-2683-1219)

Agreement Between Automated and Manual MRI Volumetry in Alzheimer's Disease: A Systematic Review and Meta-Analysis

Yi-En Quek, MPsych(ClinNeuropsych)[1], Yi Leng Fung, MPsych(ClinNeuropsych)/PhD[1], Mike W.-L. Cheung, PhD[2], Simon J. Vogrin, BAppSc(Hons)[3], Steven J. Collins, MD[3], & Stephen C. Bowden, PhD[1,3]

[1]Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia

[2]Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570

[3]Department of Clinical Neurosciences, St Vincent's Hospital Melbourne, 41 Victoria Parade, Fitzroy, Victoria 3065, Australia

Corresponding Author: Correspondence concerning this article should be addressed to Yi-En Quek, Melbourne School of Psychological Sciences, Redmond Barry Building, The University of Melbourne, Parkville, VIC 3010, Australia. Email: yien@student.unimelb.edu.au.

Grant Support: This work was supported by the Australian Commonwealth Government and the University of Melbourne.

Running Title: Automated MRI Volumetry in Alzheimer's

Agreement Between Automated and Manual MRI Volumetry in Alzheimer's Disease: A Systematic Review and Meta-Analysis

Yi-En Quek, MPsych(ClinNeuropsych)[1], Yi Leng Fung, MPsych(ClinNeuropsych)/PhD[1], Mike W.-L. Cheung, PhD[2], Simon J. Vogrin, BAppSc(Hons)[3], Steven J. Collins, MD[3], & Stephen C. Bowden, PhD[1,3]

[1]Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia

[2]Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570

[3]Department of Clinical Neurosciences, St Vincent's Hospital Melbourne, 41 Victoria Parade, Fitzroy, Victoria 3065, Australia

Corresponding Author: Correspondence concerning this article should be addressed to Yi-En Quek, Melbourne School of Psychological Sciences, Redmond Barry Building, The University of Melbourne, Parkville, VIC 3010, Australia. Email: yien@student.unimelb.edu.au.

Grant Support: This work was supported by the Australian Commonwealth Government and the University of Melbourne.

Running Title: Automated MRI Volumetry in Alzheimer's

ABSTRACT

Background

Automated magnetic resonance imaging (MRI) volumetry is a promising tool to evaluate regional brain volumes in dementia and especially Alzheimer's disease (AD).

Purpose

To compare automated methods and the gold standard manual segmentation in measuring regional brain volumes on MRI across healthy controls, patients with mild cognitive impairment, and patients with dementia due to AD.

Study Type

Systematic review and meta-analysis.

Data Sources

MEDLINE, Embase, and PsycINFO were searched through October 2021.

Field Strength

1.0T, 1.5T, or 3.0T.

Assessment

Two review authors independently identified studies for inclusion and extracted data. Methodological quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2).

Statistical Tests

Standardized mean differences (SMD; Hedges' *g*) were pooled using random-effects meta-analysis with robust variance estimation. Subgroup analyses were undertaken to explore potential sources of heterogeneity. Sensitivity analyses were conducted to examine the impact of the within-study correlation between effect estimates on the meta-analysis results.

Results

Seventeen studies provided sufficient data to evaluate the hippocampus, lateral ventricles, and parahippocampal gyrus. The pooled SMD for the hippocampus, lateral ventricles, and parahippocampal gyrus were 0.22 (95% CI −0.50 to 0.93), 0.12 (95% CI −0.13 to 0.37), and −0.48 (95% CI −1.37 to 0.41), respectively. For the hippocampal data, subgroup analyses suggested that the pooled SMD was invariant across clinical diagnosis and field strength. Subgroup analyses could not be conducted on the lateral ventricles data and the parahippocampal gyrus data due to insufficient data. The results were robust to the selected within-study correlation value.

Data Conclusion

While automated methods are generally comparable to manual segmentation for measuring hippocampal, lateral ventricle, and parahippocampal gyrus volumes, wide 95% CIs and large heterogeneity suggest that there is substantial uncontrolled variance. Thus, automated methods may be used to measure these regions in patients with AD but should be used with caution.

Keywords: automated segmentation; manual segmentation; volumetry; magnetic resonance imaging; Alzheimer's disease

INTRODUCTION

Alzheimer's disease (AD) is the predominant pathology underlying dementia in the elderly (1). The ability to detect AD prior to dementia onset, such as in the mild cognitive impairment (MCI) phase, would enable early intervention, which may help to delay cognitive and behavioral decline and maximize patient independence (2). Although typical AD may be relatively straightforward to diagnose, AD can be challenging to detect until symptoms become debilitating (3). Accordingly, it has been suggested that incorporating biomarkers in routine clinical practice can improve accurate and early AD detection (4,5). Among the proposed biomarkers, structural magnetic resonance imaging (MRI) is routinely available in the clinical setting, is non-invasive, and is relatively inexpensive. Hence, structural MRI is well placed to contribute to improving confident early AD diagnosis in routine clinical practice.

Structural MRI has enabled AD-related neurodegenerative changes to be examined in vivo. These studies have revealed a distinct hierarchical vulnerability to neuronal loss across brain regions in patients with AD, with atrophy progression correlating with upstream pathologic tau deposition (6,7). The earliest MRI-based atrophic changes are typically observed in the mesial temporal lobe structures, particularly in the hippocampus and entorhinal cortex (8-11), consistent with early memory impairment (12-17). Atrophy then extends throughout the temporal and parietal lobes and then into the frontal lobes, sparing the primary and secondary sensory areas until late in the disease process (8-11).

Structural MRI-based atrophy measures show potential clinical utility in diagnosing and predicting AD. As regions affected early in the disease process, the hippocampus and entorhinal cortex have been the most studied structures in AD structural MRI studies (18). Several studies have demonstrated that hippocampal

and entorhinal cortex volume are typically reduced in patients with MCI and with dementia due to AD relative to healthy controls (17,19-24). Hippocampal and entorhinal cortex volumes have also been shown to predict conversion to dementia in patients with MCI (24-29). Other studies have also highlighted the clinical utility in the measurement of regional volumes beyond the hippocampus and entorhinal cortex, such as the amygdala (30), parahippocampal gyrus (31), temporal neocortex (23), nucleus accumbens (32), and basal forebrain (33,34). Taken together, these findings suggest that MRI-based regional brain volume measurements could play a valuable role in supporting early AD detection and confirmation in the clinical setting (35).

The current gold standard method to measuring regional brain volumes on structural MRI is manual segmentation (36). A well-trained expert manually outlines the region of interest on each MRI slice. Volume is estimated by multiplying the delineated cross-sectional area by the distance between slices. Several commonly used medical image processing programs to conduct manual segmentation include 3D Slicer (37), Analyze (AnalyzeDirect, Overland Park, KS), ITK-SNAP (38), and ImageJ (39). Manual segmentation is highly accurate, as the trained expert can take into account anatomical variability and correct for image noise based on neuroanatomical knowledge (40). When carefully conducted, manual segmentation also exhibits high reliability within and among observers (41). However, manually tracing structures in a slice-by-slice manner is time-consuming and labor-intensive. Even a well-trained and highly experienced expert may require 1.5 hr per scan to trace a single structure such as the hippocampus (42). Manual procedures are also inherently reliant upon subjective judgements, which may result in lower reliability within and among observers, particularly when image quality is suboptimal and when

observer knowledge and experience is more limited (40,43). These limitations greatly limit translating manual segmentation to routine clinical practice.

Automated segmentation methods, which employ computer algorithms to segment a brain structure and estimate its volume, have been developed to address some of the limitations associated with manual segmentation. Diverse segmentation challenges have led to the development of numerous automated methods. These methods may be broadly categorized as (i) model-based methods, (ii) atlas-based methods, and (iii) machine learning methods (44,45). In model-based methods, an initial contour is placed within the region of interest and its boundaries are iteratively adjusted according to image-derived constraints and a priori anatomical knowledge about the location, size, and shape of the region. In atlas-based methods, a single template or multiple templates with representative segmentation (usually built on manual segmentations) is warped through various registration algorithms to the target image and the template labels are then propagated to the target image space. In machine learning methods, algorithms learn the relationships between voxels in input images by extracting image properties, such as intensity, gradient, edges, and entropy, and then generalizes this learnt expertise to label the voxels in new images. Although these methods have been described independently, it is common that multiple methods are implemented in combination in order to provide improved solutions to various segmentation problems (46-48).

Automated segmentation aims to reduce reliance on manual intervention and the biases associated with subjectivity and thus allows volume measurements to be rapidly and reliably obtained, highlighting the potential for automated methods to be integrated into routine clinical practice. However, automated methods may be less accurate than manual procedures under certain circumstances. First, automated

methods may be less accurate in segmenting brain regions that are small in size, are highly anatomically variable between individuals, or have ambiguous boundaries with neighboring regions (49,50). Second, automated methods may be less accurate in segmenting brains with pathologic morphological changes, such as those that may be seen in patients with AD (51). Third, automated methods may be less accurate when applied to MRI scans acquired at lower magnetic field strengths due to decreased gray-to-white matter tissue contrast (52). Given these potential limitations, and the increasing interest in automated segmentation methods, it is important that automated methods are validated against the gold standard manual segmentation.

To our knowledge, no systematic comparison between automated methods versus manual segmentation in measuring regional brain volumes on MRI in patients with AD has been published. Thus, the current review aimed to evaluate the agreement between automated and manual regional brain volume measurements across healthy controls, patients with MCI, and patients with dementia due to AD. A secondary objective was to investigate heterogeneity in the MRI-based automated and manual regional brain volume measurements, including by brain region, clinical diagnosis, and field strength.

METHODS

This systematic review and meta-analysis was reported in accordance with the PRISMA Statement (53) and was prospectively registered with PROSPERO (CRD42020197275).

Search Methods

Detailed search strategies and syntax were developed in collaboration with a research librarian experienced in systematic reviews. Three electronic bibliographic databases were searched: MEDLINE via OvidSP (including Ovid MEDLINE® and

Epub Ahead of Print, In-Process & Other Non-Indexed Citation and Daily) (1946 to Oct 2021), Embase via OvidSP (1947 to Oct 2021), and PsycINFO via OvidSP (1806 to Oct 2021). The search strategies that were used to search each database are given in Supplementary Table S1. The reference lists of the included studies and any relevant systematic reviews were hand-searched to retrieve additional relevant studies not identified by the database search.

Eligibility Criteria

A study was included when (i) the study participants included individuals diagnosed with MCI or dementia due to AD, (ii) the study evaluated both automated methods and manual segmentation to derive regional brain volume measurements, (iii) the MCI or dementia due to AD diagnosis was a clinical diagnosis made according to any recognized diagnostic criteria, (iv) the outcome was regional brain volume, and (v) the study provided the required data to be included in the meta-analysis.

A study was excluded when (i) the study participants were individuals diagnosed with dementia other than AD (including mixed pathologies), (ii) the MCI or dementia due to AD diagnosis was based solely on biomarker results, (iii) the study was published as an abstract only, or (iv) the study was published in a language other than English.

Study Selection

Two review authors independently reviewed the title and abstract of the studies identified by the search strategy to select potentially relevant studies for inclusion. The full text of these potentially eligible studies was then retrieved and examined to assess compliance with the eligibility criteria. Any disagreement was resolved by consensus discussion between the two review authors.

Data Extraction

One review author extracted the data from the included studies and a second review author checked the extracted data. Disagreements over extracted data were resolved by consensus discussion between the two review authors. Data were extracted according to a standardized data extraction protocol. The data extraction protocol was pilot tested on five randomly selected studies. Review authors who were involved in the data extraction were not blind to study authors, institutions, and publishing journal. Extracted data included: study characteristics (first author, year of publication), participant characteristics (setting, sample size, mean age, proportion of female participants, mean years of education, mean MMSE score), index test characteristics (region of interest, scanner manufacturer, magnetic field strength, automated method, manual segmentation protocol), reference standard characteristics (MCI diagnostic criteria, dementia due to AD diagnostic criteria), and outcome statistics (mean and standard deviation volumes, correlation between automated and manual volumes). If a study did not provide the required outcome data, the study authors were contacted to request the required data. If the study authors did not provide the required data (e.g., the study authors did not reply to the request or the study authors indicated that the requested data were unavailable), then the study was excluded from the review.

Quality Assessment

The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2; Whiting et al., 2011) was used to assess the methodological quality of each included study. The review-tailored QUADAS-2 tool is presented in Supplementary Table S2. Risk of bias was judged as “low”, “high”, or “unclear” on four domains: patient selection, index test, reference standard, and flow and timing (the flow of patients

through the study procedure). Concerns about applicability were judged as “low”, “high”, or “unclear” on three domains: patient population, index test, and reference standard. A study was judged as “at risk of bias” or as having “concerns regarding applicability” when it was judged “high” or “unclear” in one or more domains. Two review authors independently assessed the methodological quality of each included study. Cohen’s kappa statistic was calculated to assess inter-rater reliability for the quality assessment. Disagreement was resolved by consensus discussion between the two review authors.

Data Analysis

The standardized mean difference (SMD; Hedges’ *g*) was used to express the difference between automated and manual volume measurements in each of the included studies. An SMD of 1 indicates that the automated and manual volume measurements differ by 1 standard deviation. Although the comparison between automated and manual volume measurements is based essentially on a repeated-measures design (i.e., each subject serves as his or her own control), the SMD was calculated based on the raw-score standard deviation rather than the change-score standard deviation. The raw-score metric was chosen because the research interest was on comparing the volumes estimated by the two approaches rather than on the change in volume between the two approaches (54).

Most included studies reported relevant data on multiple outcomes based on the same sample. For example, some studies included multiple automated methods or multiple field strengths. Consequently, most studies contributed multiple effect size estimates that were not statistically independent. Conventional meta-analytic procedures, however, assume that the effect size estimates within a given analysis are statistically independent. To account for statistical dependencies among multiple

effect size estimates from the same sample, the robust variance estimation method was used (55). Robust variance estimation is advantageous because it permits statistically dependent effect size estimates to be included within a meta-analysis and hence minimizes data loss (e.g., by having to compute a study-average effect size estimate). Robust variance estimation also does not require knowledge of the within-study covariance structure that is required when using other multivariate meta-analytic methods. A small-sample correction was additionally applied to all analyses as recommended when sample size is small to moderate (i.e., less than 40; Tipton, 2015).

Heterogeneity was assessed using *Q* (57), τ^2 (58), and I^2 (59). The *Q* statistic tests the null hypothesis that all studies are evaluating the same effect and variations are simply caused by chance. Because the *Q* test is underpowered when the number of included studies is few, a *p*-value less than .10 was considered suggestive of significant heterogeneity. τ^2 estimates the variance of the true effect sizes between studies. I^2 describes the relative percentage variation across studies that is due to heterogeneity of effects rather than chance.

Subgroup analyses were undertaken to explore heterogeneity. A priori planned covariates included brain region, clinical diagnosis, and field strength. Because data were too sparse to include brain region as a covariate, separate meta-analyses were undertaken for each brain region with clinical diagnosis (healthy controls vs. patients with MCI vs. patients with dementia due to AD) and field strength (1.0T + 1.5T vs. 1.5T + 3.0T vs. 3.0T only) as covariates.

To estimate the between-studies variance component, the within-study correlation between effect estimates, ρ, was assumed to be .8. To investigate

whether the results were sensitive to the selected ρ value, model estimates were recalculated using different values of ρ ranging from 0 to 1 (60).

Statistical analyses were conducted using R version 4.0.3 (R Core Team, 2020), with the *metafor* (61) and *robumeta* (62) packages, and Microsoft Excel version 16.44.

RESULTS

Search Results

The search process is summarized in Figure 1. The literature search resulted in 414 records. The titles and abstracts were screened to exclude duplicates (*n* = 186) and irrelevant studies (*n* = 137). The remaining 91 records were retrieved and assessed to determine eligibility. Of these, 74 studies were excluded: eight studies because the study population was not individuals with MCI or individuals with dementia due to AD (e.g., healthy controls only, combined individuals with MCI and individuals with dementia due to AD), 10 studies due to wrong index test (e.g., automated method only, manual segmentation only, semi-automated method), 28 studies due to wrong outcome (e.g., change, shape, spatial overlap), three studies because the required data were not provided and the authors did not respond to a data request or were unable to provide the required data, and 25 studies because they were published as an abstract only. Seventeen studies were eligible according to the inclusion and exclusion criteria. For the included studies, Table 1 summarizes the participant demographic and clinical characteristics, and Table 2 summarizes the index test descriptions and characteristics.

Methodological Quality

The methodological quality assessment using the QUADAS-2 is summarized in Figure 2. All the studies were judged as at risk of bias because all the studies

were rated as having at least one domain with high risk. The inter-rater reliability for the overall judgments of risk of bias and applicability was κ = .86 (95% CI .75 to .97).

Patient Selection

All the studies were judged to be at high risk because no study enrolled a consecutive or random patient sample. Moreover, most studies also indicated potentially inappropriate exclusions (e.g., excluding patients with comorbidities, excluding patients based on MRI results). However, no study was judged as having concerns regarding applicability as the patients in the included studies matched those targeted by the review question.

Index Test

Four (24%) studies were judged to be at low risk and 13 (76%) at high risk. Ten studies did not report or did not clearly report whether volume measurements were conducted blind to patient diagnosis. Eleven studies did not assess either inter-observer or intra-observer variability. Only six studies assessed both inter-observer and intra-observer variability in the whole cohort or in a randomly selected subsample. No study was judged as having concerns regarding applicability as the automated methods used and the procedures by which manual segmentation were conducted in the included studies matched those targeted by the review question.

Reference Standard

All the studies were judged to be at low risk because all the studies used widely accepted MCI or dementia due to AD diagnostic criteria and ostensibly applied the diagnostic criteria without reference to information about brain volume measurements. No study was judged as having concerns regarding applicability as the diagnostic criteria used in the included studies matched the conditions targeted by the review question.

Flow and Timing

All the studies were judged to be at low risk because all the studies applied the same diagnostic criteria to patients and included all patients in the analyses.

Findings

Seventeen brain regions were examined across the studies. However, due to limited studies on most brain regions, summary estimates were only calculated for the hippocampus, lateral ventricles, and parahippocampal gyrus.

Hippocampus

Fourteen studies, contributing 68 effect estimates, evaluated both automated and manual hippocampal volume measurements (51,63-75). Participants totaled 256 healthy controls, 667 patients with MCI, and 235 patients with dementia due to AD. The studies were heterogenous in the automated methods used, and some studies evaluated more than one method: six studies used FreeSurfer, two studies used an SPM-based method, one study used FSL-FIRST, and seven studies used methods developed by the study authors. For the manual segmentation, six studies used previously published protocols, five studies used protocols that were adapted from previously published protocols, and three studies used manual segmentation data from the EADC-ADNI harmonized hippocampal segmentation project. SMD ranged from −1.65 to 6.08. Pooled SMD was 0.22 (95% CI −0.50 to 0.93; see Figure 4). There was evidence of significant heterogeneity between studies (Q = 4241.87, df = 67, p < .0001). The estimated within-study heterogeneity variance was τ^2 = 0.33, and the estimated between-study heterogeneity variance was τ^2 = 1.41. The estimated relative percentage variation due to within-study heterogeneity was I^2 = 18.85%, and the estimated relative percentage variation due to between-study heterogeneity was I^2 = 80.73%.

Subgroup analysis by clinical diagnosis showed that the pooled SMD did not significantly vary across healthy controls, patients with MCI, and patients with dementia due to AD ($F_{2,11}$ = 0.46, p = .646). For the healthy controls subgroup, SMD was 0.29 (95% CI −0.71 to 1.28). For the patients with MCI subgroup, SMD was 0.17 (95% CI −0.59 to 0.92). For the patients with dementia due to AD subgroup, SMD was 0.20 (95% CI −0.44 to 0.83). There was evidence of significant residual heterogeneity (Q = 3929.23, df = 65, p < .0001). The proportion of estimated within-study heterogeneity variance explained by clinical diagnosis was R^2 = .00, and the proportion of estimated between-study heterogeneity variance explained by clinical diagnosis was R^2 = .03.

Subgroup analysis by field strength showed that the pooled SMD did not significantly vary across 1.0T + 1.5T, 1.5T + 3.0T, and 3.0T only ($F_{2,11}$ = 2.42, p = .135). For the 1.0T + 1.5T subgroup, SMD was 0.30 (95% CI −0.64 to 1.25). For the 1.5T + 3.0T subgroup, SMD was −0.05 (95% CI −0.11 to 0.02). For the 3.0T only subgroup, SMD was 0.15 (95% CI −0.74 to 1.04). There was evidence of significant residual heterogeneity (Q = 4057.34, df = 65, p < .0001). The proportion of estimated within-study heterogeneity variance explained by field strength was R^2 = .00, and proportion of the estimated between-study heterogeneity variance explained by field strength was R^2 = .00.

Sensitivity analysis with assumed values of ρ ranging from 0 to 1 in .2 increments indicated that the effect size estimates, standard errors, and between-studies variance component were robust to different values of ρ.

Lateral Ventricles

Three studies, contributing five effect size estimates, evaluated both automated and manual lateral ventricle volume measurements (69,76,77).

Participants totaled 53 healthy controls, 0 patients with MCI, and 55 patients with dementia due to AD. For the automated method, one study used an SPM-based method, one study used MRI Studio, and one study used FreeSurfer. For the manual segmentation, one study used a protocol developed by the study authors and the other two studies did not indicate the protocol used. SMD ranged from −0.003 to 0.22. Pooled SMD was 0.12 (95% CI −0.13 to 0.37; see Figure 5). There was evidence of significant heterogeneity among the effect size estimates (Q = 10.53, df = 4, p = .032). The estimated within-study heterogeneity variance was τ^2 = 0.00, and the estimated between-study heterogeneity variance was τ^2 = 0.01. The estimated relative percentage variation due to within-study heterogeneity was I^2 = 0.00%, and the estimated relative percentage variation due to between-study heterogeneity was I^2 = 77.79%.

Planned subgroup analyses by clinical diagnosis and field strength were not conducted because the data were too sparse to produce valid results.

Sensitivity analysis with assumed values of ρ ranging from 0 to 1 in .2 increments indicated that the effect size estimates, standard errors, and between-studies variance component were robust to different values of ρ.

Parahippocampal Gyrus

Two studies, contributing five effect size estimates, evaluated both automated and manual parahippocampal gyrus volume measurements (63,69). Participants totaled 28 healthy controls, 18 patients with MCI, and 27 patients with dementia due to AD. For the automated method, both studies used FreeSurfer. For the manual segmentation, one study used the protocol by Burgmans, et al. (78) and the other study used a protocol developed by the study authors. SMD ranged from −1.25 to 0.41. Pooled SMD was −0.48 (95% CI −1.37 to 0.41; see Figure 6). There was

evidence of significant heterogeneity between studies (Q = 11.02, df = 4, p = .026). The estimated within-study heterogeneity variance was τ^2 = 0.26, and the estimated between-study heterogeneity variance was τ^2 = 0.00. The estimated relative percentage variation due to within-study heterogeneity was I^2 = 61.94%, and the estimated relative percentage variation due to between-study heterogeneity was I^2 = 0.00%.

Planned subgroup analyses by clinical diagnosis and field strength were not conducted because the data were too sparse to produce valid results.

Sensitivity analysis with assumed values of ρ ranging from 0 to 1 in .2 increments indicated that the effect size estimates, standard errors, and between-studies variance component were robust to different values of ρ.

Other Brain Regions

Other brain regions examined by the studies included the amygdala, CA1, entorhinal cortex, inferior prefrontal cortex, lateral occipitotemporal gyrus, midbrain, middle–inferior temporal gyrus, orbitofrontal cortex, pons, posterior cingulate cortex, precuneus, subiculum, superior temporal gyrus, and temporal lobe, in patients with AD; however, due to the limited studies on these brain regions, summary estimates were not calculated. Clerx, et al. (63) compared FreeSurfer against manual segmentation for estimating volumes of the inferior prefrontal cortex, orbitofrontal cortex, posterior cingulate cortex, and precuneus. de Flores, et al. (65) compared FreeSurfer against manual segmentation for estimating volumes of the CA1 and subiculum. Lehmann, et al. (69) compared FreeSurfer against manual segmentation for estimating volumes of the amygdala, entorhinal cortex, lateral occipitotemporal gyrus, middle–inferior temporal gyrus, superior temporal gyrus, and temporal lobe. Nigro, et al. (79) compared an automated method developed by the study authors,

the Landmark-based Automated Brainstem Segmentation (LABS), against manual segmentation for estimating volumes of the midbrain and pons. The automated methods tended to significantly overestimate the volumes of the entorhinal cortex, inferior prefrontal cortex, lateral occipitotemporal gyrus, orbitofrontal cortex, and subiculum, and significantly underestimate the volumes of the CA1 and superior temporal gyrus, relative to manual segmentation, across healthy controls, patients with MCI, and patients with dementia due to AD. There was no significant difference between the automated methods and manual segmentation in estimating the volumes of the amygdala, midbrain, middle–inferior temporal gyrus, pons, posterior cingulate cortex, precuneus, and temporal lobe.

DISCUSSION

Summary of Evidence

The current review evaluated the agreement between automated methods and manual segmentation in measuring regional brain volume on MRI across healthy controls, patients with MCI, and patients with dementia due to AD. Seventeen studies met the inclusion criteria and contributed data to the review. Across the included studies, 17 brain regions were examined. However, summary estimates were calculated only for the hippocampus, lateral ventricles, and parahippocampal gyrus due to limited studies on the other brain regions. The results showed that there was no significant difference between automated methods and manual segmentation in estimating hippocampal, lateral ventricle, and parahippocampal gyrus volumes. For all three regions, however, the 95% CIs around the pooled effect estimates were very wide, and there was evidence of significant heterogeneity of the effect estimates across the studies. These findings suggest that there is substantial uncontrolled

variance in the agreement between automated and manual volume measurements of the hippocampus, lateral ventricles, and parahippocampal gyrus.

Investigations were undertaken to explore the heterogeneity in the automated and manual volume measurements. Due to sparse data, the subgroup analyses that were planned a priori could only be undertaken on the hippocampal data. These analyses showed that the overall agreement between the automated and manual hippocampal volume measurements was invariant across both clinical diagnosis (healthy controls vs. patients with MCI vs. patients with dementia due to AD) and field strength (1.0T + 1.5T vs. 1.5T + 3.0T vs. 3.0T only). However, the 95% CIs around the effect estimates for the levels in the subgroups were all very wide. There was also significant unexplained variance across the effect estimates after accounting for the covariates of clinical diagnosis and field strength. These findings suggest that the agreement between automated methods and manual segmentation in estimating hippocampal volume is generally robust across healthy brains and brains with AD pathology as well as across field strengths. However, the wide CIs and the significant residual heterogeneity indicate that there are other sources of variability in the methodologies of the studies that contributed to the different results across the studies.

Apart from the a priori hypothesized sources of heterogeneity, one other factor that may have introduced heterogeneity in the effect estimates is the automated methods used by the respective studies. Among the included studies, the most commonly used method, FreeSurfer, was only used by six of the 17 studies. Even then, only two of the six studies used the same version of FreeSurfer. Most included studies used methods developed by the study authors or associated research groups, resulting in some methods only being used by a single study (see

Table 2). It has been shown that volume measurements between different automated methods can deviate on average by 24% (80). Hence, the diversity in automated methods across the studies may have contributed to the substantial variability in the effect estimates. Importantly, the diversity in automated methods observed in the included studies highlights the numerous methods that have been developed. The extent to which these methods are comparable requires more detailed examination and was not possible in the current review because of the limited number of studies reporting on the respective methods.

Other potential sources of heterogeneity may relate to the protocol used to guide manual segmentation of the brain regions. For brain regions that are commonly investigated, a large number of segmentation protocols are available. A previous review on protocols for hippocampal segmentation identified 71 different published protocols (81). These protocols can differ substantially in the definitions of the borders of the hippocampus, resulting in significant variability in hippocampal volume measurements across segmentation protocols (82). Segmentation protocols can also differ in the specification of the orientation of the MRI scan, leading to differences in the visualization of the structure of interest and, therefore, variability in the MRI-derived measurements (81-83, Quek Y, unpublished data). In the current review, the protocols used by the studies were heterogenous, and several studies adapted existing protocols to the individual study requirements (see Table 2). Additionally, many studies did not the report the scan orientation in which segmentation was conducted. Validating automated methods against a standard that can produce variable results across studies complicates comparison of the performance of automated methods across studies. The current review as well as previous reviews on hippocampal segmentation protocols highlight the need for more

standardized or harmonized manual segmentation procedures against which to validate automated methods (85).

Strengths, Limitations, and Other Considerations

The current review has several important strengths. First, the literature search conducted was extensive and based on a comprehensive search strategy. Multiple electronic databases were searched and no restriction on publication date was applied to the studies. In addition, primary study authors were contacted where necessary to request unpublished data in order to maximize the number of included studies. Second, given the statistical dependencies present in the data set, a robust variance estimation approach was used to synthesize the data. Conventional univariate meta-analysis is inappropriate in such a situation and would have resulted in some data loss in order to ensure statistical independence.

The current review also has some limitations that deserve comment. First, it was not possible to calculate pooled estimates on some brain regions due to limited data. Consequently, questions concerning the agreement between automated and manual volume measurements in many brain regions in patients with AD remain unresolved. Second, it was also not possible to undertake subgroup analysis according to brain region due to limited available studies. Instead, the SMDs between automated and manual volume measurements were pooled according to brain region, where possible, to evaluate the agreement between automated and manual volume measurements in each brain region. Third, all included studies were judged to demonstrate low methodological quality. This was primarily because no study enrolled a consecutive or random patient sample. Moreover, only six studies addressed both inter-observer and intra-observer reliability in the manual volume

measurements, which is essential to ensuring a reliable standard against which automated volume measurements can be compared.

One additional issue warranting consideration is that although manual segmentation is widely accepted as the gold standard against which to validate automated methods, manual segmentation is prone to variability within and among observers, particularly when conducted on low quality MRI scans or conducted by inexperienced observers or by observers with limited anatomical knowledge (40,43). However, when conducted by trained observers, manual segmentation can be highly accurately and reliable (40,41). Moreover, there currently does not exist a gold standard beyond manual segmentation. Hence, manual segmentation represents the best currently available standard against which to validate automated methods.

Clinical Implications

For automated volumetric analysis to be implemented in routine clinical practice, automated methods must be able to provide accurate and reliable volume measurements. For example, a meta-analysis examining regional atrophy in MCI estimated a mean 3.75% annual entorhinal cortex atrophy rate in patients with MCI compared to 2.41% in healthy controls and a mean 2.53% annual hippocampal atrophy rate in patients with MCI compared to 1.12% in healthy controls (86). Hence, automated methods should be highly accurate to detect the subtle volume changes that are observed in patients to provide clinically useful data. The current review showed that hippocampal, lateral ventricle, and parahippocampal gyrus volumes measured by automated methods are generally comparable to that measured by manual segmentation, with the caveat that there are certain situations under which automated methods may be potentially inaccurate. Moreover, because MRI scans acquired in the clinical setting can vary greatly in quality due to variations in imaging

hardware and acquisition protocols between and even within sites, automated methods should be shown to be robust across imaging parameters to ensure widespread applicability. The current review suggests that automated methods are generally robust across field strength. However, the performance of automated methods across other imaging parameters, such as scanner manufacturer and pulse sequence, remains to be evaluated. Overall, the current evidence suggests that automated methods are a promising tool to incorporate volumetric analysis, particularly of the hippocampus, lateral ventricles, and parahippocampal gyrus, into clinical workup. It is recommended, however, that segmentation results are reviewed by trained observers or imaging experts to verify the accuracy of the segmentations.

Research Implications

The current review highlights the numerous automated methods that have been developed to measure regional brain volume in AD cohorts. It is important that an automated method is adequately validated in order to understand more completely its ability to accurately and reliably measure regional brain volume. For example, highly selective research-oriented samples do not adequately represent the heterogeneity in patients seen in the clinical setting. Moreover, MRI scans obtained in the clinical setting tend to be more variable in quality compared to MRI scans obtained in the research setting, due to more variable MRI hardware, acquisition protocols, and subject compliance. Consequently, clinic-based validation studies will be invaluable in determining how an automated method would translate more broadly to the real-world setting.

The current review also highlights the paucity of research that has been conducted in AD cohorts to evaluate automated volume measurements of brain regions outside the hippocampus. This is perhaps unsurprising because

hippocampal atrophy is a well-established imaging biomarker in AD (4). Nevertheless, volume measurements of other brain regions that are also highly vulnerable in AD, such as the entorhinal cortex, have also shown promise in distinguishing between patients with MCI and healthy controls and in predicting conversion to dementia in patients with MCI (18,87). Further research to validate automated methods to measure the volumes of these other regions would certainly be valuable.

Conclusion

In conclusion, numerous automated methods to measure regional brain volume on MRI scans have been developed. These methods promise a relatively quick yet reliable alternative to manual segmentation. However, automated methods can be less accurate than manual segmentation and must thus be well validated. The current review suggests that automated methods are generally comparable to the gold standard manual segmentation in measuring hippocampal, lateral ventricle, and parahippocampal gyrus volumes in patients with MCI and patients with dementia due to AD. However, the agreement between automated methods and manual segmentation is associated with substantial heterogeneity, which suggests that automated methods may at times incorrectly estimate hippocampal, lateral ventricle, and parahippocampal gyrus volumes. Thus, while automated methods may be used as an alternative to manual segmentation to measure hippocampal, lateral ventricle, and parahippocampal gyrus volumes in patients with AD, these methods should be used with caution. Further work is required to elucidate the conditions under which automated methods produce suboptimal results.

REFERENCES

1. Jellinger KA, Attems J. Prevalence of dementia disorders in the oldest-old: An autopsy study. Acta Neuropathol 2010;119:421-433.
2. Hansen RA, Gartlehner G, Webb AP, Morgan LC, Moore CG, Jonas DE. Efficacy and safety of donepezil, galantamine, and rivastigmine for the treatment of Alzheimer's disease: A systematic review and meta-analysis. Clin Interv Aging 2008;3(2):211-225.
3. Knopman D, Donohue JA, Gutterman EM. Patterns of care in the early stages of Alzheimer's disease: Impediments to timely diagnosis. J Am Geriatr Soc 2000;48(3):300-304.
4. Frisoni GB, Boccardi M, Barkhof F, et al. Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. Lancet Neurol 2017;16(8):661-676.
5. Jack CR, Jr, Barkhof F, Bernstein MA, et al. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimer's & Dementia 2011;7(4):474-485.
6. Thompson PM, Hayashi KM, De Zubicaray G, et al. Dynamics of gray matter loss in Alzheimer's disease. J Neurosci 2003;23(3):994-1005.
7. Whitwell JL, Josephs KA, Murray ME, et al. MRI correlates of neurofibrillary tangle pathology at autopsy: A voxel-based morphometry study. Neurology 2008;71(10):743-749.
8. Scahill RI, Schott JM, Stevens JM, Rossor MN, Fox NC. Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of

fluid-registered serial MRI. Proceedings of the National Academy of Sciences 2002;99(7):4703-4707.

9. Whitwell JL, Przybelski SA, Weigand SD, et al. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 2007;130(7):1777-1786.
10. McDonald CR, McEvoy LK, Gharapetian L, et al. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. Neurology 2009;73(6):457-465.
11. Frisoni GB, Prestia A, Rasser PE, Bonetti M, Thompson PM. In vivo mapping of incremental cortical atrophy from incipient to overt Alzheimer's disease. J Neurol 2009;256(6):916-924.
12. Di Paola M, Macaluso E, Carlesimo G, et al. Episodic memory impairment in patients with Alzheimer's disease is correlated with entorhinal cortex atrophy. J Neurol 2007;254(6):774-781.
13. Petersen RC, Jack CR, Jr, Xu Y-C, et al. Memory and MRI-based hippocampal volumes in aging and AD. Neurology 2000;54(3):581-581.
14. Kramer JH, Schuff N, Reed BR, et al. Hippocampal volume and retention in Alzheimer's disease. J Int Neuropsychol Soc 2004;10(4):639-643.
15. Laakso MP, Hallikainen M, Hänninen T, Partanen K, Soininen H. Diagnosis of Alzheimer's disease: MRI of the hippocampus vs delayed recall. Neuropsychologia 2000;38(5):579-584.
16. Deweer B, Lehericy S, Pillon B, et al. Memory disorders in probable Alzheimer's disease: The role of hippocampal atrophy as shown with MRI. J Neurol Neurosurg Psychiatry 1995;58(5):590-597.

17. Laakso MP, Soininen H, Partanen K, et al. Volumes of hippocampus, amygdala and frontal lobes in the MRI-based diagnosis of early Alzheimer's disease: Correlation with memory functions. J Neural Transm Park Dis Dement Sect 1995;9(1):73-86.

18. Leandrou S, Petroudi S, Kyriacou PA, Reyes-Aldasoro CC, Pattichis CS. Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: A methodological review. IEEE Rev Biomed Eng 2018;11:97-111.

19. Colliot O, Chételat G, Chupin M, et al. Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 2008;248(1):194-201.

20. Juottonen K, Laakso MP, Partanen K, Soininen H. Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. American Journal of Neuroradiology 1999;20(1):139-144.

21. Frisoni GB, Laakso MP, Beltramello A, et al. Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer’s disease. Neurology 1999;52(1):91-108.

22. Du AT, Schuff N, Amend DL, et al. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J Neurol Neurosurg Psychiatry 2001;71(4):441-447.

23. De Santi S, de Leon MJ, Rusinek H, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol Aging 2001;22(4):529-539.

24. Killiany RJ, Hyman BT, Gomez-Isla T, et al. MRI measures of entorhinal cortex vs hippocampus in preclinical AD. Neurology 2002;58(8):1188-1196.

25. deToledo-Morrell L, Stoub TR, Bulgakova M, et al. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. Neurobiol Aging 2004;25(9):1197-1203.

26. Lombardi G, Crescioli G, Cavedo E, et al. Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. Cochrane Database Syst Rev 2020;2020(3).

27. Devanand DP, Pradhaban G, Liu X, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of Alzheimer disease. Neurology 2007;68(11):828-836.

28. Jack CR, Jr, Petersen RC, Xu YC, et al. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 1999;52(7):1397-1397.

29. Stoub TR, Bulgakova M, Leurgans S, et al. MRI predictors of risk of incident Alzheimer disease: A longitudinal study. Neurology 2005;64(9):1520-1524.

30. Frisoni GB, Bocchetta M, Chételat G, et al. Imaging markers for Alzheimer disease: Which vs how. Neurology 2013;81(5):487-500.

31. Visser PJ, Scheltens P, Verhey FR, et al. Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. J Neurol 1999;246(6):477-485.

32. Yi H-A, Möller C, Dieleman N, et al. Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease. J Neurol Neurosurg Psychiatry 2016;87(4):425-432.

33. Teipel S, Heinsen H, Amaro Jr E, Grinberg LT, Krause B, Grothe M. Cholinergic basal forebrain atrophy predicts amyloid burden in Alzheimer's disease. Neurobiol Aging 2014;35(3):482-491.

34. Kilimann I, Grothe M, Heinsen H, et al. Subregional basal forebrain atrophy in Alzheimer's disease: A multicenter study. J Alzheimers Dis 2014;40(3):687-700.

35. McEvoy LK, Brewer JB. Quantitative structural MRI for early detection of Alzheimer's disease. Expert Rev Neurother 2010;10(11):1675-1688.

36. Bobinski M, De Leon MJ, Wegiel J, et al. The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. Neuroscience 1999;95(3):721-725.

37. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30(9):1323-1341.

38. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116-1128.

39. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature Methods 2012;9(7):671-675.

40. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903-921.

41. Boccardi M, Bocchetta M, Ganzola R, et al. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimer's & Dementia 2015;11(2):184-194.

42. Barnes J, Foster J, Boyes RG, et al. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 2008;40(4):1655-1671.

43. Jack CR, Jr, Bentley MD, Twomey CK, Zinsmeister AR. MR imaging-based volume measurements of the hippocampal formation and anterior temporal lobe: Validation studies. Radiology 1990;176(1):205-209.

44. Safavian N, Batouli SAH, Oghabian MA. An automatic level set method for hippocampus segmentation in MR images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2020;8(4):400-410.

45. Pham DL, Xu C, Prince JL. Current methods in medical image segmentation. Annu Rev Biomed Eng 2000;2(1):315-337.

46. Wang H, Das SR, Suh JW, et al. A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. Neuroimage 2011;55(3):968-985.

47. Lötjönen JMP, Wolz R, Koikkalainen JR, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 2010;49(3):2352-2365.

48. Xue J-H, Pizurica A, Philips W, Kerre E, Van De Walle R, Lemahieu I. An integrated method of adaptive enhancement for unsupervised segmentation of MRI brain images. Pattern Recognition Letters 2003;24(15):2549-2560.

49. Xu Y, Jack CR, Jr, O'brien PC, et al. Usefulness of MRI measures of entorhinal cortex versus hippocampus in AD. Neurology 2000;54(9):1760-1767.

50. Frankó E, Insausti AM, Artacho‐Pérula E, Insausti R, Chavoix C. Identification of the human medial temporal lobe regions on magnetic resonance images. Hum Brain Mapp 2014;35(1):248-256.

51. Sánchez-Benavides G, Gómez-Ansón B, Sainz A, Vives Y, Delfino M, Peña-Casanova J. Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. Psychiatry Research: Neuroimaging 2010;181(3):219-225.

52. Kruggel F, Turner J, Muftuler LT. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. Neuroimage 2010;49(3):2123-2133.

53. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 2009;6(7).

54. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods 2002;7(1):105-125.

55. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta‐regression with dependent effect size estimates. Research Synthesis Methods 2010;1(1):39-65.

56. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods 2015;20(3):375-393.

57. Cochran WG. The combination of estimates from different experiments. Biometrics 1954;10(1):101-129.

58. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis: John Wiley & Sons: 2009.

59. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta‐analysis. Stat Med 2002;21(11):1539-1558.

60. Tanner-Smith EE, Tipton E. Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. Research Synthesis Methods 2014;5(1):13-30.

61. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 2010;36(3).

62. Fisher Z, Tipton E. Robumeta: An R-package for robust variance estimation in meta-analysis. arXiv 2015.

63. Clerx L, Jacobs HIL, Burgmans S, et al. Sensitivity of different MRI-techniques to assess gray matter atrophy patterns in Alzheimer's disease is region-specific. Current Alzheimer Research 2013;10(9):940-951.

64. Clerx L, van Rossum IA, Burns L, et al. Measurements of medial temporal lobe atrophy for prediction of Alzheimer's disease in subjects with mild cognitive impairment. Neurobiology of Aging 2013;34(8):2003-2013.

65. de Flores R, La Joie R, Landeau B, et al. Effects of age and Alzheimer's disease on hippocampal subfields: Comparison between manual and FreeSurfer volumetry. Human Brain Mapping 2015;36(2):463-474.

66. Firbank MJ, Barber R, Burton EJ, O'Brien JT. Validation of a fully automated hippocampal segmentation method on patients with dementia. Human Brain Mapping 2008;29(12):1442-1449.

67. Hurtz S, Chow N, Watson AE, et al. Automated and manual hippocampal segmentation techniques: Comparison of results, reproducibility and clinical applicability. NeuroImage: Clinical 2019;21.

68. Ishii K, Soma T, Kono AK, et al. Automatic volumetric measurement of segmented brain structures on magnetic resonance imaging. Radiation Medicine 2006;24(6):422-430.

69. Lehmann M, Douiri A, Kim LG, et al. Atrophy patterns in Alzheimer's disease and semantic dementia: A comparison of FreeSurfer and manual volumetric measurements. Neuroimage 2010;49(3):2264-2274.

70. Leung KK, Barnes J, Ridgway GR, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 2010;51(4):1345-1359.

71. Macdonald KE, Leung KK, Bartlett JW, et al. Automated template-based hippocampal segmentations from MRI: The effects of 1.5T or 3T field strength on accuracy. Neuroinformatics 2014;12(3):405-412.

72. Mulder ER, de Jong RA, Knol DL, et al. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 2014;92:169-181.

73. Safavian N, Batouli SAH, Oghabian MA. An automatic level set method for hippocampus segmentation in MR images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2019;8(4):400-410.

74. Wolf D, Bocchetta M, Preboske GM, Boccardi M, Grothe MJ. Reference standard space hippocampus labels according to the European Alzheimer's Disease Consortium-Alzheimer's Disease Neuroimaging Initiative harmonized protocol: Utility in automated volumetry. Alzheimer's & Dementia 2017;13(8):893-902.

75. Zhu H, Tang Z, Cheng H, Wu Y, Fan Y. Multi-atlas label fusion with random local binary pattern features: Application to hippocampus segmentation. Scientific Reports 2019;9.

76. Karaca O, Buyukmert A, Tepe N, Ozcan E, Kus I. Volume estimation of brain ventricles using Cavalieri's principle and Atlas-based methods in Alzheimer disease: Consistency between methods. Journal of Clinical Neuroscience 2020;78:333-338.

77. Ertekin T, Acer N, Koseoglu E, et al. Total intracranial and lateral ventricle volumes measurement in Alzheimer's disease: A methodological study. Journal of Clinical Neuroscience 2016;34:133-139.

78. Burgmans S, van Boxtel MPJ, van den Berg KEM, et al. The posterior parahippocampal gyrus is preferentially affected in age-related memory decline. Neurobiol Aging 2011;32(9):1572-1578.

79. Nigro S, Cerasa A, Zito G, et al. Fully automated segmentation of the pons and midbrain using human T1 MR brain images. PLOS One 2014;9(1).

80. Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A, Lundervold A. Evaluation of automated brain MR image segmentation and volumetry methods. Hum Brain Mapp 2009;30(4):1310-1327.

81. Konrad C, Ukas T, Nebel C, Arolt V, Toga AW, Narr KL. Defining the human hippocampus in cerebral magnetic resonance images—an overview of current segmentation protocols. Neuroimage 2009;47(4):1185-1195.

82. Geuze E, Vermetten E, Bremner JD. MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Mol Psychiatry 2005;10(2):147-159.

83. Mrzílkova J, Koutela A, Kutová M, et al. Hippocampal spatial position evaluation on MRI for research and clinical practice. PLoS One 2014;9(12).

84. Hasboun D, Chantôme M, Zouaoui A, et al. MR determination of hippocampal volume: Comparison of three methods. American Journal of Neuroradiology 1996;17(6):1091-1098.

85. Boccardi M, Bocchetta M, Apostolova LG, et al. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimer's & Dementia 2015;11(2):126-138.

86. Tabatabaei-Jafari H, Shaw ME, Cherbuin N. Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 2015;1(4):487-504.

87. Pini L, Pievani M, Bocchetta M, et al. Brain atrophy in Alzheimer's disease and aging. Ageing Research Reviews 2016;30:25-48.

Table 1

Participants: Demographic and Clinical Characteristics of the Included Studies

Study	Setting	Healthy Controls				Mild Cognitive Impairment					Dementia due to Alzheimer's Disease				
		Age	*n* (% Female)	Education	MMSE	Age	*n* (% Female)	Education	MMSE	Criteria	Age	*n* (% Female)	Education	MMSE	Criteria
Clerx, Jacobs 2013	Maastricht UMC+ Memory Clinic	64.56 ± 3.40	18 (-)	4.00 ± 1.40	28.89 ± 0.90	65.11 ± 4.50	18 (-)	4.00 ± 1.80	27.61 ± 2.30	Mayo Clinic (Petersen et al., 1999; Petersen et al., 2001)	70.59 ± 9.10	17 (-)	4.00 ± 1.90	21.18 ± 3.90	DSM-IV (APA, 1994) + NINCDS–ADRDA (McKhann et al., 1984)
Clerx, van Rossum 2013	DESCRIPA + VUmc Alzheimer Center	-	-	-	-	70.60 ± 7.60	328 (52%)	10.00 ± 3.80	27.00 ± 2.50	Revised Mayo Clinic (Petersen, 2004; Petersen et al., 1999)	-	-	-	-	-
de Flores 2015	IMAP	70.00 ± 6.40	30 (43%)	12.20 ± 4.00	29.40 ± 0.70	71.70 ± 6.00	17 (53%)	10.50 ± 3.30	27.20 ± 1.30	Revised Mayo Clinic (Petersen & Morris, 2005)	67.40 ± 9.90	18 (67%)	10.70 ± 3.80	21.40 ± 3.90	NINCDS–ADRDA (McKhann et al., 1984)
Ertekin 2016	Erciyes University Neurology Clinic	73.22 ± 3.91	18 (44%)	-	27.77 ± 1.06	-	-	-	-	-	74.75 ± 4.48	20 (35%)	-	17.95 ± 2.18	NIA-AA (McKhann et al., 2011)
Firbank 2008	Newcastle Dementia Case Register	75.00 ± 4.00	9 (33%)	-	27.00 ± 1.60	-	-	-	-	-	75.00 ± 6.00	9 (44%)	-	14.60 ± 4.40	NINCDS–ADRDA (McKhann et al., 1984)
Hurtz 2019	ADCS MCI Donepezil/Vitamin E Trial	-	-	-	-	72.58 ± 6.66	159 (45%)	15.04 ± 3.01	27.51 ± 1.83	Mayo Clinic (Petersen et al., 1999)	-	-	-	-	-
Ishii 2006	Hyogo Institute Hospital for Aging Brain and Cognitive Disorders Registry	61.30 ± 5.90	15 (73%)	-	29.90 ± 0.30	-	-	-	-	-	63.40 ± 6.90	15 (27%)	-	18.10 ± 4.10	NINCDS–ADRDA (McKhann et al., 1984)
Karaca 2020	Balikesir University Hospital Neurology Clinic	62.16 ± 8.20	25 (72%)	-	29.48 ± 0.50	-	-	-	-	-	68.32 ± 7.90	25 (80%)	-	20.28 ± 7.40	NINCDS–ADRDA (McKhann et al., 1984)
Lehmann 2010	National Hospital for Neurology and Neurosurgery Specialist Cognitive Disorders Clinic	59.70 ± 6.30	10 (50%)	-	29.80 ± 0.40	-	-	-	-	-	60.00 ± 7.60	10 (40%)	-	20.40 ± 5.80	NINCDS–ADRDA (McKhann et al., 1984)
Leung 2010	ADNI	78.60 ± 5.40	10 (40%)	-	29.50 ± 0.70	75.30 ± 8.80	10 (30%)	-	27.40 ± 1.80	ADNI (Petersen et al., 2010)	77.20 ± 6.80	10 (30%)	-	27.00 ± 2.70	NINCDS–ADRDA (McKhann et al., 1984)
Macdonald 2014	ADNI	74.45 ± 4.90	18 (61%)	-	29.40 ± 0.70	-	-	-	-	-	74.75 ± 8.10	18 (67%)	-	23.00 ± 2.10	NINCDS–ADRDA (McKhann et al., 1984)
Mulder 2014	ADNI	75.70 ± 6.10	20 (40%)	-	29.20 ± 1.14	73.00 ± 6.70	20 (35%)	-	27.70 ± 1.84	ADNI (Petersen et al., 2010)	72.60 ± 6.90	20 (45%)	-	24.40 ± 1.18	NINCDS–ADRDA (McKhann et al., 1984)
Nigro 2014	ADNI	42.70 ± 2.50	30 (50%)	-	-	-	-	-	-	-	74.00 ± 10.04	10 (40%)	-	-	NINCDS–ADRDA (McKhann et al., 1984)
Safavian 2019	ADNI	73.58 ± 7.67	12 (50%)	-	29.42 ± 0.79	69.67 ± 6.64	12 (50%)	-	26.67 ± 0.98	ADNI (Petersen et al., 2010)	71.08 ± 8.02	12 (42%)	-	22.83 ± 2.66	NINCDS–ADRDA (McKhann et al., 1984)
Sanchez-Benavides 2010	Hospital Clínic de Barcelona + Hospital del Mar	68.50 ± 8.00	41 (54%)	-	28.60 ± 1.50	73.40 ± 7.00	23 (48%)	-	25.80 ± 2.20	IPA–WHO (Levy, 1994)	75.90 ± 6.10	25 (68%)	-	21.60 ± 3.10	DSM-IV-TR (APA, 2000) + NINCDS–ADRDA (McKhann et al., 1984)
Wolf 2017	ADNI	76.00 ± 7.00	44 (50%)	16.00 ± 3.00	29.00 ± 1.00	75.00 ± 8.00	46 (41%)	16.00 ± 3.00	26.00 ± 3.00	ADNI (Petersen et al., 2010)	75.00 ± 8.00	45 (53%)	15.00 ± 3.00	21.00 ± 2.00	NINCDS–ADRDA (McKhann et al., 1984)
Zhu 2019	ADNI	75.79 ± 6.72	29 (45%)	-	-	74.24 ± 7.67	34 (41%)	-	-	ADNI (Petersen et al., 2010)	73.70 ± 8.18	36 (44%)	-	-	NINCDS–ADRDA (McKhann et al., 1984)

Note. ADCS = Alzheimer's Disease Cooperative Study; ADNI = Alzheimer's Disease Neuroimaging Initiative; DESCRIPA = Development of Screening Guidelines and Clinical Criteria for Predementia AD; DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision; IMAP = Imagerie Multimodale de la maladie d'Alzheimer à un stade Précoce; IPA–WHO = International Psychogeriatric Association–World Health Organization; MCI = mild cognitive impairment; MMSE = Mini-Mental State Examination; UMC+ = University Medical Center+; NIA-AA = National Institute on Aging-Alzheimer's Association; NINCDS–ADRDA = National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association; VUmc = Vanderbilt University Medical Center.

Table 2

Index Test: Description and Characteristics of the Included Studies

Study	Region/s of Interest	Scanner Manufacturer/s	Magnetic Field Strength/s (Tesla)	Automated Method/s	Manual Segmentation Protocol/s
Clerx, Jacobs 2013	HPC, IPFC, OFC, PCC, PCUN, PHG	Philips	3.0	FreeSurfer 4.5.0 [Atlas-Based]	HPC, IPFC, OFC, PHG: (Burgmans et al., 2011) PCC: (Jones et al., 2006) PCUN: (Ryu et al., 2010)
Clerx, van Rossum 2013	HPC	Philips + Siemens	1.0 + 1.5	Learning Embeddings for Atlas Propagation (LEAP) [Atlas-Based]	HPC: (Jack, 1994; van de Pol et al., 2007; van de Pol et al., 2009)
de Flores 2015	CA1, HPC, SUB	Philips	3.0	FreeSurfer 5.1.0 [Atlas-Based]	CA1, HPC, SUB: (La Joie et al., 2010)
Ertekin 2016	LV	Philips	1.5	Automatic Lateral Ventricle delIneatioN (ALVIN)/Statistical Parametric Mapping (SPM12) [Atlas-Based]	-
Firbank 2008	HPC	Siemens	1.0	Statistical Parametric Mapping (SPM5) [Atlas-Based]	HPC: (Jack et al., 1997; Duvernoy, 1998)
Hurtz 2019	HPC	GE + Philips + Siemens	1.5	Auto Context Model (ACM) [Machine Learning]	HPC: (Bartzokis et al., 1998)
Ishii 2006	HPC	GE	1.5	Automated Volumetry of Segmented Image System (AVSIS)/Statistical Parametric Mapping (SPM99) [Atlas-Based]	HPC: (Jack et al., 1989; Jack et al., 1992; Watson et al., 1992; Ishii et al., 1996)
Karaca 2020	LV	Philips	1.5	MRI Studio [Atlas-Based]	-
Lehmann 2010	AMYG, ERC, HPC, LOTG, LV, MITG, PHG, STG, TL	GE	1.5	FreeSurfer 4.0.3 [Atlas-Based]	AMYG: (Sheline et al., 1998) ERC: (Insausti et al., 1998) HPC: (Watson et al., 1992)
Leung 2010	HPC	GE + Philips + Siemens	1.5	Multiple-Atlas Propagation and Segmentation (MAPS) [Atlas-Based]	HPC: (Watson et al., 1992; Duvernoy, 1998)
Macdonald 2014	HPC	-	1.5 + 3.0	Hippocampal Multi-Atlas Propagation and Segmentation (HMAPS) [Atlas-Based]	HPC: (Watson et al., 1992; Duvernoy, 1998)
Mulder 2014	HPC	GE + Philips + Siemens	1.5	FreeSurfer 5.1.0 [Atlas-Based] FSL-FIRST 4.1.5 [Atlas-Based]	HPC: (van de Pol et al., 2007)
Nigro 2014	MB, PN	GE + Siemens	1.5 + 3.0	Landmark-based Automated Brainstem Segmentation (LABS) [Model-Based]	MB, PN: (Luft et al., 1999; Oba et al., 2005)
Safavian 2019	HPC	GE + Philips + Siemens	1.5 + 3.0	FreeSurfer 6.0.0 [Atlas-Based] Level-Set Method [Model-Based]	HPC: HarP (Boccardi et al., 2015)
Sanchez-Benavides 2010	HPC	GE	1.5	FreeSurfer 4.0.2 [Atlas-Based]	HPC: (McHugh et al., 2007)
Wolf 2017	HPC	GE + Philips + Siemens	1.5 + 3.0	Statistical Parametric Mapping (SPM8) [Atlas-Based]	HPC: HarP (Boccardi et al., 2015)
Zhu 2019	HPC	GE + Philips + Siemens	1.5 + 3.0	Random Local Binary Pattern (RLBP) [Atlas-Based + Machine Learning] Local Binary Pattern (LBP) [Atlas-Based] Joint Label Fusion (JLF) [Atlas-Based] Local Label Learning (LLL) [Atlas-Based + Machine Learning] Non-Local Patch-based (NLP) [Atlas-Based] Non-Local Weighted voting with Metric Learning (NLW-ML) [Atlas-Based + Machine Learning]	HPC: HarP (Boccardi et al., 2015)

Note. AMYG = amygdala; CA1 = cornu ammonis 1; ERC = entorhinal cortex; HPC = hippocampus; IPFC = inferior prefrontal cortex; LOTG = lateral occipitotemporal gyrus; LV = lateral ventricles; MB = midbrain; MITG = middle–inferior temporal gyrus; OFC = orbitofrontal cortex; PCC = posterior cingulate cortex; PCUN = precuneus; PHG = parahippocampal gyrus; PN = pons; STG = superior temporal gyrus; SUB = subiculum; TL = temporal lobe.

Figure 1

PRISMA Flow Diagram

[FIGURE 1]

Figure 2

Risk of Bias and Applicability Concerns Summary

[FIGURE 2]

Figure 3

Forest Plot of the SMDs (With 95% CIs) Between Automated and Manual MRI Hippocampus Volume Measurements

[FIGURE 3]

Note. A negative SMD value indicates that automated methods underestimated regional volume relative to manual segmentation. CI = confidence interval; SMD = standardized mean difference.

Figure 4

Forest Plot of the SMDs (With 95% CIs) Between Automated and Manual MRI Lateral Ventricle Volume Measurements

[FIGURE 4]

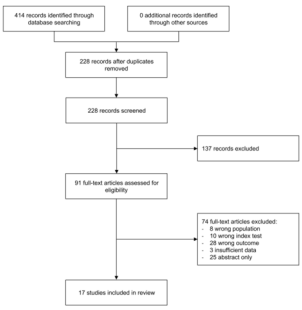
Note. A negative SMD value indicates that automated methods underestimated regional volume relative to manual segmentation. CI = confidence interval; SMD = standardized mean difference.

Figure 5

Forest Plot of the SMDs (With 95% CIs) Between Automated and Manual MRI Parahippocampal Gyrus Volume Measurements

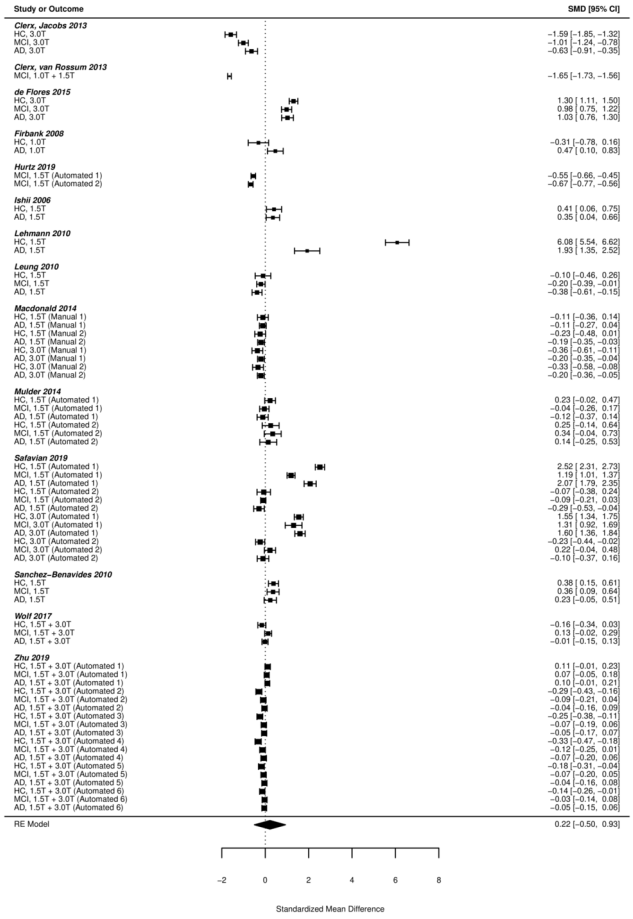
[FIGURE 5]

Note. A negative SMD value indicates that automated methods underestimated regional volume relative to manual segmentation. CI = confidence interval; SMD = standardized mean difference.

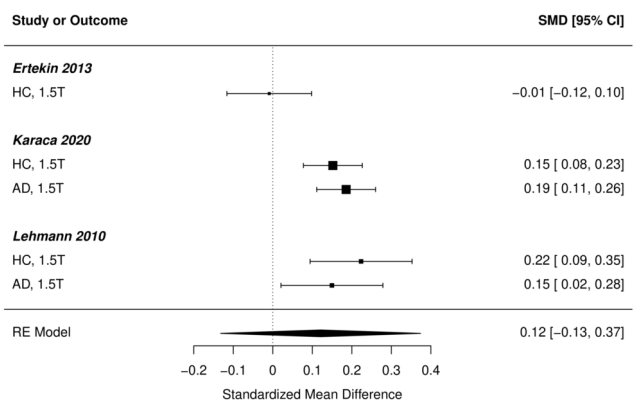


JMRI_28037_Figure 1.tiff

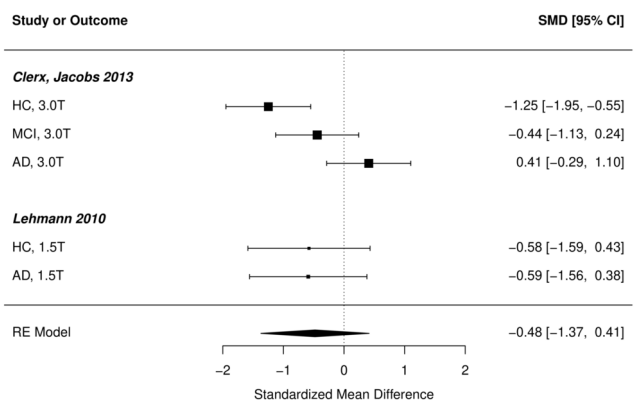
JMRI_28037_Figure 2_Revision.tiff



JMRI_28037_Figure 3_Revision.tiff



JMRI_28037_Figure 4.tiff



JMRI_28037_Figure 5.tiff