Intracranial volume in neuroimaging

Estimation and use in regional brain volume normalization

Niklas Klasson

Department of Psychiatry and Neurochemistry
Institute of Neuroscience and Physiology
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2019

Cover illustration: Work in progress by Niklas Klasson

Intracranial volume in neuroimaging

niklas.klasson@gu.se

ISBN 978-91-7833-304-2 (PRINT)
ISBN 978-91-7833-305-9 (PDF)
http://hdl.handle.net/2077/58238

Printed in Mölndal, Sweden 2019
Printed by BrandFactory AB

“Jag tror inget som jag inte vet.”

Skalman

To my parents.

Intracranial volume in neuroimaging

Estimation and use in regional brain volume normalization

Niklas Klasson

Department of Psychiatry and Neurochemistry
Institute of Neuroscience and Physiology
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ISBN 978-91-7833-304-2 (PRINT)
ISBN 978-91-7833-305-9 (PDF)

http://hdl.handle.net/2077/58238

ABSTRACT

The aim of this thesis is to validate methods for estimation of intracranial volume in magnetic resonance images and to improve our understanding of the effect of intracranial volume normalization.

To achieve the first part of the aim, 62 gold standard estimates of intracranial volume were generated by manually segmenting 1.5 T T1-weighted magnetic resonance images. These estimates were then used to validate a more work-efficient manual method that is frequently used in neuroimaging research. We also proposed an even more work-efficient method for situations where only a strong linear association between estimate and gold standard are required (rather than a strong agreement). Finally, we evaluated the validity of a frequently used automatic method for estimation of intracranial volume. To achieve the second part of the aim, we presented mathematical functions that predict the effect of intracranial volume normalization on the mean value and variance of the brain estimates and their Pearson's correlation to intracranial volume.

We found that segmentations of one intracranial area every 10th mm in magnetic resonance images will result in valid estimates of intracranial volume (intra-class correlation with absolute agreement to gold standard estimates >0.998). The segmentation of two intracranial areas and the estimation of the perpendicular intracranial width will result in estimates with strong linear association to gold standard estimates (Pearson's correlation >0.99). It was also shown that FreeSurfer's automatic estimates of intracranial volume risk being biased by total brain volume. Further, the presented mathematical functions closely predicted the effect of intracranial volume normalization on certain statistics of brain estimates, both in a simulation and compared to actual data from other studies. All these findings contribute to an improved intracranial volume estimation and a better use of intracranial volume in regional brain volume normalization.

Keywords: magnetic resonance imaging, intracranial volume, normalization

SAMMANFATTNING PÅ SVENSKA

Den här avhandlingen har två syften. Det första syftet är att validera metoder för estimering av skallhålans volym i magnetkamerabilder. Det andra syftet är att utöka vår förståelse inom medicinsk bildanalys för vad som sker vid normalisering för skallhålans volym.

För att uppfylla det första syftet i avhandlingen gjordes manuell utlinjering av volymen av 62 skallhålor i 1.5 T T1-viktade magnetkamerabilder. Detta gjordes med en ytterst utförlig metod för att få referensvolymer att använda vid validering av andra mer användarvänliga metoder. Dels utvärderade vi en manuell metod som används flitigt i hjärnavbildningsforskning, dels en metod som vi själva föreslår för det fall man endast efterfrågar estimat av skallhålans volym med starkt linjärt samband till referensvolymer (snarare än en stark likhet). Slutligen validerade vi också en automatisk metod för estimering av skallhålans volym som ofta används i hjärnavbildningsforskning. För att uppfylla det andra syftet presenterade vi matematiska funktioner som förutsäger effekten av normalisering för skallhålans volym på estimat av regionala volymer. De matematiska funktionerna beskriver hjärnestimatens förväntade medelvärde, varians och Pearsons korrelationskoefficient till skallhålans volym efter normalisering.

I vår första studie fann vi att segmentering av areor av skallhålan med 10 mm mellanrum ger valida estimat av dess volym (intraklasskorrelation till våra referensvolymer >0.998). I vår andra studie fann vi att estimat baserat på två areor av skallhålan samt skallhålans bredd hade ett starkt linjärt samband till våra referensvolymer (Pearsons korrelation >0.99). I den tredje studien visade vi att FreeSurfer-estimat av skallhålans volym, som erhålls automatiskt, är beroende av den totala hjärnvolymen och därför kan vara vilseledande vid fall av hjärnatrofi. I vår fjärde studie visade vi att de matematiska funktioner som presenterades väl kunde predicera effekten av normalisering för skallhålans volym. Prediktioner gjordes både på simuleringar och faktiska data från tidigare studier. Sammantaget bidrar alla dessa fynd till att förbättra estimeringen av skallhålans volym utifrån magnetkamerabilder samt dess användning för normalisering av regionala hjärnvolymer.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Klasson Niklas, Olsson Erik, Rudemo Mats, Eckerström Carl, Malmgren Helge, Wallin Anders. Valid and efficient manual estimates of intracranial volume from magnetic resonance images.
BMC Medical Imaging. 2015; 15:5.

II. Klasson Niklas, Olsson Erik, Eckerström Carl, Malmgren Helge, Wallin Anders. Delineation of two intracranial areas and the perpendicular intracranial width is sufficient for intracranial volume estimation.
Insights into Imaging. 2018;9(1):25-34.

III. Klasson Niklas, Olsson Erik, Eckerström Carl, Malmgren Helge, Wallin Anders. Estimated intracranial volume from FreeSurfer is biased by total brain volume.
European Radiology Experimental. 2018; 2:24.

IV. Klasson Niklas, Olsson Erik, Eckerström Carl, Malmgren Helge, Wallin Anders. Statistics of brain estimates normalized by intracranial volume.
Manuscript.

CONTENT

ABBREVIATIONS V

1 INTRODUCTION 1

1.1 Dementia diseases 1

1.2 Structural magnetic resonance imaging 4

1.3 Analysis of structural magnetic resonance images 6

1.4 Interpretation of structural brain segmentations 8

1.5 Intracranial volume normalization 11

1.6 Manual estimation of intracranial volume 15

1.7 Automatic estimation of intracranial volume 20

1.8 Knowledge gaps 24

2 AIM 26

2.1 Specific aims 26

3 MATERIAL AND METHODS 27

3.1 The Gothenburg MCI study 28

3.2 Study participants 28

3.2.1 Healthy controls 28

3.2.2 Patients 29

3.2.3 Exclusion criteria 31

3.3 MR examination 31

3.4 Sample selection 32

3.4.1 Participant demographics 33

3.5 Image preprocessing 35

3.6 Manual segmentation 40

3.6.1 Segmentation tool 40

3.6.2 Segmentation protocol 42

3.6.3 Intra- and inter-rater reliabilities ... 43

3.7 Image masks ... 45

3.8 Paper I ... 46

3.9 Paper II ... 47

3.10 Paper III ... 48

3.11 Paper IV ... 49

3.12 Statistics ... 52

4 Results ... 62

4.1 Paper I ... 62

4.2 Paper II ... 63

4.3 Paper III ... 64

4.4 Paper IV ... 65

5 Discussion ... 66

5.1 Method validation ... 66

5.1.1 Significance testing ... 68

5.1.2 Effect size ... 69

5.1.3 Adjustment for multiple comparisons ... 70

5.1.4 Sample selection ... 71

5.2 Manual estimation of intracranial volume ... 72

5.3 Estimation of intracranial volume using FreeSurfer ... 76

5.4 Effects of intracranial volume normalization on brain estimates ... 80

5.4.1 Reduced coefficient of variation ... 81

5.4.2 Reduced linear association to intracranial volume ... 85

5.4.3 Reduced estimation reliability ... 89

5.5 Effects of intracranial volume normalization on a third factor ... 92

5.5.1 Gender differences ... 93

5.5.2 Diagnostic accuracy ... 95

5.6 Final thoughts 98

6 CONCLUSION 101

7 FUTURE PERSPECTIVES 102

ACKNOWLEDGEMENTS 103

REFERENCES 104

APPENDIX 111

ABBREVIATIONS

BET	Brain extraction tool (software tool)
CDR	Clinical dementia rating (clinical rating scale)
CI	Confidence interval (statistical estimate)
CSF	Cerebrospinal fluid
DICOM image	Digital imaging and communication in medicine image (file format)
Et al.	*Et alii*/and others
eTIV	Estimated total intracranial volume (estimate from FreeSurfer)
EXIT	Executive interview (cognitive testing)
FAST	FMRIB automated segmentation tool (software)
FLAIR	Fluid-attenuated inversion recovery (MRI sequence)
FMRIB	Oxford center for functional MRI of the brain
FSL	FMRIB software library (software package)
GDS	Global deterioration scale (clinical rating scale)
ICA	Intracranial area
ICV	Intracranial volume
I-FLEX	Investigation of flexibility (cognitive tests)
ITK-SNAP	Insight segmentation and registration toolkit-SNAP (software)
MATLAB	Matrix laboratory (software package)
MCI	Mild cognitive impairment
MIDAS	Medical image display and analysis software
MIST	Medical image segmentation tool (software)
MMSE	Mini-mental state examination (cognitive tests)
MNI Display	Montreal neurological institute Display (software)
MNI305	Montreal neurological institute 305 (a head atlas)
MR	Magnetic resonance
MRI	Magnetic resonance image
n	Number of observations
Nifti image	Neuroimaging informatics technology initiative image (file format)
NLSS	Non-local spatial STAPLE (ICV estimation method)
OPLS	Orthogonal projections to latent structures (statistical method)

PD-w	Proton density-weighted (MRI sequence)
PhD	*Philosophiae doctor*
PIVUS	Prospective investigation of vasculature in Uppsala seniors (study cohort)
p-value	Probability of an observation given a null hypothesis (statistical estimate)
r	Pearson's correlation coefficient
RBM	Reversed brain mask (software tool)
SPM	Statistical parametric mapping (software package)
STAPLE	Simultaneous truth and performance level estimation (MRI analysis tool)
STEP	Stepwise comparative status analysis (cognitive tests)
T	Tesla (unit for magnetic field strength)
T1-w	T1-weighted (MRI sequence)
T2-w	T2-weighted (MRI sequence)

Variables used in equations

$b, b_{1,} b_2,$	*Brain estimates: all, from sample 1, from sample 2*
icv, icv_1, icv_2	*Intracranial volume estimates: all, from sample 1, from sample 2*
n_1, n_1	*Number of observations in sample 1, and in sample 2*
C_b, C_{icv}	*Coefficient of variation for brain and intracranial volume estimates*
$\bar{b}, \overline{icv}$	*Mean value of brain and intracranial volume estimates*
s_b, s_{icv}	*Standard deviation of brain and intracranial volume estimates*
s_b^2, s_{icv}^2	*Variance of brain and intracranial volume estimates*
b_{norm}	*ICV normalized brain and intracranial volume estimates*
$\overline{b_{norm}}$	*Mean of ICV normalized brain estimate*
$r_{b,icv}$	*Pearson's correlation between brain and intracranial volume estimates*
z	*z value from a standard normal distribution*

1 INTRODUCTION

This thesis is about the estimation and use of intracranial volume (ICV) in neuroimaging and more specifically in structural magnetic resonance (MR) imaging. While the topic is broad and applicable to a number of areas in psychiatry and neurology, my interest came through dementia research. My PhD studentship has been in a research group specialized in dementia diseases where I was to analyze an existing set of medical images. However, initial discussions with coworkers sparked my interest in ICV. Brain volumes differ between individuals due to the size of the head. Larger heads naturally contain larger brains. In dementia disease research, we want to separate the healthy from the ill before the illness is obvious and one way we try to achieve this is by using the size of regional brain volumes. However, as the size of these volumes vary with the size of one's head, we instead risk ending up separating those with large heads from those with small heads. This risk is often accounted for in dementia research by entering ICV into the statistical analyzes, but how this is done varies and seems to be poorly understood. While my goal for long was to continue with analyzing the medical images to learn more about dementia diseases once I understood how we should use ICV to account for head size variability, eventually these plans were put on ice. My entire thesis ended up being just about ICV. Still, as my interest in ICV came from research about dementia diseases, I will introduce my research to the reader through this context.

1.1 DEMENTIA DISEASES

Dementia refers to a syndrome of pronounced cognitive impairment beyond what is expected by normal aging and that reduces the capacity to perform activities of daily living. There are a number of causes of dementia, such as traumatic brain injury, infections, drug misuse, and more commonly dementia diseases[1,2]. The most common dementia diseases are Alzheimer's disease, vascular dementia, mixed dementia (combined Alzheimer's and cerebrovascular disease), Lewy-body dementia, and frontotemporal dementia. No common denominator separates the dementia diseases from other causes of dementia. However, the dementia diseases generally include progressive cognitive decline along with progressive neuropathological changes[1,2]. In Table 1, I present some typical characteristics of some of the dementia diseases.

Table 1. Characteristics of dementia diseases

Disease	Symptoms	Brain damage
Alzheimer's disease	Impaired memory and impaired learning ability are early signs of Alzheimer's disease. Later on, fine motor skills (movement), language ability, and eventually social skills may also be affected. Depression, apathy, irritability, and agitation are also common symptoms.	Hippocampal and parietotemporal atrophy are early signs of Alzheimer's disease.
Frontotemporal dementia	Behavioral changes and/or language impairments. For example, lessened interest in socializing, less restraints, impaired planning/organizing ability, poor judgement. Difficulties with findings words or understanding single words. Grammatical errors and limited vocabulary.	Atrophy in the frontal lobe, the anterior temporal lobe, and sometimes the parietal lobe.
Vascular dementia	Reduced cognitive processing speed. Impaired sustained, selective and otherwise complex attention. Impaired executive cognitive functions (such as problem solving). Personality and mood changes and depression are other symptoms.	Infarcts, hemorrhages, white matter lesions.

Typical characteristics of three common dementia diseases[2].

As brain damage is irreversible and dementia diseases typically are progressive, it is important to detect these diseases as early as possible, preferably before they affect the patient’s daily life. By early detection, potential treatments will have a greater impact on the patient’s life. Many of the patients that are referred to a dementia specialist have an impaired cognitive function that does not yet substantially affect their daily living. Such cognitive impairment is called mild cognitive impairment. Even though patients with mild cognitive impairment do not get a dementia disease diagnosis at the time of examination, about 5–10% of them will be diagnosed with a dementia disease for each year to come[3]. Still, after five years, about 60% of these patients have not progressed in their cognitive impairment[3] and many will remain in mild cognitive impairment long after that.

In accordance with the definition of dementia, differences in cognitive function have been detailed using basic cognitive testing[4] or more advanced neuropsychological tests[5]. Differences have also been shown using regional brain volumes estimated from MR images[6] and traces of Aβ42 (a certain peptide) and tau (certain proteins) found in cerebrospinal fluid[7] and on positron emission tomography[8]. When using these markers to try to predict conversion to dementia (or to some dementia disease), a strong diagnostic accuracy is often seen[9-11]. However, the diagnostic accuracy tends to be weaker in the earlier stages of disease. For example, a lower diagnostic accuracy has been shown using neuropsychological tests in patients with subjective cognitive impairment compared to patients with objective cognitive impairment[12].

There are ways to improve our markers for dementia diseases. One way is simply to redefine the diseases. For example, including presence of Aβ42 as a necessary diagnostic criterion for possible Alzheimer’s disease will increase the specificity (ability to tell who are not diseased) of Aβ42 as a marker for this diagnosis. It is also possible to come up with new markers through new technology or by applying existing technology in a new way. Lastly, it is possible to improve existing markers by improving methodology, either how we measure the markers or how we use them for analysis. This thesis focus on the latter approaches and more specifically on the estimation and use of ICV to improve brain volume estimates as markers for disease.

1.2 STRUCTURAL MAGNETIC RESONANCE IMAGING

At the diagnosis of dementia diseases, computed tomography or structural magnetic resonance (MR) imaging can be used to rule out other causes of dementia. These other causes might for example be brain tumor or subdural hematoma. MR imaging may also strengthen specific dementia diagnoses, for example by the presence of atrophy in the temporal lobe (sign of Alzheimer's disease) or white matter changes (sign of vascular disease).

The quality (resolution, signal-to-noise ratio and image contrast) in MR images mainly depends on the strength of the magnetic field of the MR scanner and the time used to do the scan. With longer scanning time, better image quality is achievable[13]. However, with longer scanning sessions comes the risk of the patients moving in the scanner. Movements may drastically lower the image quality and result in image artifacts. The strength of the magnetic field is measured in tesla (T) where one tesla is about 20,000 times the strength of the earth's field at the surface[13]. In today's clinical settings, 1.5 T and 3 T MR scanner are used, but the 1.5 T scanners are being phased out.

The scan parameter settings in MR acquisition determine how tissues appear in the resulting images. Two main types of scan sequences are the T1- and T2-weighted ones. T1-weighted images are generally thought to be optimal for maximizing the contrast in the images between gray and white brain matter, but does less well in separating the skull from the cerebrospinal fluid. T2-weighted images have an inverted grayscale and lower contrast between gray and white brain matter than T1-weighted images, but separate the skull from the cerebrospinal fluid better and is useful for visualizing white matter changes and brain tumors. A T1-weighted MR acquisition is visualized in Figure 1 on the next page.

During an MR examination, a number of MR acquisitions with different scanner settings are usually produced. Each acquisition takes about 2–6 minutes and a full examination in dementia disease evaluation about 30 minutes. Such an MR examination costs about 5000 Swedish kronor (year 2018).

MR acquisitions are often converted into in an image format called DICOM (Digital Imaging and COmmunication in Medicine) when analyzed outside of the clinical setting. Other image formats exist as well, such as the NIfTI (Neuroimaging Informatics Technology Initiative) format. However, from here on we will simply refer to DICOM images from a MR examination as MR images.

A MR acquisition is often saved as a set of MR images where each image represents a slice of the three-dimensional structure that was scanned (see Figure 1). Besides the image data, the MR images contain information about for example how distance in the images is related to distance in real space. While the smallest element in a normal digital image is called a pixel, the smallest element in a MR image is called a voxel. This difference is due to the three-dimensionality of the MR images. Just as with grayscale pixels, each voxel does only contain one color value. The color value of a voxel is often visualized as a grayscale intensity that is related to, but not specific to, some tissue type in the brain/head. What tissue a grayscale intensity represents will depend on the scanner setting and scanner variability.

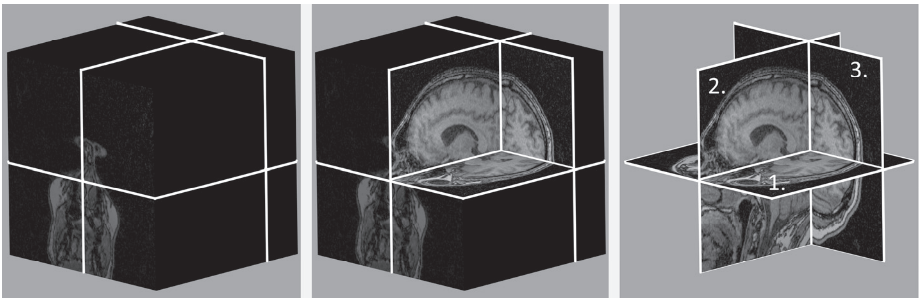


Figure 1. 3D visualization of a MR acquisition. The lefthand image shows a representation of the whole MR acquisition that is constituted by many millions of voxels (small rectangular boxes). In the middle image, the scanned brain is revealed (by removing voxels). In the righthand image, one transversal (1), sagittal (2), and coronal (3) MR image is shown. It is common to examine MR acquisitions through MR images in one of these three orientations.

1.3 ANALYSIS OF STRUCTURAL MAGNETIC RESONANCE IMAGES

To get a medical opinion guided by the findings in the MR images, the images are visually analyzed by a radiologist. As the quality of the analyzes varies depending on who did them, the medical opinion is in danger of varying in quality too. To minimize this risk in dementia disease evaluation, it has been suggested to use certain rating scales when visually analyzing medial temporal lobe atrophy, global cortical atrophy, and white matter changes[14]. By using the suggested rating scales during the visual analysis, one gets criteria for what should be analyzed and how.

Visual rating scales often give only a rough assessment of the state of the brain while manual or automatic segmentations can be used to get continuous measures that enable a higher level of differentiation. Segmentation refers to the demarcation of a specific structure in the MR images by which for example the volume or area of a structure can be calculated. Manual segmentations are often performed by demarcating the structure using some drawing software specialized for MR images. There are also program packages that automatically segment MR images. The higher differentiation achievable by MR segmentations makes them useful in neuroimaging research. However, segmentations only give estimates of the absolute size of brain regions or lesions. With visual rating scales, change in a brain region may also be gathered from a single MR acquisition. This is possible by a visual comparison of the brain region of interest to other regions in the brain and by knowledge about how the region “should look” under different circumstances. The possibility to estimate change (for example brain atrophy) with visual rating scales is probably one reason why they still are used in clinical settings.

The accuracy of both manual segmentation and visual rating scales depends on the rater’s ability to follow guidelines in the assessment reliably. Even from the most skillful raters, some errors can be expected. With automatic segmentation, the procedure of the segmentation can be described and followed rigidly. Thus, with automatic software it is possible to achieve perfect reliability of segmentations. The use of automatic software also reduces the workload and time needed to do the segmentations. Yet another advantage is that automatic software are not affected by visual illusions (see Figure 2 on the next page). However, today’s automatic software generally depends in several ways on visual assessments of the MR images with all its flaws. Visual assessments are needed for constructing/training the software, to evaluate it, and to check for gross errors in the segmentation process. Still, automatic segmentations have already replaced manual segmentation in neuroimaging research. Three reasons to the widespread use of automatic software are 1) big data sets are often being analyzed, which would be painful to analyze manually, 2) automatic methods have enabled more people (non-experts) to perform brain segmentations, and 3) research is much easier to replicate when automatic methods are being used.

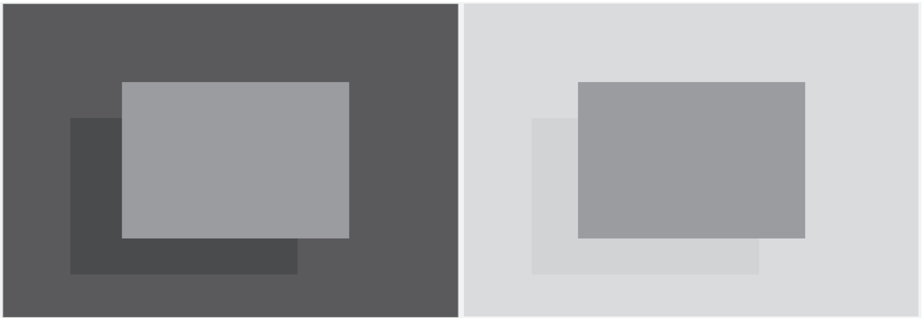


Figure 2. Visual illusion. The middle rectangle in the left part of the image might seem brighter than that in the right part. However, both rectangles are equally bright (it is just the context that differs). Similar visual illusions might interfere with our ability to, for example, correctly differentiate tissue types in visual/manual ratings of magnetic resonance images.

1.4 INTERPRETATION OF STRUCTURAL BRAIN SEGMENTATIONS

It is common to segment MR images in order to estimate the volume, area, or length of a brain region. Less commonly, other features of the brain region are estimated too, such as texture[15] or shape[16]. All these estimates will vary by artificial variance due to estimation (user or method related) errors and fluctuations in the MR imaging. They will also vary due to physiological factors such as cell density, water content, presence of protein assemblies, inflammation and more. For brain estimates from structural MR images, we can presently only speculate on how all these factors come into play.

Let us say that we detect a 1 ml difference between two estimates of hippocampal volume and that the MR acquisitions are from the same participant who was examined twice within an hour using the same MR scanner. We cannot know why the estimates differ. One possible interpretation, given the short amount of time between the examinations and that the same participant is studied, is that the difference is due to estimation error. If the same difference was detected between two MR acquisitions from different participants where one participant was scanned two years later than the first (but on the same MR scanner), we might prefer a different interpretation. Let us say that the participant with the larger hippocampal volume is a 23-year-old healthy male while the other participant is an 85-year-old female with Alzheimer's disease. One possible interpretation of the volume difference still is estimation error. Other interpretations are loss of neurons due to Alzheimer's disease, loss of neurons due to age, dehydration due to age, different head sizes due to gender, imaging artifacts due to fluctuations in the MR scanner and so on. All these factors will potentially affect the brain estimates. With so many potential explanations, a brain estimate is hard to interpret, especially so without knowing its context. Using other MR techniques such as magnetic resonance spectroscopy or magnetic resonance fingerprinting[17] further information is possible to gain about the specific brain regions.

Still, it is also possible to investigate the association of brain estimates to other factors. If, for example, we estimate hippocampal volume in a large sample of participants, we expect variation in the estimated volumes due to a lot of factors. We might hypothesize that one such factor is hearing ability. If we also measure hearing ability in the sample, we can evaluate if the variability in this ability is associated with the variability of the hippocampal volumes. In this way, we might come to the conclusion that the size of hippocampal volume is associated with hearing ability. However, if we do find such an association it does not mean that the one affects the other. A third factor, such as age, could affect both hearing ability and hippocampal volume and cause the association found between these estimates. Further, just because an association is seen in our sample, there is not necessarily an association in the population (but the probability of that can be evaluated using statistical tests).

When evaluating associations to brain estimates, it is often possible to calculate the amount of variability in the brain estimates that a certain factor explains in the sample. For example, in a study by Barnes *et al.*[18], gender explained about 17% of the total variance in total brain volume in their sample. Thus, about 83% of the total variance was still unexplained. Another possibility is to describe, with some mathematical function, how the brain estimates depend on the factor of interest. In the same study, Barnes *et al.*[18] showed that increased age was associated with reduction in hippocampal volume by a factor of 0.36%/year (after adjusting for gender, ICV, and MR scanner upgrade). Yet another way to interpret the brain estimates is in terms of how it affects the probability of having a disease. This is made possible by expressing the proportion of diseased participants compared to the number of healthy participants in the sample as a function of the size of the brain estimates. By doing so, the size of the brain estimates becomes a marker of disease. By deciding at which probability an individual should be considered diseased, brain estimation can even be transformed into a yes-or-no diagnostic tool. The diagnostic accuracy of such a tool is often judged by its sensitivity (percent of diseased participants correctly diagnosed) and specificity (percent of healthy participants considered healthy). The diagnostic accuracy that is achievable using a certain brain estimate depends on how much variability of the estimate that can be explained by diagnostic status (or interchangeably how well the estimate explains the variability in diagnostic status).

Besides gender[18,19], age[18,20], and psychiatric diseases[21], the size of different brain estimates have been shown to be associated to a number of different factors. Factors such as heritability[22], chronic stress[23], aerobic fitness[24], bipolar disorder[25], becoming a taxi driver in London[26] or a medical student in Munich[27]. While the causality of some of these associations might be questionable, yet another association that is not controversial is that between regional brain volume and whole brain volume.

ICV is often seen as a proxy for the size of the whole brain at its peak (premorbid brain volume). About 10–50% of the variance in regional brain structures can be explained by ICV[18,28]. For example, it has been shown that ICV explains about 5–15% of the variance in the volume of nucleus accumbens[28], 9–15% in hippocampal volume[18,28], 15–25% in the volume of amygdala[18,28], and 40–50% in the volume of thalamus[28]. ICV also explains about 15–35% of the variance in most neocortical volumes[28].

1.5 INTRACRANIAL VOLUME NORMALIZATION

In its broadest sense, ICV normalization is done to adjust brain estimates for interindividual differences related to head size/premorbid brain volume. A reasonable clarification of this statement is that

> *ICV normalization is done to reduce the proportion of the total variance of a brain estimate that is predicted by ICV, using some statistical model that supposedly describes some true relationship between ICV and the brain region.*

It is through this perspective that I will discuss ICV normalization. I will often refer to the reduction of variance mentioned above as a reduction of "unwanted" variance.

By reducing unwanted variance in a brain estimate, we might improve upon our understanding of some phenomenon under study in relation to the brain region. The effect of the reduction will differ depending on whether the unwanted variance is independent of the phenomenon under study or not.

By reducing independent unwanted variance by ICV normalization, we might facilitate the detection of a difference between two samples or an association between the phenomenon under study and the brain estimate. This allows for the use of smaller samples or for making a statistical inference based on more subtle associations or differences (with retained sample sizes). The opposite risks being true if we reduce unwanted variance that is dependent on the phenomenon under study. This might still be useful. When all variance that is explained by ICV is removed, we can draw conclusions about the phenomenon as if ICV were a constant.

As seen in Section 1.4, between 10–50% of the total variance in a regional brain volume is explained by ICV (when using linear regression), and can potentially be removed by ICV normalization. It might seem unnecessary to remove as little as 10% of the total variance in the estimated volume, but it could have a large impact on a research study.

For example, let us assume that we want to compare the volume of nucleus accumbens between two samples. We expect that the mean volume in one of the samples is about 440 mm^3 with a standard deviation of 70 mm^3 (from Voevodskaya *et al.*[28]). In the other sample, we expect a similar standard deviation, but want to evaluate if there is a difference in mean volume between the two samples of 5% or more. Then, for a statistical power of 0.8 and using an independent samples t-test (with pooled variance), 160 participants would be needed in both samples. However, if just 10% of the variance in both samples are explained by ICV[28], the expected standard deviation after a successful normalization would roughly be 67 mm^3 (= $(70^2 * 0.9)^{0.5}$). With this smaller standard deviation (and assuming that the normalization would not affect the mean volumes), we would instead need 147 participants in each sample. After ICV normalization, we would thus need 26 participants less in total. Just for the MR examinations, we would be able to save 130,000 SEK (at a cost of 5000 SEK/examination). It would also save some discomfort for 26 individuals that the study otherwise could have brought them.

In research, it has been common to use one of three ICV normalization methods. These methods are 1) least-squares normalization[29,30], 2) inferred least-squares normalization[31,32], and 3) proportion normalization[33,34].

Using least-squares normalization, a simple linear regression is deployed with ICV as the independent variable and the brain estimates as the dependent variable. From this regression analysis, the regression coefficient is used to normalize the brain estimates. This is done using the function

$$b_{i,norm} = b_i - k(icv_i - \overline{icv})$$

Here $b_{i,norm}$ is the normalized brain estimate i, b_i the unnormalized estimate i, k the regression coefficient, icv_i the ICV from the same participant, and $\overline{icv}$ the mean ICV in the whole sample. A similar way of applying least-squares normalization is to analyze the residuals from the simple linear regression. One slight difference compared to using the above function is that the above function adjusts the residuals so that the mean of the brain estimates is unchanged by the normalization. Another slight difference is that when not using the above function, it is common to add further covariates to the regression analysis at once. However, I will refer to both these procedures as least-squares normalization.

Using inferred least-squares normalization, the same function is used as for least-squares normalization, but the regression coefficient is calculated from a subsample before normalizing the whole sample. This method is commonly preferred over least-squares normalization when it is believed that the phenomenon of interest is associated with ICV in some part of the sample (even if just by chance). The regression coefficient is calculated in a subsample where this association is believed to be absent or otherwise negligible. By doing so, one avoids the risk of reducing variance of interest during ICV normalization. Often, the regression coefficient is calculated using a sample of healthy controls before normalizing the whole sample.

Using proportion normalization, the brain estimates are simply divided by ICV. An advantage with proportion normalization over the least-squares methods is that it can be done for single individuals without needing a sample for which to calculate the regression coefficient. As mentioned by O'Brien *et al.*[35], the interpretation of proportion normalized brain estimates depends on the relation between the units of the numerator (the brain estimates) and the denominator (ICV). If both are measured in mm^3, the proportion normalized estimates will be unitless and could be interpreted as percentages of the intracranial volume. However, if the regional brain estimates are areas (mm^2) or thicknesses (mm), the proportion normalized estimates will have a unit of mm^{-1} or mm^{-2} respectively. These units are less easy to interpret. Using least-squares or inferred least-squares normalization, the unit of the brain estimates will remain the same after normalization.

Further, when using least-squares normalization, the interpretation of the normalized brain estimates is made as if ICV was constant between individuals. When using inferred least-squares normalization, the interpretation of the normalized brain estimates is as if ICV was constant between individuals if not for the phenomenon of interest. For proportion normalization, no such reservation needs to be made[35] and can probably only be legitimately made if there is a proportional relationship between the brain estimates and ICV.

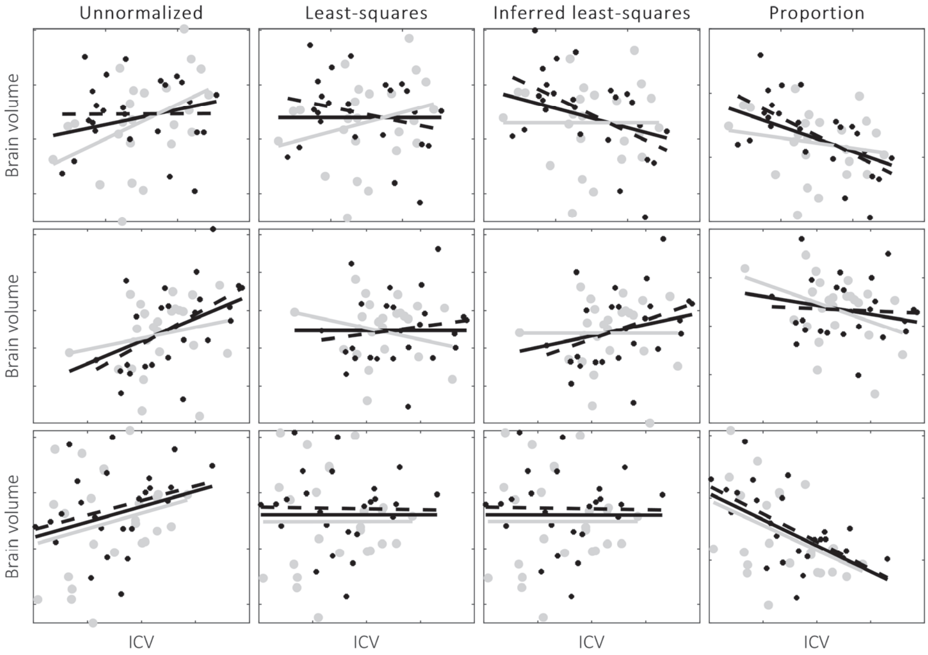


Figure 3. Examples of three different normalization approaches. In the left column are scatter plots of three different samples (one sample in each row) from simulated data. The x-axis shows the intracranial volume (ICV) of the participants in the samples and the y-axis a certain brain volume. The solid black line shows the association between ICV and the brain volume in the total sample. The slope of this line is the regression coefficient used during least-squares normalization. All three samples have been divided randomly into two subsamples (gray and black dots). The solid gray line shows the association seen between ICV and the brain volume in the gray subsample and the dashed black line the association seen in the black subsample. In this example, the slope of the solid gray line is the regression coefficient used during inferred-least squares normalization. As seen in the second column, the slope of the black line is zero after least-squares normalization. As seen in the third column, the slope of the gray line is zero after inferred least-squares normalization. In the fourth column, proportion normalization is used.

As exemplified in Figure 3 on the previous page, the effect of the different normalization approaches on brain estimates is quite complex. Many studies have therefore explored how the different ICV normalization approaches affect for example the linear association of the brain estimates to ICV[28], variance reduction[36], diagnostic accuracy[28] and reliability[37]. I will mention some of these studies in more detail in Section 5 (Discussion). In Paper IV, we try to describe the expected effect of the different ICV normalization approaches.

1.6 MANUAL ESTIMATION OF INTRACRANIAL VOLUME

The skull consists of three layers, namely the outer table, the diploë and the inner table. While the diploë, a porous layer containing red bone marrow, is easy to detect in T1-weighted MR images (as a bright layer) both the outer and the inner table are dark and indistinguishable from cerebrospinal fluid. This complicates the demarcation of the inner surface of the skull. Instead, the dura mater is used to trace this border whenever possible. The dura mater is closely attached to the skull and is often easy to detect in T1-weighted images as a white contour where the brain is separated from the skull by cerebrospinal fluid. When the brain is close to the skull the contour of the brain is demarcated instead since the dura mater cannot be distinguished from the brain tissue there. In Section 3.6.2, a sagittal MR image with the mentioned landmarks is displayed.

The estimation of ICV in MR images is mainly done using T1-weighted images even though it is easier to separate the skull from cerebrospinal fluid in T2-weighted images (and possibly in proton density weighted images too[38]). The reason for this is that T1-weighted images are almost exclusively used when segmenting regional brain volumes. By also estimating the ICV in the T1-weighted images, one avoids the inclusion of an extra MR acquisition during the MR examination. Whitwell *et al.*[39] also point out that by estimating the ICV on the same acquisition as the brain estimates, one avoids the risk that the ICV and the brain estimate will diverge due to different "image-acquisition factors".

It is fairly straightforward to segment the intracranial vault following the dura mater, but at some locations the segmentation becomes a bit ambiguous. One example is at the foramen magnum, an opening in the occipital bone through which the spinal cord passes. In a sagittal view, it can be hard to tell exactly where one should draw the line traversing the foramen magnum. By using guidelines for what to do at such locations, the segmentations will become more reliable and easier to replicate. Probably the most used guidelines for manual segmentation of the intracranial vault are those included in a study by Eritaia *et al.*[40] (described in Section 3.6.2). Other less common guidelines exist as well and new ones are often introduced too. In Table 2, I cite three different guidelines, two of which are used in more than one study. To my knowledge, there is no guideline published with the stated intention to be used as such by others. Rather, the guidelines are actually just descriptions of how the ICV segmentations were performed in the respective studies.

Manual segmentation of the whole intracranial vault is burdensome. Using the guidelines by Eritaia *et al.*[40] in MR images with 1 mm^3 voxels, one segmentation takes about 2.5 hours. To reduce the time needed, several less burdensome estimation methods have been developed. For example, Mathalon *et al.*[37] use a method where the height of the intracranial vault is estimated in an unspecified coronal MR image and an area of the intracranial vault (ICA) estimated in one transversal MR image (referred to as the index slice). The two estimates are then combined by the function 4/3*(height/2)*area to get an estimate of ICV. A similar method based on four ICAs is used in Eckerström *et al.*[41]. One ICA or an average of a few ICAs have also been used as estimates of ICV[42-44]. Further, it has been common to use head circumference as a proxy for premorbid brain volume[45,46].

Table 2. Guidelines for manual segmentation of intracranial volume

From	MR sequence	Orientation	Guidelines
Jenkins *et al.*[47]	1.5 T T2-w	Transversal	"The inner boundary of the calvarium, which includes the brain, meninges, and cerebrospinal fluid, was outlined…", "The inferior plane through brainstem was… [determined by] the level of the lowest slice that included cerebellar tissue."
Nordenskjöld *et al.*[48]	1.5 T PD-w	Transversal	"include all brain tissue and CSF [(cerebrospinal fluid)] inside the skull; include all dural sinuses; exclude the bilateral cavernous sinus and trigeminal cave; stop and do not include the brain stem when the occipital condyles are clearly visible"
Hansen *et al.*[36]	1.5 T T1-w	Transversal	"[Draw] along the outer surface of the dura mater using the lowest point of the cerebellum as the most inferior point. …no active exclusion of sinuses or large veins. The pituitary gland was excluded by drawing a straight line from the anterior-to-posterior upper pituitary stalk."

Examples of three different guidelines for ICV segmentation in the research literature. The sequences used were either T1-weighted (T1-w), T2-weighted (T2-w) or proton density weighted (PD-w).

Perhaps the most frequently used manual ICV estimation method is to segment the intracranial vault in every xth image and then multiply the total volume of the segmented slices by x[49-51]. By doing so, the time needed to segment an ICV will approximately be reduced by a factor x. For example, segmenting the intracranial vault in every 10th image would take about 15 minutes instead of 2.5 hours. Eritaia *et al.*[40] evaluated how the validity of such estimates depends on x. The evaluation was done for estimates based on segmenting every second sagittal MR image up to every 50th sagittal MR image. The conclusion was that estimates of ICV will be almost as good as a full segmentations of the intracranial vault if at least every 10th sagittal MR image is segmented (the intra-class correlation with absolute agreement between these estimates and full segmentations was >0.999). Since the publication by Eritaia *et al.*, estimates by every 10th sagittal MR image have even been used when evaluating ICV estimates using other methods[52,53].

A few studies[39,54-58] used some modified version of the method evaluated in Eritaia *et al.*[40]. For example, Whitwell *et al.*[39] used every 10th transversal ICA and linear interpolation to estimate ICV. They refer to Eritaia *et al.*[40] to justify their own ICV estimation approach. However, there are two important differences that make this justification questionable. First, Eritaia *et al.* evaluated the use of every 10th sagittal ICA, not transversal ones. There is less symmetry between transversal ICAs from the most superior to the most inferior point of the intracranial vault than there is between sagittal ICAs from one lateral point of the intracranial vault to the other. This could potentially make ICV estimates calculated from every 10th transversal ICA less valid than those calculated using every 10th sagittal ICA. Secondly, Eritaia *et al.* used nearest neighbor interpolation (also known as piecewise constant interpolation) and not piecewise linear interpolation. It is likely that linear interpolation is a better option than nearest neighbor interpolation. We investigate this in Paper I.

In Table 3 on the next page, I list a number of different manual ICV estimation methods that have been described in the research literature.

Table 3. Manual estimates of total intracranial volume

Method	n	Software	MR sequence	Correlation to whole ICV	Inter-rater	Intra-rater
Every 10th transversal ICA[56]	11	Analyze	1.5 T T1-w	–	–	0.965[b]
10 midcranial transversal ICA[a,56]	11	Analyze	1.5 T PD-w	–	–	0.997[b]
10 midcranial transversal ICA[56]	11	Analyze	1.5 T T2-w	–	–	0.999[b]
Every transversal ICA[a,56]	11	Analyze	1.5 T PD-w	–	–	0.998[b]
Every transversal ICA[a,56]	11	Analyze	1.5 T T2-w	–	–	0.994[b]
Every transversal ICA[48]	40	SmartPaint	1.5 T PD-w	–	0.999[c]	0.999[c]
Every transversal ICA[36]	10	ITK-SNAP	1.5 T T1-w	–	–	0.99[d]
Manually edited ICV from FSL[e,59]	10	FSL/ITK-SNAP	3 T T1-w + T2-w	–	>0.91[b]	>0.91[b]

The table continues on the next side.

One midsagittal ICA[44]	23/47[f]	MRIcro	1.5 T T1-w	0.89	0.97[b]	0.96[b]
2-4 midsagittal ICAs[44]	23	MRIcro	1.5 T T1-w	0.93-0.95	–	–
One midsagittal ICAs[42]	40/10[g]	Analyze	1.9 T T1-w	0.88	0.976 [b]	–

Examples of manual estimates and (for some) their Pearson's correlation to segmentations of the whole intracranial vault (whole ICV). Many of the methods reported are already whole ICV estimates since every intracranial area (ICA) was segmented. Intra- and interrater reliabilities are reported as Pearson's correlations if not otherwise noted. The sequences used were either T1- (T1-w), T2- (T2-w) or proton density weighted (PD-w). n is the number of MR acquisitions used. Midsagittal ICA: the ICA in sagittal orientation in the middle of the brain where the cerebral aqueduct is most prominent. When "2–4 midsagittal ICAs" is stated, the two, three or four sagittal ICA closest to the midsagittal plane are included.

[a] *semi-automatic approach*

[b] *intra-class correlation (possibly without absolute agreement)*

[c] *probably Pearson's correlation*

[d] *intra-class correlation with absolute agreement*

[e] *ICV segmentations were retrieved automatically by the software tool set FMRIB Software Library (FSL)[60] before being manually edited.*

[f] *23 MR acquisitions were used for the comparison to full segmentations, 47 MR acquisitions were used for the intra- and interrater reliability calculations*

[g] *40 MR acquisitions were used for the comparison to full segmentations, 10 MR acquisitions were used for the interrater reliability calculation*

1.7 AUTOMATIC ESTIMATION OF INTRACRANIAL VOLUME

As the dura mater (a thin bright but inconsistent line) is what guides the manual segmentations of the intracranial vault in T1-weighted MR images, it is not an easy task to create an automatic segmentation equivalent. Rather, automatic segmentation approaches have avoided the use of the dura mater. One common approach is to add together the estimated total brain volume and an estimate of the subarachnoid cerebrospinal fluid volume. This approach is often used via the tissue classification acquired when using SPM[48,56,61]. Another common way is to estimate the intracranial volume based on how the MR images are scaled in size when aligned to a head atlas (such an atlas is roughly speaking a volume of MR images from one [or multiple] head scans). This approach is for example used in FreeSurfer[48,56,62]. As it is hard to separate the skull from the cerebrospinal fluid, the first approach risks including parts of the skull and excluding cerebrospinal fluid that should have been included in the segmentation. The other approach risks being dependent on other things than the intracranial vault. What these other things might be depends on what mainly guides the alignment of the MR images to the head atlas.

In Table 4 on the next page, I present results from some comparisons between automatic and manual ICV estimation. Generally, ICV estimates from automatic methods tend to have a strong linear association to manual estimates of ICV. However, it is easy to achieve ICV estimates with rather high correlations to manual segmentations. Even if it estimates total brain volume rather than ICV, the Pearson's correlation to the manual segmentations will be about 0.9 (= the correlation between ICV and total brain volume[18]). Also, just by segmenting one ICA (compared to about 140 ICAs for a full segmentation of the intracranial vault), Pearson's correlations around 0.88–0.89 can be expected[42,43]. Therefore, a fair estimate of ICV should at least have a Pearson's correlation of 0.9 to thorough manual segmentations. Finally, depending on how the estimates are to be used, it is not necessarily enough to just have estimates with strong linear association to the actual ICV. A good volumetric agreement might also be necessary.

Hansen *et al.*[36] point out that it is easy to think naively that more accurate ICV estimates would automatically result in more effective ICV normalization. In other words, poor accuracy does not by default imply poor ICV normalization performance. In their study, Hansen *et al.* found that the least accurate method (eTIV [estimated total intracranial volume] from FreeSurfer) was the best at reducing variance in many regional brain volumes when using least-squares normalization. However, if the ICV estimate is not accurate, do we really normalize by ICV? In the example from Hansen *et al.*, does eTIV only reduce variance explained by ICV or does it also reduce variance due to something else (which the more accurate ICV estimates does not estimate)? We investigate this possibility in Paper III.

Table 4. Automatic compared to manual estimates of intracranial volume

Software	MR sequence	Manual reference	n	Pearson's correlation	Percentage error
FreeSurfer 4.5.0[a, 36]	1.5 T T1-w	Every transversal ICA	30	0.96[b]	7.3±3.7[c]
FreeSurfer 5.1.0[48]	1.5 T T1-w	Every PD-w transversal ICA	399	0.94	~5.9[d]
FreeSurfer 5.3[59]	3 T T1-w	Manually edited ICV from FSL	80	–	–6.9±11.0
FreeSurfer 5.3.0[63]	T1-w	Every transversal ICA	25	0.84	–2.3±7.8
FreeSurfer 5.3.0[57]	1.5 T T1-w	Every 10th transversal ICA	286	0.90	3.7±5.2
FreeSurfer[64]	T1-w	Semi-manual ICV estimation	20	0.95	5.9±3.2[c]
FSL (BET)[e 53]	3 T T1-w	Every 10th sagittal ICA	5	0.99[b]	0.5±2.4[c]
FSL (BET)[e 53]	1.5 T T1-w	Every 10th sagittal ICA	5	0.95[b]	–4.2±2.4[c]
FSL 5.0.4 (atlas scaling)[63]	T1-w	Every transversal ICA	25	0.92	–15.7±3.6
FSL (atlas scaling)[64]	T1-w	Semi-manual ICV estimation	20	0.92	–

The table continues on the next side.

NLSS[64]	T1-w	Semi-manual ICV estimation	20	0.99	1.4±1.5[c]
SPM5 (RBM)[e,f] [53]	3 T T1-w	Every 10th sagittal ICA	5	0.99[b]	0.4±1.7[c]
SPM5 (RBM)[e,f] [53]	1.5 T T1-w	Every 10th sagittal ICA	5	0.98[b]	–0.9±3.3[c]
SPM5 (sum tissue)[e,f] [53]	1.5 T T1-w	Every 10th sagittal ICA	5	0.88[b]	–3.0±8.8[c]
SPM5 (sum tissue)[e,f] [53]	3 T T1-w	Every 10th sagittal ICA	5	0.82[b]	9.5±2.0[c]
SPM8 (RBM)[36]	1.5 T T1-w+T2-w	Every T1-w transversal ICA	30	0.99[a]	0.1±1.7[a]
SPM8 (RBM)[36]	1.5 T T1-w	Every transversal ICA	30	0.97[a]	–0.6±2.7[a]
SPM8 (sum tissue)[63]	T1-w	Every transversal ICA	25	0.94	2.2±3.6
SPM8 (sum tissue)[48]	1.5 T T1-w	Every PD-w transversal ICA	399	0.86	~20.9[d]
SPM8 (sum tissue)[57]	1.5 T T1-w	Every 10th transversal ICA	288	0.76	13.9±8.3
SPM12 (sum tissue)[57]	1.5 T T1-w	Every 10th transversal ICA	288	0.97	–2.8±2.5
SPM12 (sum tissue)[64]	T1-w	Semi-manual ICV estimation	20	0.95	3.6±2.3[c]

Estimates of intracranial volume (ICV) from automatic methods compared to manual reference segmentations. The sequences used were either T1- (T1-w) or T2- (T2-w) weighted. FSL is a tool set for analysis of medical images[60]. BET in FSL estimates the inner surface of the skull. SPM is a tool set for analysis of medical images[61]. "sum tissue" implies that a sum of tissue classes is used. RBM estimates ICV through a rough brain mask created in SPM that is aligned to the head in native space[53]. NLSS[64] segments ICV as an image mask. ICA: intracranial area

[a] *I did not include results from modified methods*

[b] *intra-class correlations (possibly without absolute agreement)*

[c] *not ordinary percentage errors;*

[d] *calculated from mean values*

[e] *I only present the results from the case when "FAST" bias field correction was applied*

[f] *only with the default settings*

1.8 KNOWLEDGE GAPS

In year 2000, Eritaia *et al.*[40] evaluated the use of ICV estimates based on the segmentation of every x^{th} sagittal ICA. Many studies have since segmented every 10th sagittal ICA to estimate ICV, as proposed by Eritaia *et al.*. Estimates of every 10th sagittal ICA have even been used to validate other ICV estimation methods[52,53]. As the study by Eritaia *et al.* is so heavily relied on, it would be desirable to replicate it to confirm their results. Further, many variants of the proposed method have also been used over the years, for example segmenting every 10th transversal ICA. These variants needs to be evaluated too.

A number of automatic ICV estimation methods have been suggested over the years and many of them are found to have fair to good validity compared to manual segmentations. However, many of these methods differ in how ICV is estimated. Depending on the estimation approach, errors that we do not expect in manual estimates might be present in the automatic ones. The automatic methods should be evaluated with this in mind.

It is still unclear exactly how brain estimates are affected by the different normalization approaches. Knowledge about how the mean value, variance, and Pearson's correlation to ICV of regional brain estimates tend to be affected by normalization would help in a number of situations. Firstly, it may help settling which method is optimal in which settings. Secondly, it would increase our ability to interpret the effect of the different normalization methods. This would for example make it easier to compare findings from studies that use different normalization methods. Last but not least, we could include the probable effects of ICV normalization in our power calculations.

Some specific questions are raised by earlier findings. Arndt *et al.*[65] and Mathalon *et al.*[37] both showed that the reliability of brain estimates was reduced by ICV normalization. However, at least some of the reduced reliability seemed to be due to reduced true score variance. The question is how much the reduced reliability is due to an introduced estimation error and how much of it is due to such reduced true score variance. Does the answer differ between the normalization methods?

Sanfilipo *et al.*[66] found that least-squares normalization was not affected by ICV estimation errors, but they did not evaluate errors affecting the Pearson's correlation between the brain estimates and ICV. It is important to evaluate the effect of such errors too. If, indeed, least-squares normalization is not affected by estimation errors, then the reduced reliability in least-squares normalized brain estimates would solely be due to reduced true score variance.

Inferred least-squares normalization is applied when it is believed that the association between brain estimates and the phenomenon under study is to some degree explained by ICV. Therefore, after inferred least-squares normalization, we expect that some association will remain between the brain estimates and ICV, which has also been shown[28]. However, we do not know if the remaining association is related to the phenomenon under study or if it is association left by chance (or a combination of both). By not knowing which is true, it is hard to assess the value of an inferred least-squares normalization.

2 AIM

The general aim of this thesis is to validate methods for ICV estimation using MR images and to improve our understanding of the effects of ICV normalization in neuroimaging research.

2.1 SPECIFIC AIMS

Paper I – To validate a commonly used method for manual ICV estimation.

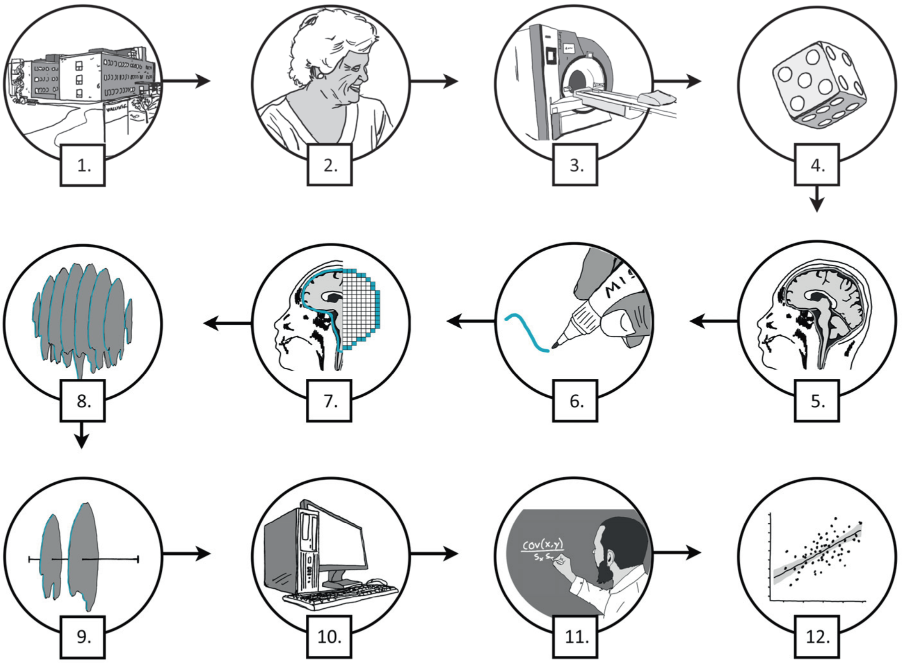
Paper II – To determine if the segmentation of one or two ICAs is enough to achieve adequate estimates of ICV.

Paper III – To show certain shortcomings in a commonly used method for automatic ICV estimation.

Paper IV – To describe how certain statistics (mean, variance, and Pearson's correlation to ICV) of brain estimates are affected by ICV normalization.

3 MATERIAL AND METHODS

In this section, I will describe the material and methods used in Papers I-IV. I will do so in twelve steps. In the first ten steps, I will describe the process of the three first studies. These studies all spring from the same manual estimates of ICV made from MR examinations from the Gothenburg MCI study. In the two last steps, I will describe the fourth study and the statistics used in all four studies. The twelve steps are illustrated below and include brief information about: 1) The Gothenburg MCI study. 2) The participants of the Gothenburg MCI study. 3) The MR examinations made. 4) The sample selection made. 5) The preprocessing of the MR images. 6) The manual segmentation of the intracranial vault. 7) Conversion of the manual segmentations into image masks. 8) The first study in which we evaluate the use of equidistant ICAs to estimate ICV. 9) The second study in which we evaluate the use of one or two ICAs to estimate ICV. 10) The third study in which we evaluate the use of FreeSurfer to estimate ICV. 11) The fourth study in which we present composite functions that describe the probable effect of ICV normalization on certain statistics. 12) A summary of the statistics used in all four reports.



3.1 THE GOTHENBURG MCI STUDY

The Gothenburg MCI study is a longitudinal study initiated in 1999 to learn more about the different phases of dementia diseases. The study includes patients from a memory clinic in Mölndal, Sweden, and healthy controls. Since its start, participants have been followed through five separate examinations over 10 years, first at inclusion and then 2, 4, 6, and 10 years after inclusion. The participants are examined using MR imaging, neuropsychological evaluation, cerebrospinal fluid sampling, and blood sampling. For a full description of the study design and a review of its previous findings, see Wallin *et al.*[67,68]. Some information about the study participants and the MR imaging is presented below.

The Gothenburg MCI study follows the declaration of Helsinki and the study is approved by the ethical review board in Gothenburg (diary number L091-99, 1999; T479-11, 2011).

3.2 STUDY PARTICIPANTS

3.2.1 HEALTHY CONTROLS

Healthy controls are mostly recruited through seniors' organizations, but a few are relatives to the patients included in the Gothenburg MCI study. To be included as a healthy control, the person must not have objective or subjective cognitive decline. A nurse establishes the cognitive status of the person before inclusion by the use of basic cognitive tests and by asking about subjective impairments. When there are any uncertainties, the nurse also consults medical doctors. Further, the persons must be between 50 and 79 years old to be included as a healthy control and have a mini-mental state examination[69] (MMSE, see Section 3.2.2) score above 26.

3.2.2 PATIENTS

Patients are included after being referred to the memory clinic in Mölndal, Sweden. The physicians at the clinic decide which patients should be asked to participate. Patients might for example be too ill to participate in such an extensive study. To be included, the patient must be 50–79 years old, have a MMSE score above 18, and have an ongoing cognitive decline since at least six months. The cognitive decline may either be self-reported or reported by informants.

The patients are classified into one of four groups following the global deterioration scale[70,71] (GDS). Patients classified as GDS 1 are regarded as having no cognitive impairment and are not included in the study. GDS 2 is described as subjective cognitive impairment and GDS 3 as mild cognitive impairment. In the original scale, GDS 4 means mild dementia and the scale then continues up to GDS 7 with more severe stages of dementia. In the Gothenburg MCI study, GDS 4 is used as the end stage. Thus, in Papers I–III, GDS 4 refers to dementia and not just to mild dementia. However, most of the patients classified as GDS 4 in the study have mild rather than more severe dementia.

Reisberg *et al.* give examples of characteristics that patients of each GDS stage may have[71]. A patient with GDS 3 may for example have problems with demanding work or with getting lost in new areas, while a patient with GDS 4 may have problems recalling recent events and handling finances. However, Reisberg *et al.* does not say how precisely to classify a patient to either of the stages. To be able to do the classification in a systematic way, the Gothenburg MCI study applies an in-house classification algorithm. Scores from four different tests that capture cognitive and daily living capacity are used in the algorithm. The tests are MMSE[69], investigation of flexibility (I-FLEX, a modified EXIT-test[72]), stepwise comparative status analysis[73] (STEP), and clinical dementia rating[74-76] (CDR).

MMSE was designed to assess cognitive function in patients with cognitive impairment. It consists of eleven tests that among other things include asking the patient to state where the examination is being held, to recall a few words, and to follow some simple instructions. EXIT was designed to assess executive cognitive function in patients that might have impaired such function. I-FLEX, which is a modified version of EXIT, consists of seven tests covering counting the number of instances of a given object in a picture, naming in one minute as many words as possible that begin with a given letter, and following simple instructions. STEP was designed to assess symptoms in mild to moderate dementia and to connect these symptoms to different brain syndromes that the patient might suffer from. STEP consists of 50 questions about the patient's symptoms that are each to be graded from zero to three by the rater. In the GDS algorithm, only the scores from question 13–20 in STEP are used. These questions cover among other things if the patient is able to remember objects that have been shown 5 minutes earlier, if she is able to mention similarities between two objects that are named (for example car and bicycle), and if she seems to have a reduced vocabulary and/or is talking slowly. CDR was designed to rate the stage of dementia by cognitive dysfunction. CDR consists of six cognitive and behavioral categories that are scored from zero to three in six steps. To determine the score of a given category, the rater follows certain guidelines for describing the state of the patient, but the guidelines are not explicit on how to make the assessments. The sum of the six scores are then used to assess the cognitive function of the patient[75].

Patients classified as GDS 4 (or above) are further classified as having either Alzheimer's disease, vascular dementia, mixed dementia, frontotemporal dementia, Lewy-body dementia, primary progressive aphasia, or dementia *non ultra descriptus*. The latter diagnosis is used when no specific dementia diagnosis is suitable. Mixed dementia is a diagnosis used when sufficiently many signs or symptoms of both Alzheimer's disease and vascular dementia are present. A physician performs the classification by taking into account the patient's medical history, clinical symptoms and cerebral white matter lesion burden. The different diagnostic criteria are given in[67].

3.2.3 EXCLUSION CRITERIA

Both healthy controls and patients are excluded from the Gothenburg MCI study if they have a systemic disease, somatic disease or psychiatric disorder that may affect their cognitive functioning. Alcohol or substance abuse, and confusion caused by drugs are further exclusion criteria.

3.3 MR EXAMINATION

Between 1999 and 2004, MR examinations were performed in the Gothenburg MCI study using a 0.5 T MR scanner. From 2005, 1.5 T scanners were used. With a few exceptions, the latter examinations were done using a Siemens (Healthineers, Erlangen, Germany) Symphony scanner. The MR examinations were done at Mölndal's Hospital, Mölndal, Sweden and took about 30 minutes/examination. Only examinations from the 1.5 T Symphony scanner were used in Papers I–III. A number of scanner sequences were available including a three-dimensional T1-weighted sequence, a two-dimensional FLAIR sequence, and a two-dimensional T2-weighted sequence. Of these, we only used the three-dimensional T1-weighted sequence. This was for three reasons. First, none of the other sequences includes the whole intracranial vault. Secondly, a T1-weighted sequence was used in the study replicated in Paper I. Thus, by using a T1-weighted sequence, any difference found between the original and our replication study should less likely be due to choice of image sequence. Lastly, the automatic method evaluated in Paper III requires a T1-weighted sequence.

The T1-weighted sequence we used was a magnetization-prepared, rapid gradient echo sequence. The acquisition parameters were: inversion time 820 ms; repetition time 1610 ms; echo time 2.38 ms; flip angle 15°; field of view 250 x 203 mm; matrix 512 x 416; acquisition pixel spacing 1.0 x 1.0 mm; reconstruction pixel spacing 0.49 x 0.49 mm; slice thickness 1 mm; spacing between slices 1 mm (no interslice gap); receiver bandwidth 220 Hz/pixel; number of slices 192; acquisition time 1.7–2.4 minutes; coil type body transmit. The examinations were performed between year 2005 and 2008. All MR images were anonymized.

3.4 SAMPLE SELECTION

The Gothenburg MCI study uses convenience sampling to include participants. In Paper I–III, a subsample was selected using stratified random sampling from the larger sample of participants in the Gothenburg MCI. This subsampling was made with two inclusion criteria. First, that 1.5 T MR acquisitions should be available. Secondly, that FreeSurfer results for these acquisitions should be available. Erik Olsson stratified the participants that met the inclusion criteria into healthy controls, patients with dementia, and other patients. After the stratification, Olsson made a random selection such that half of the participants would be healthy controls and half participants with dementia. Olsson did this selection so that I would be blinded to participant age, patient/control status, gender, and cognitive status when manually segmenting ICV.

The two inclusion criteria were chosen for two reasons. We chose the first criterion as we, in Paper I, were to replicate a study that used T1-weighted images. The second criterion was chosen to take advantage of the fact that we already had FreeSurfer segmentations on most T1-weighted images, segmentations that I already had corrected for gross errors. Using these segmentations, the evaluation of the estimation of ICV in FreeSurfer (in paper III) would need a bit less work.

When performing a replication study, it is recommended to have a larger sample than the original study[77,78]. The study by Eritaia *et al.*[40] was done on 30 normal controls. Therefore, we wanted to include at least 60 participants in our sample. Considering the risk of having to exclude participants later for various reasons, we settled for a primary sample size of 70 participants.

The inclusion of both healthy controls and participants with dementia was done in order to be able to evaluate if FreeSurfer's estimates of ICV may be affected by brain atrophy. While healthy elderly certainly will have some brain atrophy, a wider range of atrophy is secured by also including demented participants.

Unfortunately, only 32 healthy controls and 27 patients with dementia were available with 1.5 T T1-weighted MR images and with successful FreeSurfer analyses at the time. Therefore, 11 participants with subjective or mild cognitive impairment were included too.

3.4.1 PARTICIPANT DEMOGRAPHICS

Eight of the 70 participants were eventually excluded, as the whole intracranial vault was not captured in their MR images. The demographics of the remaining participants are presented in Table 5. The demographics of the excluded participants are presented in Table 6.

When comparing the remaining and excluded participants, there was no statistically significant difference in age, education or MMSE. However, there was a significant difference in gender. A majority of the remaining participants were females while most of the excluded participants were males. This difference could be due to the exclusion criterion that the whole intracranial vault is not covered in the T1-weighted MR images. As males have larger intracranial vaults and the image matrix has a given size, there is an increased risk for males not to get the whole intracranial vault covered.

Due to the small number of excluded participants, we must consider the risk of type II errors (not rejecting the null hypothesis when it is false) in the above comparisons. However, besides the gender inequality, we did not see any further reasons to suspect that there actually are differences also in age, education or MMSE between the remaining and excluded participants.

Among the remaining 25 GDS 4 patients, ten were diagnosed as having Alzheimer's disease. Further, six had dementia *non ultra descriptus*, two had mild cognitive impairment (rather than dementia), two had mixed dementia, two had frontotemporal dementia, one had vascular dementia, and one had primary progressive aphasia. Finally, one patient did not retrieve a dementia diagnosis due to complex medical history (and was at later examinations classified as GDS 3 and 2). Among the excluded participants, one of the GDS 4 participants had vascular dementia and the other dementia *non ultra descriptus*.

Of the remaining 62 participants, 26 had been examined at inclusion (year 0), 22 at year two, 12 at year four, and two at year six. Of the excluded participants, four had been examined at inclusion, one at year two, and three at year four. No participant had more than one examination included in the sample.

Table 5. Study demographics for remaining participants

Group belonging	n	Gender (m/f)	Age	Education	MMSE
All participants	62	23/39	66.1±8.0	11.0 (6.0,23.0)	28.5 (16,30)
Healthy controls	29	8/21	66.4±7.5	11.5 (7.0,15.0)	30 (27,30)
Patients, GDS 2–3	8	4/4	66.7±8.2	12.0 (6.5, 20)	28.5 (26, 29)
Patients, GDS 4	25	11/14	65.5±8.8	10.0 (6.0, 23.0)	25 (16,30)

Age is presented as mean age in years and standard deviation. Education in years and mini-mental state examination (MMSE) score are presented as medians followed by minimum and maximum values. GDS = global deterioration scale; n = number of participants; m = number of males; f = number of females.

Table 6. Study demographics for excluded participants

Group belonging	n	Gender (m/f)	Age	Education	MMSE
All participants	8	6/2	66.1±9.7	11.5 (8.0,18.0)	28 (22, 30)
Healthy controls	3	2/1	70.0±10.8	12.0 (11.0,14.0)	28 (28,30)
Patients, GDS 2–3	3	2/1	59.9±10.5	11.0 (11.0,18.0)	29 (28,29)
Patients, GDS 4	2	2/0	69.6±3.6	10.0 (8.0,12.0)	25 (22,28)

Age is presented as mean age in years and standard deviation. Education in years and mini-mental state examination (MMSE) score are presented as medians followed by minimum and maximum values. GDS = global deterioration scale; n = number of participants; m = number of males; f = number of females.

3.5 IMAGE PREPROCESSING

Before starting the segmentation of the ICV, the T1-weighted images were pre-processed. The pre-processing consisted of a reformatting of the voxels and intensity adjustments.

The first step of the pre-processing was a reformatting of the voxels from 0.49x0.49x1 mm voxels to cubic 1 mm^3 ones. In the expression for voxel size, 0.49x0.49 is the pixel spacings and x1 the spacing between slices. The reformatting was done using the MATLAB (Mathworks, Natick, MA, USA) function *interp3* with linear interpolation. We did the reformatting for three reasons. 1) To get an image resolution similar to that of the study by Eritaia *et al.*[40] where the voxel dimensions were 0.938x0.938x1.5 mm. 2) To be able to talk about the position in the MR images in mm and in slices interchangeably. This was useful as Eritaia *et al.* talked about the position in slices while we figured it would be less ambiguous to talk about it in mm. A segmentation of every 10th slice will give different reliabilities depending on whether the slice thickness is 1 mm or 2 mm. 3) To reduce the number of ICAs to segment. Without the reformatting, an MR acquisition would require about 5 hours of segmentation instead of 2.5 hours (which was the case now). Linear interpolation was chosen to reformat the voxels as it is easy to understand and as we thought that intensity errors in a few voxels due to the interpolation would have only a small impact on the segmentation of the whole intracranial vault (where about 1500 000 voxels are included). With knowledge that we have since acquired about the pros and cons of the different interpolation methods, we would probably have used a more advanced interpolation method. For the visualization of the images on the computer screen, we used cubic interpolation, which is probably a better option.

After the reformatting of the voxels, each MR acquisition was intensity adjusted so that its mean intensity was close to the mean intensity of all MR acquisitions. The 10% darkest and brightest voxels were not included when calculating the means. The mean intensity would otherwise vary between MR acquisitions depending on the size of the participants' heads (the dark background occupies more space the smaller the head). The adjustment was done by first finding the maximum intensity in each MR acquisition (Max_i). We then multiplied Max_i with 0.1 and 0.9 (we assumed that the intensity range is continuous) to get the 10th and 90th percentiles of the intensity range. The mean of the intensity range within the 10th and 90th percentiles was then calculated (=$\overline{Max_{\iota,10-90}}$). An equivalent mean was calculated for the intensities within these percentiles in all MR acquisitions (=$\overline{Max_{All,10-90}}$). Lastly, the intensity of each voxel was multiplied by $\frac{\overline{Max_{\iota,10-90}}}{Max_{All,10-90}}$. We did this intensity adjustment to reduce the risk of getting segmentations that varied because of differences in the overall brightness of the MR acquisitions. The effect of this intensity adjustment is illustrated in Figure 4 on the next page.

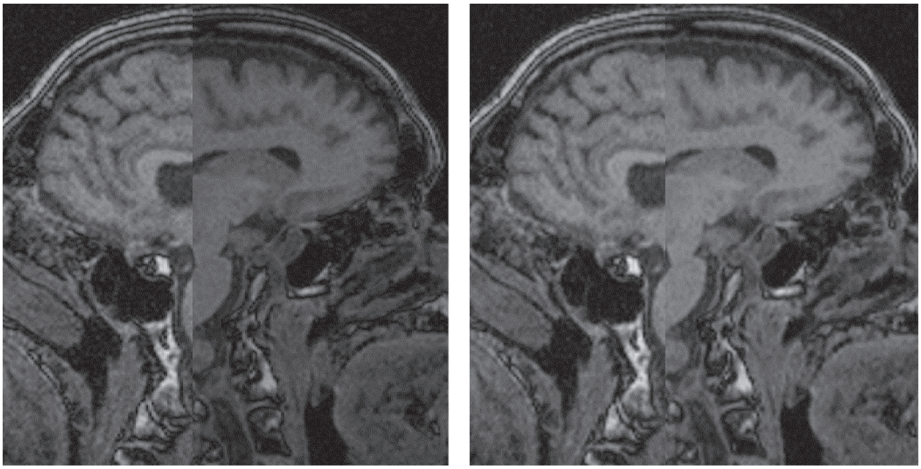


Figure 4. Intensity adjustment. The left half of both MR images is from participant A and the right half from participant B. The left image shows the original image brightnesses and the right image the brightnesses after the initial intensity adjustment.

A second intensity adjustment was done by manipulating the color map of the images. The intensities of grayscale images can be seen as indices to a given palette of colors. The palette of colors is also called a color map. The T1-weighted images used in Papers I–III had a range of intensities with values between 0 and 4095. Normally, the value 0 is used as an index for black and the highest value (in this case 4095) as an index for white. The intensity values between 0 and 4095 are then indices for different shades of gray. By manipulating the color map, the intensity values of the original image is kept intact while their visualization/interpretation is changed. In Figure 5a, I give an example where I have altered the color map so that the intensity values of the image become indices for a range of colors (and not just shades of gray). The actual adjustment of the color map for the study was made to make the dura mater easier to detect. This was done by altering the color map so that the intensity values of the 10% of the voxels with smallest and largest intensity values were set to indices for black and white respectively. The indices of the remaining 80% of the voxels were then linearly distributed to shades of gray (Figure 5c). By this change, the contrast between these intensities will increase, which makes them easier to tell apart from each other. While the intensity values in the upper and lower 10% of the voxels become indistinguishable, I judged this a tolerable loss of information when segmenting ICV.

The color map was also gamma corrected (Figure 5d). A gamma correction transforms the relationship between the brightness of the shade of gray and the indices from a linear to a non-linear relationship. We chose to use the gamma function $y = 12 * \left(\frac{x}{12}\right)^{0.8}$, where y is the new shade of gray, 12 is the bit depth of the MR images, x is the previous shade of gray, and 0.8 is the gamma value. When the gamma value is less than one, the overall brightness of the image will increase. The gamma correction was made to make it easier to tell the dark cerebrospinal fluid apart from dark (dense) bone. It made the noise in the cerebrospinal fluid become more distinct relative to the fairly noise free bone.

As the adjustments to the color map do not affect the actual image data, I had the possibility to switch between the original color map and the adjusted color map during the segmentation of the intracranial vault. This helped me to do more well-informed choices in ambiguous areas. I performed the pre-processing steps using the software MIST (see 3.6.1).

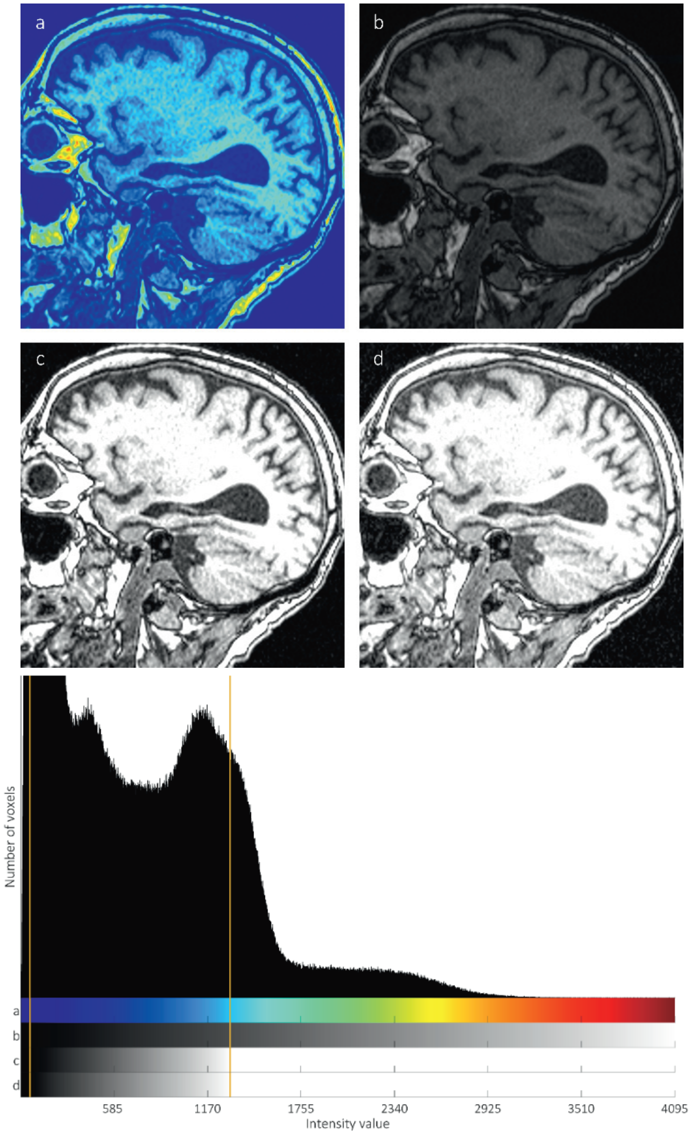


Figure 5. Color map adjustment. An MR image with four different color maps and a histogram of the number of voxels in the whole MR acquisition as a function of the voxel intensities (0–4095). The color maps are: a) intensities mapped to colors. b) intensities mapped to matching shades of gray. c) map b after compressing the 10% darkest and brightest voxels (outside of the orange lines). d) map b with the compression and the gamma correction (the color map that we used).

3.6 MANUAL SEGMENTATION

The whole intracranial vault was segmented on 62 MR acquisitions. The segmentations were done by me and took about 2.5 hours per acquisition. On average, 136 sagittal images were segmented for each acquisition. The segmentation of one image constitutes a border that essentially follows the inner surface of the intracranial vault. The border encircles an area, which we refer to as an intracranial area (ICA). To calculate the ICV in the simplest possible way, one has just to multiply the sum of all ICAs by the image thickness. During all segmentations, including intra- and inter-rater segmentations, the raters had no knowledge about the participants (for example no information about age, gender, or cognitive status). In the following subsections, I will describe the segmentation of ICV in more detail. The volume of the manual segmentations of the whole intracranial vault will be referred to as the gold standard ICV in comparison to other methods.

3.6.1 SEGMENTATION TOOL

The segmentations of the whole intracranial vault were made using MIST (Medical Image Segmentation Tool). MIST is a tool for manual segmentation of MR images that I developed with input from Erik Olsson. The development of MIST started as updates of a previous, similar software called Hipposegm, first created by Magnus Borga, that Helge Malmgren’s research team in Gothenburg used between 2000 and 2013. MIST is written in MATLAB and includes features such as image rotation, brightness and contrast adjustment, and three-dimensional visualization of segmentations. The development of the software has been on hold for some time, but I might come to publish the source code in the future. Until now, only Erik Olsson and I have been using MIST in research. Erik Olsson and Carl Eckerström have also used the earlier software, Hipposegm. MIST is used in[79] and in Paper I, while Hipposegm is used in[9,10,41,80]. We chose to use MIST as segmentation tool as we are familiar with it, it is user friendly and we can adapt it depending on the purpose. Most manual segmentation tools should be able to reach a similar accuracy in the segmentations, but the effort and time needed to do so might vary between these tools. With MIST we knew that high accuracy was reachable without too much effort.

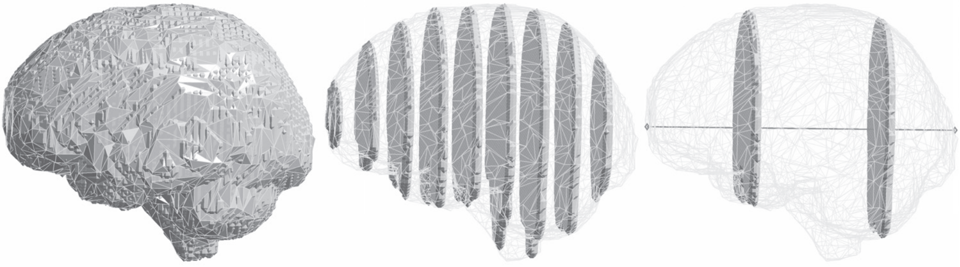


Figure 6. The left image shows the intracranial vault (ICV) segmented on every intracranial area (ICA) in a low-resolution magnetic resonance acquisition. The middle image shows the ICV segmented on every 10th ICA and the right image two ICA with the perpendicular diameter of the intracranial vault.

When segmenting the intracranial vaults, I used a Wacom DTU-2231 display. On the screen, each image was visualized in sagittal orientation and scaled to a quarter of its true area. That is, one mm on the screen represents 2 mm in the imaged space. The screen interpolation was done using the *interp3* function in MATLAB. We used cubic interpolation that gives a smoother feel to the images compared to when using nearest neighbor or linear interpolation. The scaling in size when visualizing the images was chosen to reduce the time needed for the segmentation while keeping the precision of the segmentations at a high level. The segmentations were done at screen resolution, which is higher than the image resolution. Segmentation at screen resolution makes it easier to segment areas as one wishes. It also allows for an indirect way to correct for partial volumes effects (tissues with different intensities within one voxel making the voxel intensity a mixture of the actual tissue intensities) in the MR images, as fractions of voxels can be included in the segmentation.

3.6.2 SEGMENTATION PROTOCOL

When segmenting the whole intracranial vault, we followed the guidelines described by Eritaia *et al.*[40]. To our knowledge, these guidelines are the most frequently used guidelines for manual ICV segmentation in MR images (151 citations according to Scopus 2018-10-29). As we were to replicate the study by Eritaia *et al.* in Paper I, it was also the obvious choice of guidelines for us. The guidelines consists of a visualization of a midsagittal ICA (the intracranial area in sagittal orientation where the cerebral aqueduct is most prominent) where a few landmarks are pointed out. I replicate this visualization in Figure 7 on the next page. The most useful landmarks for the segmentation are the dura mater, the cerebral contour, the undersurface of the frontal lobe, the dorsum sellae, clivus, and the posterior and anterior arch of the atlas. Further landmarks that Eritaia *et al.* mention are the scalp, the diploë, and the outer and inner table of the skull. When doing the segmentations, I start at the midsagittal ICA and then continue laterally. In the preparation for the segmentations, I made a pictorial guide over Eritaia's guidelines including a few other landmarks, such as the cerebral aqueduct. The pictorial guide is included in Appendix A.

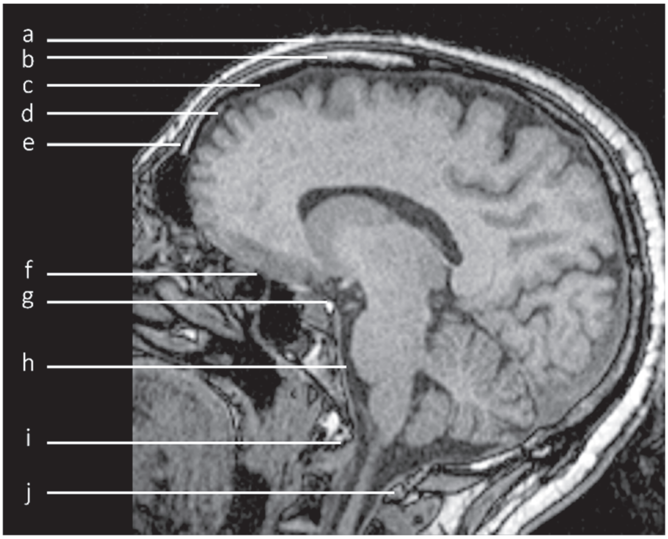


Figure 7. Segmentation landmarks. a) scalp, b) diploë, c) inner table of the skull, d) dura mater, e) outer table of the skull, f) undersurface of the frontal lobe, g) dorsum sellae, h) clivus, i) anterior arch of the atlas, j) posterior arch of the atlas.

3.6.3 INTRA- AND INTER-RATER RELIABILITIES

Three intra- and inter-rater reliabilities were assessed in Paper I and II. In Paper I, we assessed the rater reliabilities when segmenting every 10th ICA and every 40th ICA respectively. In Paper II, we assessed the rater reliabilities when only segmenting one midsagittal ICA. We did not assess the rater reliabilities when segmenting the whole intracranial vault, as it would take more time than we had at our disposal. We also figured that if the rater reliabilities when using every 10th ICA were high, they should be high when segmenting the whole vault as well. We thought so for two reasons. First, the more areas that are segmented, the likelier it is that the average error is not due to chance. Secondly, odd images that might be hard to segment will have less impact on the estimate when segmenting more areas. For the same reasons, there also is a risk that the reliability drops when one segments fewer than every 10th ICA.

That is why we wanted to evaluate the rater reliability for every 40th ICA and for only one midsagittal ICA.

I did the intra-rater segmentations six months after the initial segmentations. Before beginning, I randomly divided the 62 MR acquisitions into two equal batches. On one batch, I was to segment every 10th ICA and on the other batch every 40th ICA. For the segmentations, I used the already preprocessed MR images. The rater reliability will not differ between using already preprocessed images and redoing the preprocessing. Once it is decided what preprocessing to do, it is performed just by running a MATLAB script. The segmentations of every 10th ICA took about 13.5 minutes and every 40th ICA about 4.5 minutes.

I started the segmentations of every 10th and 40th ICA on the midsagittal ICA. Thus, I was able to use these midsagittal ICAs to assess the intra-rater reliability when only segmenting one midsagittal ICA (which was done in Paper II).

Inter-rater segmentations were done by Simon Skau. Before this, Skau had no previous experience of MR image segmentation. To start with, Skau was trained by first introducing the pictorial guide for ICV estimation and by making him familiar with MIST. Skau got a few MR acquisitions that were not included in the study to practice with at home. As an evaluation, I also let Skau segment the intracranial vault on four MR acquisitions that were not included in the study sample. Skau was to segment the whole intracranial vault on one of these acquisitions, every second ICA on another acquisition, and every 10th ICA on the remaining two acquisitions. Afterwards, I examined the segmentations for errors and discussed places with Skau where I would have done it differently as well as steps that he thought were ambiguous. As Skau's segmentations were of high quality, Skau started to segment the MR acquisitions included in the sample. The segmentations were done on already preprocessed images and in the same batches that I used. The segmentations of every 10th ICA took Skau about 22.5 minutes/acquisition. The segmentations of every 40th ICA took him about five minutes/acquisition.

Skau was instructed to begin the segmentations at the midsagittal area. Thus, I was able to use these midsagittal ICAs to assess the inter-rater reliability when only segmenting one midsagittal ICA (which was done in Paper II).

Intra- and inter-rater reliabilities were calculated by comparing the new segmentations to the same subsamples of ICAs from the initial segmentations of the whole intracranial vaults.

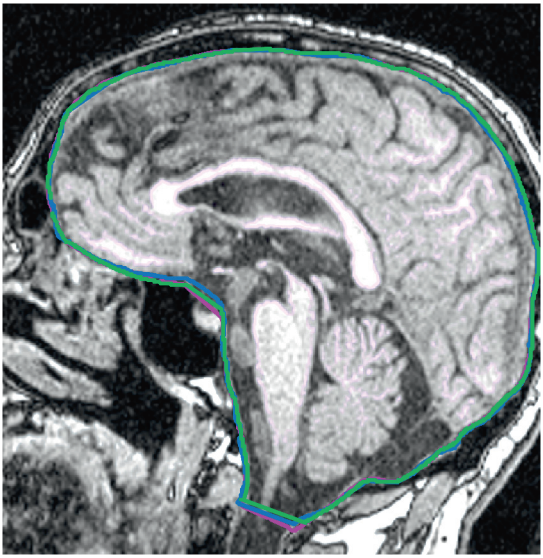


Figure 8. Example of the original segmentation of a midsagittal intracranial area (green) with the intra- (pink) and inter-rater (blue) segmentations overlaid.

3.7 IMAGE MASKS

The segmentations of the whole intracranial vault were reconstructed into binary image masks. These masks were used in Papers I–II as they enabled us to divide the whole segmentation into coronal, sagittal or transversal ICAs. Doing so, we were able to evaluate, not only different sets of ICAs to estimate ICV, but also how the orientation of these ICAs affects the estimation.

A binary image mask is a binary matrix with the same size as the original image. A one in the matrix tells that the object of interest is present in the given voxel. A zero tells that the object is not present in the given voxel. The segmentations were reconstructed into image masks using the MATLAB function *inpolygon* for each ICA. *Inpolygon* finds all voxels inside the border of the segmented object of interest. The found voxels are set to ones in the binary image mask. The process continues until all ICAs have been gone through. By doing this reconstruction, the resolution of the segmentation is lowered to the resolution of the images. However, the absolute percentage errors of the image masks compared to the actual segmentations were very small. The average absolute percentage error was 0.07%. In Figure 9, a segmentation of an ICA and its reconstructed image mask is shown.

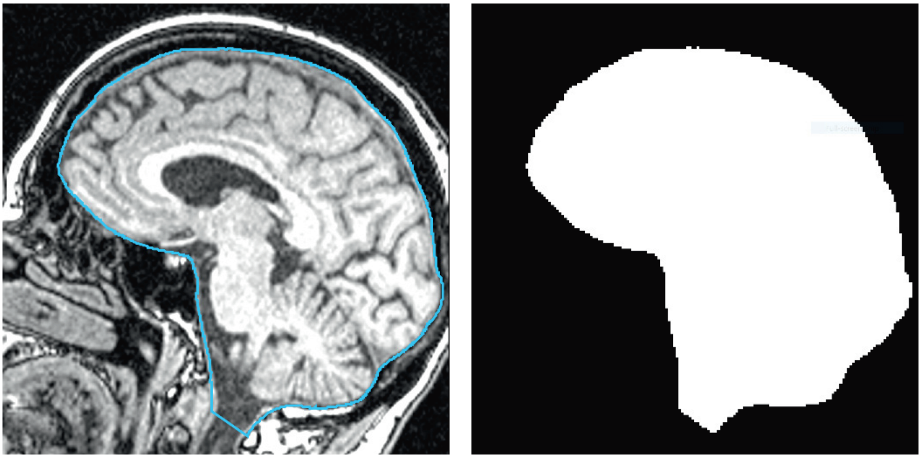


Figure 9. A segmented intracranial area (left image) and the binary image mask of this segmentation (right).

3.8 PAPER I

In Paper I, we evaluate whether estimates based on every second up to every 50th ICA are enough to get valid estimates of ICV when using MR images with $1\ \text{mm}^3$ voxels. We designed the study to replicate a study by Eritaia *et al.*[40], but with two additions. In contrast to the study by Eritaia *et al.*, who only evaluated the use of sagittal ICAs, we also added an evaluation of the use of coronal and transversal ICAs. Further, Eritaia *et al.* used a nearest neighbor interpolation in order to get a volume estimate from the equidistant ICAs. In our study we also investigated whether better estimates could be achieved with either piecewise linear interpolation or a cubic spline interpolation. With fewer ICAs, the choice of interpolation method could have a great impact on the validity of the estimates as more information has to be estimated through interpolation.

3.9 PAPER II

In Paper II, we evaluate whether it is enough to segment one or two ICAs to get valid estimates of ICV. We did this evaluation using four different methods. 1) Using only the midsagittal ICA as an estimate of ICV. To use one ICA as an estimate of ICV has already been suggested by Ferguson *et al.*[42] and Nandigam *et al.*[44]. 2) Using one ICA multiplied by the intracranial width perpendicular to the ICA. 3) Using the sum of two ICAs with the same orientation multiplied by the intracranial width perpendicular to the ICAs. 4) Using two ICAs and a shape-preserving piecewise cubic interpolation. Except for the first case, it is not given where in the intracranial vault to segment the ICAs. Therefore, for methods 2–4, the 62 MR acquisitions were randomly divided into a training and an evaluation set. Using the training set, the validity when using each possible combination of ICAs was evaluated. The combination that resulted in the most valid estimate (according to certain chosen criteria) in the training set had its validity recalculated using the evaluation set.

In Paper I, when segmenting every 50th ICA, about 2–3 ICAs were segmented at total to estimate an ICV. In Paper II, we thus continued the evaluation of estimating ICV with a small number of ICAs. An important difference is that we dropped the requirement of equidistant ICAs. That is, we did not require that there is a distance of 50 mm between the ICAs.

3.10 PAPER III

In Paper III, we examine FreeSurfer's[81] estimate of ICV (eTIV[62], estimated Total Intracranial Volume). The calculation of eTIV is based on a method proposed by Buckner *et al.*[52]. Using this method, ICV is approximated by how much the MR images are scaled in order to align them to a head atlas. Roughly speaking, a head atlas is a set of MR images of one (or multiple) MR acquisition(s) that cover the head.

As eTIV is calculated based on the alignment of the head in the MR images to a head atlas, there is no underlying segmentation. Further, as this alignment is done using information about the whole head[82], the variability in eTIV will depend on more than just the intracranial vault. As the brain constitutes a large part of the head, it is likely to affect the alignment and therefore eTIV. With such a dependence, eTIV risks becoming biased by total brain volume and possibly by global atrophy. Such a bias could have negative effects in for example dementia research, where we expect global atrophy to occur. To evaluate this risk in Paper III, we used the gold standard ICV, eTIV from FreeSurfer, and estimates of the total brain volume from FreeSurfer. While it would probably have been even better to use manual estimates of total brain volume, total brain volume estimation in FreeSurfer is based on a more rigorous method than eTIV; the brain is actually segmented in the MR images. For this reason, it was also possible to do some error correction to these segmentations to improve the estimated total brain volume.

When planning for the study, we had FreeSurfer analyses available from FreeSurfer version 5.1.0. During my work with the study, FreeSurfer was updated, so I decided to rerun the analyses. In Paper III, FreeSurfer version 6.0.0 was used on a MacPro 3.1 with two quad-core Intel Xeon processors and Mac OS X version 10.8.5. Both eTIV and total brain volumes were obtained from the aseg.stats files that are generated during the Freesurfer analyze. There are two options for the choice of total brain volume estimate in the aseg.stats file. We chose to use the output BrainSegVol as the estimate of total brain volume. BrainSegVol includes the cerebrum, the cerebellum, all ventricles, some cerebrospinal fluid, the optic chiasm, and vessels, but not the brain stem or the dura mater[83]. The other option is BrainSegVolNotVent, which will not include the ventricles, cerebrospinal fluid or choroid plexus[83]. BrainSegVol was used over BrainSegVolNotVent as we thought that a shrinkage of the outer perimeter of the brain relative to the intracranial surface would have the most influence on eTIV. BrainSegVolNotVent will also be affected by ventricular enlargement and thus relatively less by cortical atrophy.

3.11 PAPER IV

In Paper IV, which is a manuscript, we introduce mathematical functions that predict the mean, variance, and Pearson's correlation to ICV of brain estimates normalized by ICV. As arguments, the functions take statistical properties of the brain estimates and the ICV prior to normalization, such as mean and variance. The functions differ depending on the normalization approach. In Paper IV, we focused on three common normalization approaches: 1) least-squares normalization, 2) inferred least-squares normalization, and 3) proportion normalization. The mathematical functions are presented in Table 7 on page 51.

Besides introducing the mathematical functions, we also evaluated their use in predicting data from a simulation and from two previous studies that also evaluated the effects of ICV normalization. Our evaluation served two purposes. First, to show that the functions actually predict the mean, variance, and Pearson's correlation to ICV after normalization. Secondly, to get a feeling of what kind of prediction uncertainty and errors to expect when using the functions. The two studies, the results of which we predict, were chosen as they report all values needed for making the predictions and for comparing them to the actual results.

The simulation was done by creating a family of one million bivariate normal distributions that describe hypothetical populations with two given variates, a regional brain volume and ICV. Each bivariate distribution was created by randomly determining the mean and variance of the two variates, as well as the Pearson’s correlation between the variates. The randomness was constrained so that the two variates would resemble a brain estimate and ICV to some degree. Then, from each bivariate distribution, two samples were drawn, each with a random size between 10 and 1000. For the whole family of such sample pairs, we first predicted what would happen when normalizing the first variate by the other, and then evaluated what actually happened. In Figure 10, I have visualized an example of a bivariate normal distribution and a sample drawn from it.

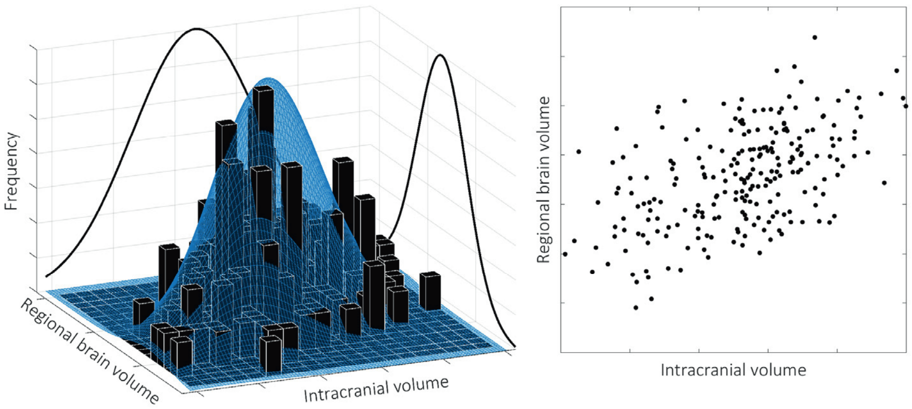


Figure 10. Randomly created bivariate normal distribution. The black lines in the left graph shows two normal distributions where mean, variance and Pearson’s correlation have been set randomly but constrained to resemble the distributions of a regional brain volume and intracranial volume in a population. Together these two distributions form a bivariate normal distribution, also illustrated as a blue surface. From the bivariate normal distribution, a sample of 232 cases has been drawn randomly. The frequency distribution of the sample is illustrated by the bars. In the right graph, the regional brain volume in the sample is plotted as a function of the intracranial volume.

Table 7. Composite functions from Paper IV

Normalization	$\overline{b_{norm}}$	$s^2_{b_{norm}}$	$r_{b_{norm},icv}$
Least-squares	$\bar{b}_1$	$s^2_{b_1} - r^2_{b_1,icv_1} s^2_{b_1}$	0
Inferred least-squares	$\bar{b}_2 - \frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}} (\overline{icv_2} - \overline{icv_1})$	$s^2_{b_2} + \frac{r^2_{b_1,icv_1} s^2_{b_1} s^2_{icv_2}}{s^2_{icv_1}}$ $- \frac{2 r_{b_1,icv_1} r_{b_2,icv_2} s_{b_2} s_{b_1} s_{icv_2}}{s_{icv_1}}$	$\frac{\left(r_{b_2,icv_2} s_{b_2} s_{icv_1} - r_{b_1,icv_1} s_{b_1} s_{icv_2}\right)}{s_{icv_1} \left(s^2_{b_2} + \frac{r^2_{b_1,icv_1} s^2_{b_1} s^2_{icv_2}}{s^2_{icv_1}} - \frac{2 r_{b_1,icv_1} r_{b_2,icv_2} s_{b_2} s_{b_1} s_{icv_2}}{s_{icv_1}} \right)}$
Inferred least-squares (CI)	$\bar{b}_1 \pm \frac{(0.5+z)(1+r) s_{b_1}}{\sqrt{n_1}}^{\dagger}$	$0 \text{ to } s^2_{b_1} \left(1 - r^2_{b_1,icv_1} + 1.5z \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}} \right)^{\dagger}$	$\pm(1+z)\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}^{\dagger}$
Proportion	$\frac{\bar{b}}{\overline{icv}} - \frac{r_{b,icv} s_b s_{icv}}{\overline{icv}^2} + \frac{\bar{b}^2 s^2_{icv}}{\overline{icv}^3}$	$\frac{\overline{icv}^2 s^2_b + \bar{b}^2 s^2_{icv} - 2\bar{b}\,\overline{icv}\, r_{b,icv} s_b s_{icv}}{\overline{icv}^4}$	$\frac{r_{b,icv} C_b C_{icv} - C^2_{icv}}{C_{icv} \sqrt{C^2_b + C^2_{icv} - 2 r_{b,icv} C_b C_{icv}}}$

Composite functions of the mean ($\overline{b_{norm}}$), variance ($s^2_{b_{norm}}$), and Pearson's correlation to intracranial volume ($r_{b_{norm},icv}$) with functions for intracranial volume (icv) normalization. b_1 is the brain measurements from the sample used to calculate the regression coefficient and b_2 those from a second sample. icv_1 is the icv from the sample used to calculate the regression coefficient and icv_2 those from a second sample. s_b is the standard deviation of b. s_{icv} is the standard deviation of icv. $r_{b,icv}$ is the Pearson's correlation coefficient between b and icv. C_b is the coefficient of variation of b. C_{icv} is the coefficient of variation of icv. z is a z value chosen from a standard normal distribution to determine the size of the confidence interval (CI). n_1 and n_2 are the sample sizes of the two samples used during inferred least-squares normalization. †The composite function expressed as the lower and upper bound of a CI. The probability of the CI is given by the area left of the z value. The function is not mathematical proven, but found through fitting an exponential model to simulated data.

Proofs for the composite functions in Table 7 that give point estimates in the cases of least-squares and inferred least-squares normalization are presented in Appendix B. The composite functions for inferred least-squares normalization that give confidence intervals were derived by finding functions that best explained simulated data. These functions are only preliminary. The composite functions for proportion normalization were retrieved from[84-86].

3.12 STATISTICS

In Paper I–IV, we used a number of different statistical methods. These were 1) chi-square test of independence, 2) Mann-Whitney U test, 3) t-tests, 4) Kruskal-Wallis test, 5) Pearson's correlation, 6) partial correlation, 7) intra-class correlation, 8) Jaccard index, 9) confidence intervals for differences between Pearson's correlations, and 10) delete-x Jackknife resampling. I will describe all these things in short before continuing with our use of them.

1) The chi-square test of independence is a non-parametric test with regard to frequencies of observations when classified by two categorical variables. The test is used to evaluate if the two variables are independent. The null hypothesis is that the two variables are independent.
2) The Mann-Whitney U test is a non-parametric test used to test if two independent samples come from populations with different values of a given property. The null hypothesis is that the samples come from populations with exactly the same values of the given property.
3) The t-test is a parametric test that can be used both to test if two independent samples come from populations where the mean values of a given property differ (independent samples t-test)

and if the mean values of two measures differ in the same population (paired samples t-test). The null hypothesis in the independent samples t-test is that the samples are from populations with the same mean value. The null hypothesis in the paired samples t-test is that the difference in mean value is zero between the two measures in the population.

4) The **Kruskal-Wallis test** is a non-parametric test to test if the values in multiple independent samples come from populations with different values of a given property. The null hypothesis is that the samples come from populations with exactly the same values of the given property.
5) The **Pearson's correlation** is a measure of linear association between two continuous variables. The Pearson's correlation can be used in a parametric test to test if there is any linear association between two variables in a population. The null hypothesis is that there is no linear association between the two variables in the population.
6) **Partial correlation** is a measure of linear association between two continuous variables when ruling out the influence of a third continuous variable. Partial correlation can be used in a parametric test to test if there is a linear association between two variables in the population when ruling out the influence of the third variable. The null hypothesis is that there is no linear association between the two variables in the population that is not explained by the third variable.

7) **Intra-class correlation** is a measure of association or agreement and has a number of different configurations[87]. We have used a variant of intra-class correlation called the two-way random effects model for single measurements and absolute agreement. Using this intra-class correlation, both the linear association and the agreement between estimates may be assessed in one index (that will range from –1 to 1 as with Pearson's correlation). The intra-class correlation can be used in a parametric test to test if this index is zero in the population. The null hypothesis is that there is no agreement/association in the population (that the index is zero).
8) **Jaccard index** is the ratio of the intersect of two measurable objects divided by their union. The index range from 0 to 1 where 1 indicates an exact similarity. The index is often used to calculate the similarity between segmentations, but can also be used to calculate the similarity between vectors[88] (which is what we do).
9) **Confidence intervals for differences between overlapping Pearson's correlations[89]** can be used to test the difference between two Pearson's correlations that involve a common variable. This, is a parametric test. The null hypothesis is rejected if the 95% confidence interval of the difference contains zero. The null hypothesis is that there is no difference between the two correlations in the population.
10) **Delete-x Jackknife resampling**[90,91] is a method where subsamples are created by removing x observations from the total sample of

> observations. This is done so that all possible combinations of subsamples are created. For example, if applying a delete-one Jackknife resampling on a total sample of 100 observations, 100 unique subsamples are created with 99 observations in each.

Papers I–III are based on the same sample of 70 participants. We chose to include 70 participants with Paper I in mind, the replication study. When doing a replication study it is recommendation to have a larger sample size than the original study[77,78]. In the original study by Eritaia *et al.*[40], 30 participants were included and we figured that twice the sample size would be enough. Thus, we did not determine the sample size by a power calculation, which would have been better. Further, as we thought about Papers II–III already during the sample selection, we should have made power calculations for these studies too, to see if 70 participants were enough. We did not plan for these studies in that extent, but just perceived them as potential bonus studies.

While we included 70 participants, eight participants had to be excluded because their intracranial vault was not fully covered in the MR images. To see if the underlying population of the remaining and excluded participants differed, statistical tests were performed on four demographic variables. These variables were the participants' age, education, MMSE, and gender. If there is a statistical significant difference between the excluded and remaining participants, there has been a systematic exclusion and the remaining sample is no longer random (if it ever was). For gender, a chi-square test was used. For the three other variables, Mann-Whitney U tests were used. Even if age and education might be normally distributed, normality is hard to assess for the small number of excluded participants. Histograms of the participants' age, education, MMSE, and gender are shown in Figure 11 (on the next page). The alpha value was set to 0.05 for all these tests and no correction for multiple comparisons was done. Correction for multiple comparisons might even be unwise in the above tests, as it would increase the risk of type II errors when this risk is already substantial due to the small sample of excluded participants.

Not correcting for multiple comparisons does increase the risk of a type I error (rejecting a true null hypothesis), but an incorrect rejection would only make us overly cautious in the interpretation of following results. A type II error would on the other hand make us less cautious in our interpretation (when there is a need for cautiousness).

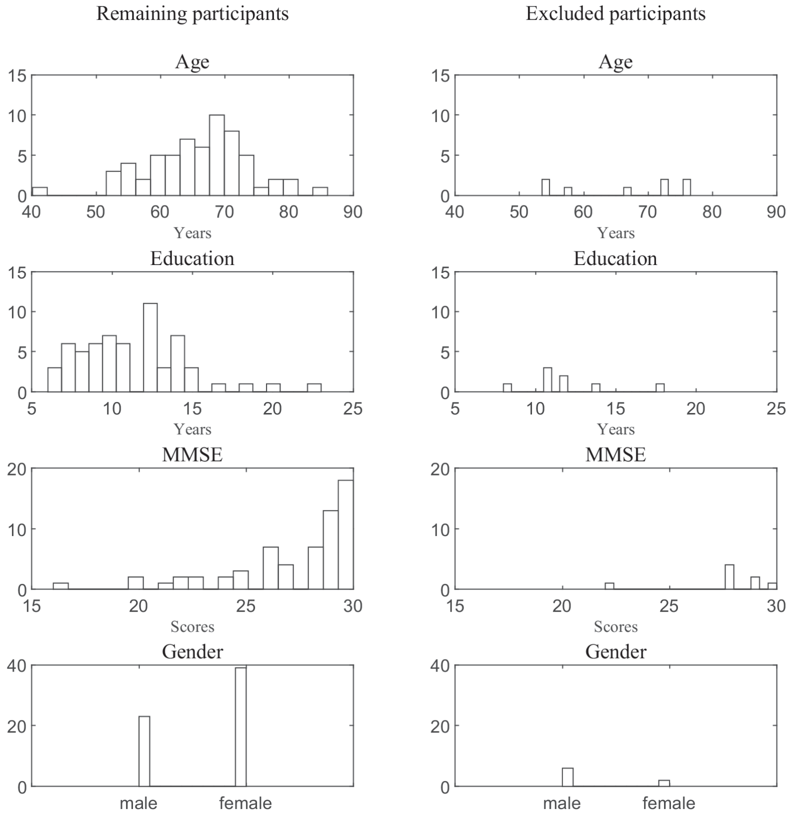


Figure 11. Characterization of the remaining (left column) and excluded (right column) participants shown as histograms of four variables (age, education, MMSE score and gender)

After manually segmenting the whole intracranial vault, we assessed the validity and reliability of these segmentations. Validity was assessed by three tests. The first test was the Pearson's correlation between the estimated volumes and age. The ICV should be robust to aging[92] and we expected that any such correlation in the sample would be small enough to be ruled out by chance alone. That is, that the correlation would not be statistically significant. The second test was made to evaluate if the average ICV differed by gender. This test was done using an independent samples two-tailed t-test. Here, we did expect a difference, as males generally have larger heads than females. Lastly, we evaluated if there was any difference in ICV between healthy controls, GDS 2, GDS 3, and GDS 4 patients. This test was performed using the Kruskal-Wallis test for all participants, but also for both genders separately. While we do expect that at least GDS 4 patients have smaller brain volumes than healthy controls due to brain atrophy, we do not expect such a difference in ICV. Therefore, this last test gives a rough indication whether the manually segmented ICV might be biased by brain atrophy. However, such an interpretation is a bit bold, as we might not even detect a statistically significant difference in total brain volume between the different groups. In the early stages of dementia, the atrophy is small and local. Therefore, to actually detect a statistically significant difference in brain volume between the groups, one might have to use regional brain volume and, in the statistical analysis, account for other factors that may affect the regional brain volume, such as gender or age.

Intra- and inter-rater reliability were calculated for segmentations of every 10th ICA, every 40th ICA, and for the midsagittal ICA. The reliabilities were calculated using intra-class correlation using a two-way random effects model for single measurements and absolute agreement. In Paper I, we write that a two-way *mixed* effects model for single measurements and absolute agreement is used. However, using either of these methods, the exact same analysis is done and the difference between the two models lies in how to interpret the results[87]. The mixed effects model does not allow for generalization of the found rater reliabilities to other raters, which we do want. It was a mistake to say that these intra-class correlations were of mixed rather than random effects.

The alpha values in all these tests were set to 0.05. No correction for multiple comparisons was done in the above tests, which increases the risk of us making a type I error. As we do not expect any statistically significant Pearson's correlation between ICV and age, a type I error would make us question the validity of the ICV, examine the segmentations in more detail, and be more careful in interpreting associations to these segmentations. The same is true for the Kruskal-Wallis test where we neither expect any statistically significant difference. However, for the test of a gender difference, where we expect a statistically significant difference, a type I error would make us ignore the mentioned precautions.

In Paper I, we compared estimates of ICV based on subsamples of equidistant ICAs to the gold standard ICV. We did these comparisons using Jaccard index, Pearson's correlation, percentage error, and intra-class correlation using a two-way random effects model for single measurements and absolute agreement. We included intra-class correlation as it was used in the study by Eritaia *et al.*[40]. Pearson's correlation and percentage errors were included as these measures are used in many other studies evaluating methods for ICV estimation. Jaccard index was included upon request of a reviewer. For each linear spacing, image orientation, and interpolation method that was evaluated, 2000 sets of 62 estimates (one for each MR acquisition) were chosen. The estimates in these sets varied through a random choice of the first ICA in the series of ICAs from the given MR acquisition (which also determines which the rest of the equidistant ICAs will be). The 95 percentiles of these comparisons (2000 for each linear spacing, orientation, and interpolation method), presented in Paper I thus describe how the validity of the estimates varies due to the random choice of the start ICA. No significance tests were performed for any of these comparisons.

In Paper I, the difference in mean absolute percentage errors between using cubic spline interpolation and nearest neighbor interpolation for estimating a volume from the segmented subset of ICAs was evaluated using paired samples two-tailed t-test. One t-test was done for each possible orientation of the ICAs (coronal, sagittal, and transversal) as well as for each spacing between ICAs (from 2 mm between areas to 50 mm between areas). The alpha value was set to 0.05. At total, 147 (3*49) t-tests were done, but no correction for multiple comparisons was done. Instead of probability values (p-values), 95% confidence intervals of the differences were presented. If zero is not within the 95% confidence interval of a difference, the difference is statistically significant. Such a difference implies that there actually is some difference in the underlying population too. However, as no correction for multiple comparisons was done, the confidence intervals are narrower than they would have been with such a correction. Therefore, it might be wise to interpret the results from this analysis with care. In addition, just because there is a difference between methods, it is not necessarily of interest. One also has to judge the size of the difference.

In Paper II, the validity of ICV estimates from four different manual methods were evaluated in comparison to the segmentations of the whole intracranial vault. These comparisons were done using Pearson's correlation and percentage error calculations. We were primarily interested in finding out whether estimates from these methods had a strong linear association to the gold standard ICV. The percentage errors were calculated more for descriptive purposes. No significance tests were performed for the found Pearson's correlations, but their confidence intervals were reported. Confidence intervals were also calculated for the differences between the correlations[89] found between methods 2–4 and the gold standard ICV. This was done upon request of a reviewer. A total of 72 such confidence intervals were reported. None of the reported confidence intervals were corrected for multiple comparisons.

In Paper III, our null hypothesis was that there is no partial correlation between eTIV and total brain volume after controlling for gold standard ICV. The variance left after such controlling is our estimate of the error in eTIV. If eTIV is not biased by a certain factor, the estimated error should be independent of that factor. Thus, if our null hypothesis is rejected, one interpretation is that eTIV is biased by total brain volume. Another interpretation is that FreeSurfer's estimate of total brain volume is biased by eTIV.

As the gold standard ICV also contains some error (for example rater error), the variance left in eTIV after controlling for gold standard ICV could also be due to the error in gold standard ICV. To take this risk into account, we assumed that the Pearson's correlation between gold standard and true ICV should be similar to the correlations found between estimates of ICV using only a few ICAs and the gold standard ICV (~0.99, Paper II). With this in mind, we made a simulation to evaluate the risk of committing a type I error in the partial correlation analysis if eTIV is not biased by total brain volume. Through the simulation we confirmed that the risk of a type I error should be about 5% when using an alpha value of 0.05, using a delete-two Jackknife resampling[90,91] and performing the statistical significance test on the Jackknife replicate with the lowest partial correlation.

Instead of using a delete-two Jackknife resampling, we could have adjusted the alpha value to make sure that the risk of a type I error would be around 5%. However, as partial correlation is sensitive to outliers, we thought that the delete-two Jackknife resampling would be a better way to make the adjustment.

In Paper III, we also calculated the Pearson's correlations between ICV, total brain volume, and eTIV, but we did not perform significance tests for these correlations. We also assessed the difference between eTIV and the manual ICV. These differences were also not statistically tested.

In Paper IV, Pearson's correlations and percentage errors were calculated between actual and predicted statistics. We thought that the Pearson's correlation and percentage errors complemented each other in that if the agreement is bad, then there still is a chance that there is a good linear association. If intra-class correlation with absolute agreement is used, a low correlation may indicate either a low linear association, a low volumetric agreement or both. We did not do any significance tests for the comparisons in Paper IV.

We calculated all statistics in Paper I-IV using different versions of MATLAB.

4 RESULTS

In Papers I–III, a sample of 70 participants was included. Eight of these participants were excluded because their intracranial vault was not covered in the MR images. No statistically significant differences were seen between the remaining and excluded participants regarding age (p-value = 0.789), education (p-value = 0.407) or MMSE (p-value = 0.933). However, a significant difference was seen in the proportion of males (p-value = 0.041).

The whole intracranial vault was segmented for each of the 62 remaining participants. The validity of these segmentations was assessed by statistically evaluating three different properties. First, there was no statistically significant Pearson's correlation between age and the volume of the segmentations (p-value = 0.376). Secondly, a highly significant difference in mean volume was seen between males and females (p-value < 0.001). Lastly, the size of the volumes could not be shown to differ between the healthy controls, GDS 2, GDS 3, and GDS 4 patients ($\text{p-value}_{\text{all}}$ = 0.977, $\text{p-value}_{\text{males}}$ = 0.672, $\text{p-value}_{\text{females}}$ = 0.458). All these results were in line with our expectations.

Intra- and inter-rater reliability were calculated for segmentations of every 10th ICA, every 40th ICA, and for the midsagittal ICA. The intra-rater intra-class correlations for segmenting every 10th ICA and every 40th ICA were both 0.996. For the inter-rater segmentations, these intra-class correlations were 0.991 and 0.987 respectively. For the segmentation of the midsagittal ICA, the intra- and inter-rater Pearson's correlations were 0.997 and 0.995 respectively.

4.1 PAPER I

In Paper I, we evaluate if estimates based on every second up to every 50th ICA are enough to get valid estimates of ICV when using MR images with 1 mm^3 cubic voxels. We found not only that with larger spacing between the ICAs, the validity of the ICV estimates got worse, but also that the choice of start position for the segmentations had a larger impact with larger spacing (see the distribution of the percentile curves in Figure 3, Paper I). The way in which the validity worsened varied depending on the orientation of the ICAs and on the interpolation method used when estimating ICV. At small linear spacings, below 10 mm between the ICAs, there was not much of a difference in validity between the different configurations of the method. At linear spacings beyond 10 mm, estimates calculated using piecewise linear interpolation started to underestimate ICV and by that also got worse intra-class correlation with absolute agreement to our gold standard ICV. For estimates calculated using nearest neighbor interpolation, there was rather a drop in the Pearson's correlation to the gold standard ICV, which was also reflected in the intra-class correlation. When using cubic-spline interpolation, both a good absolute agreement in volume and a high Pearson's correlation were maintained at larger linear spacings. Using cubic spline interpolation, the validity of the estimates at larger linear spacing (>15 mm) was better when using coronal or sagittal ICAs than when using transversal ICAs. This observation is strengthened by the t-tests of the difference in mean absolute percentage error of estimates when using cubic spline compared to nearest neighbor interpolation. For both coronal and sagittal ICAs, the percentage error at larger linear spacings were consistently smaller when using cubic spline interpolation. For transversal ICAs, which of the interpolation methods that resulted in estimates with the smallest percentage errors varied somewhat irregularly with linear spacing.

4.2 PAPER II

In Paper II, we evaluate whether it is enough to segment one or two ICAs to get valid estimates of ICV. Four different methods were evaluated. In Method 1, a single midsagittal ICA is used as an estimate of ICV. In Method 2, a single ICA is multiplied by the intracranial width perpendicular to the ICA. In Method 3, the sum of two ICAs with the same orientation is multiplied by the intracranial width perpendicular to the ICAs. Finally, in Method 4, two ICAs and a shape-preserving piecewise cubic interpolation is used. For Method 1, the Pearson's correlation to our gold standard estimates was 0.904. For Methods 2–4, the correlations ranged between 0.951 and 0.970, between 0.992 and 0.998, and between 0.989 and 0.997 respectively, depending on the orientation of the ICAs. The strongest correlations found for Methods 3 and 4 were found when using sagittal ICAs. For Method 3, the optimal positions of the two sagittal ICAs was at 17.5% and 64% of the perpendicular intracranial width. For Method 4, the two sagittal ICAs should be at 12% and 64% of the perpendicular intracranial width. Estimates from Method 4 using coronal ICAs had the lowest mean percentage error (–1%) compared to the gold standard ICV. However, the standard deviation of these percentage errors was slightly larger than that of estimates using sagittal ICAs. The Pearson's correlations for estimates from Method 4 using coronal ICAs to the whole ICV segmentations were about 0.99. The difference in correlation found between the estimates from Method 4 when using sagittal and coronal ICAs seems to be statistically significant, however, this is when not correcting for multiple comparisons.

4.3 PAPER III

In Paper III, we evaluate if eTIV from FreeSurfer is biased by total brain volume. Our null hypothesis was that no partial correlation between eTIV and total brain volume should remain after controlling for gold standard ICV. To test this null hypothesis, 1891 partial correlations were calculated using a delete-two Jackknife resampling. Then, the Jackknife replicate with the lowest partial correlation was selected in order to do the significance test. The partial correlation between eTIV and total brain volume in this replicate was 0.290, which was statistically significant (p-value = 0.026). Therefore, we rejected our null hypothesis. The median partial correlation for all Jackknife replicates was 0.355. The Pearson's correlation between eTIV and gold standard ICV was 0.960. The Pearson's correlation between eTIV and total brain volume was 0.923. The Pearson's correlation between gold standard ICV and total brain volume was 0.921.

4.4 PAPER IV

In Paper IV, we introduced composite functions of three statistics and three ICV normalization functions. The statistics were the mean, variance and Pearson's correlation to ICV. The ICV normalization functions were those for least-squares, inferred least-squares, and proportion normalization. For inferred least-squares normalization, two types of composite functions were presented. First, functions that predict the mentioned statistics of the normalized data as single values and require information from both the sample from which the regression coefficient is calculated and the sample to normalize. Secondly, functions that predict confidence intervals and require information only from the sample from which the regression coefficient is calculated. In total, twelve composite functions were presented. Using these composite functions, we predicted the mean, variance, and Pearson's correlation to ICV of brain estimates normalized by ICV, using data from two previous studies and a stochastic simulation. The predicted statistics were then compared to the actual post-normalization statistics from these studies and the simulation, using Pearson's correlation and absolute errors. The predicted and actual statistics differed only slightly. The largest deviation between predicted and actual statistics was found for the composite function for variance after proportion normalization. In comparison to actual data from one of the previous studies, these predictions had a mean absolute percentage error of 4.22% and a Pearson's correlation to the actual variance after normalization of 0.981. In comparison to actual data from the simulation, the mean absolute percentage error was 1.10% and the Pearson's correlation 1.000. The predictions from the other composite functions had Pearson's correlations to the actual statistics of above 0.99. The mean absolute errors for these predictions were also small. For the composite functions of statistics after inferred least-squares normalization that predicted confidence intervals, 93.1–94.1% of the actual statistics were within the predicted 95% confidence intervals.

5 DISCUSSION

All four papers included in this thesis include method validation. In the first part of this discussion, I will therefore focus on some general aspects of such validation. In the second part of the discussion, I will look at the evaluations of manual ICV estimation methods in Papers I–II and the evaluation of FreeSurfer's ICV estimate (eTIV) in Paper III. Finally, in the last part of the discussion, I will discuss the effects of ICV normalization and its use in neuroimaging research. I will do so in relation to our findings in Paper IV.

5.1 METHOD VALIDATION

In method validation, validity and reliability are important properties to consider. Validity tells us if estimates are close to the true value. Reliability tells us how similar the estimates are when estimating the same thing a number of times. Often the true value is unknown, and the validity impossible to tell. Then it is common to evaluate the agreement of the method with some reference (or "gold standard") method with supposedly high validity. Agreement is essentially the same thing as validity but closeness is assessed compared to other estimates rather than to true values.

In Paper I, we evaluate the agreement between certain estimates and a gold standard and in Paper IV the validity of certain predictions compared to the true values. To assess the agreement/validity we use a number of different methods including Pearson's correlation. However, it should be noted that Pearson's correlation is a measure of linear dependence and not of agreement. Bland and Altman[93] list a number of reasons why Pearson's correlations are inappropriate when assessing the agreement between two methods. Instead, they suggest the use of what is known as the Bland-Altman plot.

Three other measures of agreement are intra-class correlation with absolute agreement, Jaccard index and percentage error. These three are all used in

Paper I. The Jaccard index is especially useful when assessing the agreement of image masks in structural MR images. For example, if two objects that should agree have the same volume, the percentage error will be 0%, but the Jaccard index of these objects will vary depending on their spatial overlap in image space. The estimation methods evaluated in Papers I–II do not result in image masks. Therefore, we do not use Jaccard index for such masks. In Paper I, we instead used another variant of the Jaccard index that measures the overlap of paired elements between two vectors[88]. The two vectors in this case consists of the new ICV estimates and our gold standard ICV.

For some purposes, it is enough that estimates have a strong association to gold standard estimates while a strong agreement is not necessary. In Paper II, our interest was in finding work-efficient methods to get estimates that are to be used in linear regression models, such as least-squares normalization. When a covariate is added in a linear regression just to correct the model for it, it is enough if the linear association between the covariate and the true value (or gold standard) is strong.

When evaluating the validity of eTIV in Paper III, it was also just linear association that was considered. The linear association evaluated was to an external variable (total brain volume) when controlling for our gold standard ICV (using a partial correlation analysis). We hypothesized that the linear association between eTIV and our gold standard ICV should explain any linear association between eTIV and total brain volume.

Evaluations of method reliability were only included in Papers I–II. In Paper I, this was done by mapping the variability related to the random selection of which ICA to start the segmentation with (see Section 5.2). Intra- and interrater reliabilities were also calculated for segmenting every 10th and every 40th ICA. In Paper II, we only included intra- and interrater reliabilities for one of the four evaluated methods, which is a limitation. In Paper III, we were only interested in a specific aspect of eTIV and did not attempt to evaluate its reliability. However, as the eTIV calculation is automatic and to our knowledge without any influence from any random factor, the same eTIV should be produced every time the same set of MR images are analyzed. Therefore, eTIV should have perfect reliability. For the same reason, the composite functions evaluated in Paper IV should also have perfect reliability.

In the following subsections, I will discuss the interpretation of validity assessments.

5.1.1 SIGNIFICANCE TESTING

In Papers I–II and IV, most validity assessments were done without testing for statistical significance. This might seem worrisome as significance tests are standardly used to draw conclusions about the population. We could for example have used significance tests to test the Pearson's correlations and volumetric differences that we report. By such statistical tests we could have answered the questions "Given that there is no *correlation*/*difference* in the population, what is the probability to get the detected or a *stronger*/*larger correlation*/*difference* in a random sample of our sample size?". If the p-value is low enough, we would reject the null hypothesis that the *correlation*/*difference* in the population is zero.

Theoretically, there almost certainly is some *correlation*/*difference* between any two different estimates in any two samples (and in the population)[94]. And even the slightest *correlation*/*difference* between two estimates will be statistically significant if the sample size is large enough. Thus, if we fail to reject the null hypothesis that there is no *correlation*/*difference* in the population, it is a sign of a too small sample given the effect size seen in it. If we on the other hand do reject the null hypothesis, we only learn the direction of *correlation*/*difference* in the population.

If, prior to the significance test, we believe that the *correlation*/*difference* in the population could actually be zero, the rejection of the null hypothesis would tell us that our observation is unlikely given a zero difference. However, such knowledge is of little use when comparing methods that should estimate the same quantity. As Bland and Altman[93] puts it about significance tests of Pearson's correlations in this context, "The test of significance may show that two methods are related, but it would be amazing if two methods designed to measure the same quantity are not related. The test of significance is irrelevant to the question of agreement.".

5.1.2 EFFECT SIZE

To examine agreement or association between estimates, it is important to use effect size. By the effect size, we can judge the amount of agreement or association. For example, given that the difference in estimated hippocampal volume between two estimation methods is normally distributed, our best estimate of this difference in the population is the mean difference in our sample. If the residuals from a simple linear regression between these two types of hippocampal volume estimates are normally distributed, our best estimate of their Pearson's correlation in the population is the Pearson's correlation in our sample. In Paper I, we report effect sizes of our estimates of agreement/association.

Even if the mean difference and the Pearson's correlation in our sample are our best estimates of the corresponding values in the population, we cannot be sure that they are correct. The smaller sample we have, the less certain we can be about the correctness of our estimates. To make this uncertainty more apparent, one should include the confidence intervals for the effect sizes. A 95% confidence interval is the result of applying an interval estimate statistics such that with repeated sampling, 95% of the found intervals would in the long run contain the true population value. The width of a 95% confidence interval is not a perfect measure of uncertainty (as it will vary between samples), but may at least give some understanding of the uncertainty (especially when intervals are available from a number of studies).

A good agreement in the population might also be insufficient evidence of overall agreement. There could be a small average difference between the estimates and gold standard estimates in the population, but a large variability. If the difference is normally distributed, our best estimate of its variance in the population is its variance in our sample. It tells us *something* about how much the difference between estimates will vary in the population. It is also possible to calculate confidence intervals for the variance[95]. This kind of analysis was not included in any of the papers.

Confidence intervals were included for at least some of the effect sizes in all papers except in Paper III. In Paper III, we were interested in knowing whether there was a significant partial correlation at all rather than in assessing the size of such a correlation. Still, we should have included the confidence intervals from the partial correlation analysis. In the other papers, we included confidence intervals when we thought this was appropriate, but we should have included them for all effect sizes.

5.1.3 ADJUSTMENT FOR MULTIPLE COMPARISONS

When using statistical significance testing one can correct for the use of multiple statistical comparisons to avoid an increased risk of committing type I errors. This comes with the drawback of an increased risk of committing type II errors. It is common to accept this trade-off, as we rather prefer not to draw conclusions to drawing premature and false ones.

As mentioned in Section 5.1.2, to interpret effect size, we should include confidence intervals to specify our uncertainty. For this specification of uncertainty to be correct, we should also adjust it for the number of comparisons made. To adjust the confidence intervals for multiple comparisons, the alpha value can be adjusted. One possibility is to use Bonferroni correction. Then, the alpha value is divided by the number of comparisons[96].

We did not correct any of the confidence intervals presented in Papers I–IV. Due to lack of correction, the presented confidence intervals are smaller than should otherwise be expected.

For Pearson's correlations that are as strong as those described in Papers I–II and IV, the corresponding confidence intervals will not be affected much by correction for multiple comparisons (as long as the sample sizes are not very small). For example, if the Pearson's correlation is 0.99 in a sample of 50 participants, the 95% confidence interval is between 0.982–0.994. By using Bonferroni correction on the alpha value (0.05) for 1000 comparisons (corrected alpha = 0.05/1000), the new confidence interval becomes 0.968–0.997. By correcting for 100,000 tests, the new confidence interval becomes 0.957–0.998. However, the effect of correction is much more apparent if the correlations or sample sizes are small.

5.1.4 SAMPLE SELECTION

One of the cornerstones of statistical analysis is randomly drawn samples. It is by having random samples that we can assume that the mean difference in our sample is a fair estimate of the mean difference in the population we aim to study. It also makes the questions asked during significance testing meaningful (see examples of questions in Section 5.1.1). Most statistical tests or measures that are used to generalize a finding to the underlying population assume that one uses random samples.

While we used a stratified random sampling approach for the data included in Papers I–III, this sampling was done from a larger convenience sample. Further, eight participants had to be excluded as the whole intracranial vault was not covered in their MR images. The proportion of males was significantly larger in the group of excluded participants compared to the remaining ones. Both these factors lead to concerns about how well statistical inference is applicable using the data (as the sampling error not necessarily is by chance). A problem with convenience samples is for example that they tend to be less variable than the population[95]. The exclusion of more males also brings concerns about the representativeness of the current sample.

Despite the sample selection, I do think that the findings in Papers I–III are both trustworthy and generalizable. I can of course not be sure of this. A weak sign of the adequacy of the data in Paper I is that our results agreed well with those of the study that we replicated[40]. Of course, we might have "benefited" from a smaller variability in the sample (compared to the population) due to the use of a convenience sample. For Paper III, I think that the presence of a theoretical explanation of a bias in eTIV supports the validity of the study results. While a theoretical framework does not improve the quality of the data, it may help in the reasoning about for example the generalizability of the findings (rather than just being limited to statistical inference).

5.2 MANUAL ESTIMATION OF INTRACRANIAL VOLUME

One of the most frequently used manual estimation methods in neuroimaging research is one evaluated by Eritaia *et al.*[40] where every 10th sagittal ICA is segmented to estimate the ICV. As the study by Eritaia *et al.* has had such an impact on ICV estimation in neuroimaging and still had never been replicated before, our aim in Paper I was to replicate it.

In our replication study, the results were very similar to those presented by Eritaia *et al.*[40]. The similar results in both studies support the use of every 10th sagittal ICA as a valid ICV estimation procedure. We also found that when segmenting at least every 15th ICA, neither the orientation of the segmentations nor the interpolation method used played any crucial role for the validity of the estimates. As most studies that use linearly spaced ICA to estimate ICV segment every 10th ICA, the orientation of the ICAs or the interpolation method used will not matter much.

In our evaluation, the slice thickness of the MR images was one mm and there was no slice gap between the images. Therefore, in our evaluation, the segmentation of every 10th ICA is equivalent to the segmentation of one ICA every 10th mm. Slice thickness and slice gap are important to have in mind when estimating ICV using linearly spaced ICA. For example, if the MR images have a slice thickness of 2 mm and no slice gap, the use of every 10th ICA will be roughly equivalent to segmenting one ICA every 20th mm when using one mm thick MR images. I say "roughly equivalent", as thicker images generally are less noisy but have larger partial volume (see explanation on page 42) effects than thinner images (which could affect the validity in some direction).

As both we and Eritaia *et al.*[40] found an intra-class correlation with absolute agreement close to one (>0.999) between gold standard ICV and estimates based on every 10th ICA, it seems reasonable to use the latter estimates as gold standard when evaluating new methods (which has been done already[52,53]). However, the reported intra-class correlations in our study and the study by Eritaia *et al.* do not take into account estimation errors (user or method related). When evaluating the intra- and interrater reliabilities for estimates based on every 10th ICA, we found the intra-class correlation (with absolute agreement) to be around 0.991–0.996. Thus, one should remember that the actual validity of these estimates will be restrained by any estimation error introduced by the raters.

With larger linear spacing between the ICAs, both we and Eritaia *et al.*[40] found that the ICV estimates become less reliable. In Figure 3 in Paper I, the reduced reliability is seen as widened percentile curves of the different validity measures with increased linear spacing. The main reason for the reduced reliability with larger linear spacing is probably the random choice of the position of the first ICA to segment. Let me exemplify. Our segmentations of the whole intracranial vault in average contained 136 sagittal ICAs. When using every second ICA to estimate ICV, the position of the first ICA to segment is determined randomly to be either the first or the second ICA in the intracranial vault (from either the right or the left side of the head). If we by chance are directed to start our segmentation at the first ICA in the intracranial vault, then our ICV estimate will include 68 ICAs on average. If we are directed to start the segmentation at the second ICA, our ICV estimate will include the same number of ICAs on average. Further, these two possible sets of ICAs will include very similar area sizes. However, when using every 50th ICA to estimate ICV, the position of the first ICA to segment is determined randomly between the first and 50th ICA in the intracranial vault. If we then are directed to start the segmentation at the first ICA, our ICV estimate will include three ICAs on average. If we are directed to start at the 50th ICA, our ICV estimate will include two ICAs on average. Further, the different sets of ICAs will probably include differently sized areas. Therefore, if we do not let chance direct which ICA to start the segmentation at, it is probable that the reliability does not drop as much with increased linear spacings.

In Paper I, some of the combinations of every 50th ICA reached intra-class correlations with absolute agreement beyond 0.98 to gold standard ICV. Possibly, even higher levels of intra-class correlation can be expected if the position of the first ICA to segment is determined by a given position in the intracranial vault. Further, the validity might also increase by not using linearly spaced ICAs but rather defining the position of all ICAs that should be included in the ICV estimate. Then the segmentation of two ICAs could be enough to get valid estimates of ICV. In Paper II, we evaluated this possibility.

We found that the sum of two ICAs segmented at optimal locations multiplied by the perpendicular intracranial width can result in estimates with very good Pearson's correlation to gold standard ICV. Irrespective of the orientation of the ICAs, Pearson's correlations above 0.99 were found. The strongest Pearson's correlation (of 0.997) was found for estimates using sagittal ICAs. This can be compared to a Pearson's correlation of 0.88–0.89 previously found when segmenting one ICA[42,44], and 0.93–0.95 when segmenting 2-4 midsagittal ICAs[44]. When using sagittal ICAs, the optimal location were found at 17.5% and 64% of the perpendicular intracranial diameter (the width of the intracranial vault from the participant's right side of the head to her left side).

We choose to describe the location of an ICA in the intracranial vault as a percentage of the diameter of the intracranial vault perpendicular to the ICA. We did so assuming that the intracranial vaults are somewhat similar in shape between individuals. Then, similar locations in the intracranial vaults should be found at similar percentage values. As the validity of the estimates was evaluated in a separate subsample from the one used to find the optimal locations, our assumption seems to be correct. However, the assumption might not hold in samples/populations where the shape of the skull might deviate from the general population. For example, in patients with earlier skull trauma or with hydrocephalus.

The fact that the optimal positions of sagittal ICAs were at opposite sides of the intracranial vault could be by chance, but could also be a consequence of bilateral asymmetry in some of the intracranial vaults included in the study. If all intracranial vaults had been perfectly symmetrical, a solution with both ICAs at the same side of the midline would have been equally good. When using the position indices from either the right or the left side of the head randomly between the participants, the Pearson's correlation to gold standard ICV did drop slightly, which could also be an indication of some bilateral asymmetry in the intracranial vault. To what extent the found optimal solution can handle more obvious asymmetries is a matter for further research.

In Paper II, we only aimed at finding simple estimates with high Pearson's correlation to gold standard ICV. We did so, as such estimates would be applicable in linear regression models where an ICV estimate should be included, such as least-squares normalization. We also thought that if absolute volume differences were of interest, then other estimates of ICV would be better. Still, smaller volumetric errors than those presented in Paper II could be achievable for the four methods evaluated using some other way to combine the basic quantities into one single estimate. For example, when only segmenting one ICA and estimating the perpendicular diameter, the algorithm used by Mathalon *et al.*[37] could be applied. Then the estimate would be calculated as 4/3*(diameter/2)*ICA instead of just by ICA*diameter.

5.3 ESTIMATION OF INTRACRANIAL VOLUME USING FREESURFER

One of the most commonly used estimates of ICV is eTIV from FreeSurfer. eTIV is calculated based on the alignment of MR images to a head atlas and not on a segmentation of the images. As the alignment is done to the whole head[62,82], it is not unlikely that the variability in eTIV will depend on more than just the intracranial vault. When eTIV is used in research it is hardly an active choice of the best ICV estimation method, but more likely a passive choice as it is included among the other estimates provided when using FreeSurfer. Often the ICV estimate is not of particular interest (other as a potential adjustment factor) so one might pay less attention to how it is actually determined.

A somewhat unfair, but illustrative example of how eTIV is calculated can be constructed by removing parts of the head in the MR images to be analyzed and evaluate how this affects eTIV. The unfairness with this illustration is that FreeSurfer expects the whole head to be included in the MR images, hence we must not expect perfect estimates when this is not the case.

Nonetheless, using the 62 MR acquisitions included in Papers I–III, I ran two new analyzes using FreeSurfer. In the first analysis, I provided FreeSurfer with MR images where the intensities of all voxels more than 10 mm beyond the border of the manual ICV segmentation had been set to zero. This mainly removed intensities in the throat and neck. In the second analysis, intensities of voxels more than 3 mm beyond the manual segmentations had been set to zero, removing most of the head outside of the dura mater. In Figure 12, I visualize an example of the MR images in this experiment (before and after setting intensities to zero) and how the atlas alignment in FreeSurfer was affected.

In the first analysis, the correlation between ICV and eTIV increased from 0.960 to 0.963. The absolute percentage error of eTIV decreased from 4% to 3.6%. The removal of the throat and neck from the MR images seems to improve eTIV slightly. The throat and neck are not included in the atlas, hence removing them from the MR images might have made the alignment a bit more accurate.

In the second analysis, the correlation between ICV and eTIV decreased from 0.960 to 0.923. The absolute percentage error increased from 4% to 12% and went from an overestimation of ICV in the initial analysis to an underestimation. The reason for this is probably that the alignment tries to fit the whole atlas to the intensities in the images to analyze, which now do not contain any non-zero intensities outside the skull. The result was that fat and skin in the atlas were aligned to the dura mater in the images to analyze.

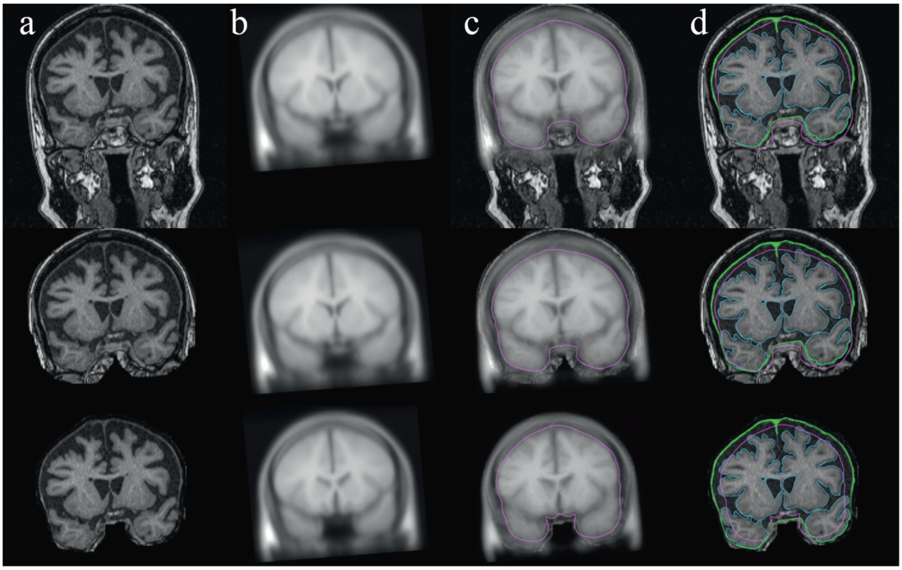


Figure 12. Column a) a T1-weighted image of a participant with frontal lobe dementia; Column b) the MNI305 atlas aligned to the T1-weighted image; Column c) the T1-weighted image with the aligned atlas overlaid and the atlas intracranial surface demarcated (pink contour); and Column d) the T1-weighted image with the brain surface (blue contour), the atlas intracranial surface (pink contour), and the actual intracranial surface (green contour) demarcated. The first row illustrates the initial analysis, the second row illustrates the analysis where intensities more than 10 mm beyond the borders of the manual ICV segmentations were set to zero in the T1-weighted images, and the third row where intensities more than 3 mm beyond the manual segmentations were set to zero.

As the brain constitutes a large part of the head, it is likely to affect the atlas alignment and with that eTIV. With such a dependency, eTIV risks becoming biased by total brain volume. Further, as eTIV derives from a linear alignment of the input MR images to the MNI305 atlas[62], it is not unlikely that the error in eTIV is linearly related to the difference between brain volume and ICV. In Paper III, we empirically showed that there is an association between total brain volume and eTIV that cannot be explained by gold standard ICV. Our interpretation of this association was that eTIV is indeed biased by total brain volume. However, an association does not entail a causal relationship. The association we found could as well be due to the estimation of eTIV somehow being affected by (or affecting) the estimation of total brain volume in FreeSurfer (in a way that cannot be explained by gold standard ICV). Another possibility is that some third variable acts as a confounder.

The MR images included in the study were from examinations made between year 2005 and 2008. Due to the time span between the examinations there is a risk that the MR images vary due to scanner variability. It could be that eTIV and total brain volume are affected by this scanner variability in a way that the gold standard ICV is not (or the other way around). If so, the found partial correlation could be due to a scanner artifact in combination with a susceptibility to this artifact either in the estimation of eTIV and total brain volume or in the gold standard ICV. When inspecting the data, I noticed that there were in fact significant Pearson's correlations ($r = 0.27$–0.31) between acquisition date and both eTIV and gold standard ICV, but no significant correlation between date and total brain volume ($r = 0.21$). Thus, I should have adjusted for acquisition date in my analysis. To test if such an adjustment would have had any effect on our conclusion about eTIV, I reran the analysis adjusting for acquisition date. The delete-two Jackknife replicate with the lowest correlation had a correlation of 0.311 (p-value < 0.02) between eTIV and total brain volume after correcting for gold standard ICV. The median partial correlation of all Jackknife replicates was 0.383. Thus, the results presented in Paper IV holds also when adjusting for scanner variability.

Total brain volume explained 85% of the variance in eTIV (=0.921^2 * 100). However, after controlling for gold standard ICV, total brain volume only explained 1% of the variance in eTIV ([1-0.96^2] * 0.355^2). This tells us that between 1 and 85% of the variance in eTIV in our sample could be due to total brain volume (rather than to ICV). The high correlation between eTIV and ICV is the reason behind the large uncertainty. Due to this uncertainty, other studies must be conducted in order to determine to what extent eTIV is affected by total brain volume. Our study merely points out a theoretical bias in eTIV and shows that it is indicated empirically. It would be better to evaluate the possibility of a bias using longitudinal data. Then the association between change in brain volume and change in eTIV can be evaluated more directly.

Two longitudinal studies had previously been published that evaluated if change in total brain volume was associated with change in eTIV[48,56]. Nordenskjöld *et al.*[48] examined this association in 53 elderly participants that had MR examinations made at an age of 75 years and then five years later. Pengas *et al.*[56] examined the association in 11 patients with semantic dementia that were followed 1.62 years on average (range was 1–3 years). Nordenskjöld *et al.* detected a decrease in total brain volume of 2.51% on average while Pengas *et al.* noted a decrease of 3.22% (1.99%/year * 1.62 years). Neither Nordenskjöld *et al.* nor Pengas *et al.* found any association between change in brain volume and change in eTIV. In the study by Pengas *et al.*, the Pearson’s correlation of 0.515 between the two was just non-significant (the presented p-value is 0.05) and they interpret this as a “trend toward correlation”. While it is not correct to interpret a small p-value as showing a trend towards something, it does tell us that it is fairly improbable to see such a correlation in the sample if there is no correlation in the population.

In Paper III, we discuss how the different study samples might result in different findings and our concern with how the MR scanner upgrade might have affected the findings in the study by Nordenskjöld *et al.*[48]. Yet another example is that Nordenskjöld *et al.* evaluate how absolute change in ICV is associated to relative change in brain volume. I think it would have been better to either evaluate the association between relative *or* absolute changes (and not some combination of them).

In my opinion, to be sure about the existence of a total brain volume bias in eTIV, further studies must be conducted. In the meantime, the use of a more study specific brain atlas might improve the overall validity of eTIV. A better morphological similarity between the data to be analyzed and the atlas should improve the overall alignment, which could benefit eTIV too. Further, the average difference between the volume of the intracranial vault and the total brain volume will be similar in the atlas and the MR images to analyze, which in average should reduce any total brain volume bias in eTIV. The effect of different atlases on a predecessor to FreeSurfer's eTIV is mentioned in a paper by Buckner *et al.*[52]. There they state that when using an atlas of only young people the older participants tended to get overestimated ICV. For this reason, Buckner *et al.* used an atlas based on both young and old adults. FreeSurfer on the other hand use the MNI305 atlas[62] that is based on only young people[97]. Therefore, one could expect that eTIV from FreeSurfer would benefit from a change of atlas when estimating ICV of people with possibly atrophied brains. With the potential influence of the brain volume on eTIV remaining, individual cases still would risk getting poor estimates.

5.4 EFFECTS OF INTRACRANIAL VOLUME NORMALIZATION ON BRAIN ESTIMATES

A number of previous studies have evaluated the effect of ICV normalization on brain estimates. In the present section, I will describe some of these studies and how the composite functions presented in Paper IV might help to interpret the previous findings. I will also use the composite functions to prove certain properties of the different ICV normalization approaches. Remember that the composite functions for proportion normalization are only approximate, so the proofs including these functions will also only have approximate conclusions.

Further, this is ongoing research and Paper IV will be revised before it is submitted for publication.

The provisional nature of this section is actually one reason why it is so long compared to other parts of the discussion: I need feedback on my ideas. Another one, of course, is that the possibility of normalization with ICV is one of the main rationales for estimating ICV.

5.4.1 REDUCED COEFFICIENT OF VARIATION

While I (in Section 1.5) defined ICV normalization in short as the adjustment of brain estimates to reduce the proportion of total variance that is explained by ICV, a reduction of the variance *per se* is not necessarily an indication of a successful normalization. If it was, we could simply divide our brain estimates with a number larger than one. The larger the value we would choose, the more successful would the normalization be. As proportion normalization is division by a large number (or more correctly a large variable), it will automatically cause a large drop in variance. To evaluate if proportion normalization actually reduces variance in any interesting way, the coefficient of variation could be used instead (a measure of variance relative to the mean). As the least-squares normalization does not affect the mean value of the brain estimates, both the variance and the coefficient of variation can be used to evaluate variance reduction due to the normalization.

Hansen *et al.*[36] evaluated how proportion and least-squares normalization decrease the sample size needed to detect a 2% difference in the mean volume between two samples. They made the evaluation for eleven brain regions using hypothetical samples based on real data. Further, the evaluation was made using a number of automatic estimates of ICV (including eTIV from FreeSurfer). Their calculation assumed that the difference is still 2% of the mean volume after normalization. Consequently, the lowest sample size will be achieved by the normalization with the largest reduction of the coefficient of variation.

It was found that least-squares normalization reduced the sample size needed with either of the automatic ICV estimates. Proportion normalization reduced the sample sized needed in all cases but one. The exception was hippocampal volume normalized with eTIV. With least-squares normalization by eTIV, the sample size needed to detect a 2% difference in hippocampal volume was reduced from 406 to 284 participants. With proportion normalization by eTIV, the sample size was increased to 412 participants. The reduction in sample size generally tended to be larger after least-squares normalization than after proportion normalization. Thus, it seems as if least-squares normalization tends to reduce the coefficient of variation more than proportion normalization does.

Through the composite functions presented in Paper IV, we can help in the interpretation of these findings. For example, it might seem surprising that in order to detect a difference in hippocampal volume between two samples, larger sample sizes are required after proportion normalization using eTIV, but not when normalizing other brain estimates. As the sample size calculations in the study by Hansen *et al.* will only depend on which method that reduces the coefficient of variation most, we begin with looking at the variance of a brain estimates (b) after proportion normalization (the variables are explained in the Abbreviations section and in Section 3.11 [Table 7]):

$$\frac{\overline{\imath c v}^2 s_b^2 + \bar{b}^2 s_{icv}^2 - 2\bar{b}\,\overline{\imath c v}\, r_{b,icv} s_b s_{icv}}{\overline{\imath c v}^4}$$

This function can be rewritten as

$$\overbrace{\frac{s_b^2}{\overline{\imath c v}^2}}^{1} + \overbrace{\frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\,\overline{\imath c v}\, r_{b,icv} s_b s_{icv}}{\overline{\imath c v}^4}}^{2}$$

We then see that the proportion normalization has two effects upon the variance of the brain estimate (s_b^2). The first effect is that it is scaled by $\overline{\imath c v}^2$ and the second effect is the addition of $\frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\,\overline{\imath c v}\, r_{b,icv} s_b s_{icv}}{\overline{\imath c v}^4}$. As we are interested in a reduction in the coefficient of variation and not in a scaling of the variance due to a constant (the mean ICV), the function may be simplified by multiplication by $\overline{icv}^2$. We then get

$$s_b^2 + \frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\,\overline{icv}\,r_{b,icv} s_b s_{icv}}{\overline{icv}^2}$$

It is now obvious that the coefficient of variation of the brain estimate will be reduced by proportion normalization only if

$$\bar{b}^2 s_{icv}^2 < 2\bar{b}\,\overline{icv}\,r_{b,icv} s_b s_{icv}$$

This tells us that as long as the Pearson's correlation between the brain estimates and ICV is negative, the coefficient of variation and the sample size needed to detect a given difference in mean volume will be increased by proportion normalization. However, if the Pearson's correlation is positive, which it will be in real data, there might be a reduction in both the coefficient of variation and the sample size. Assuming that the Pearson's correlation is positive, we rewrite the above function to

$$\overbrace{\frac{1}{2r_{b,icv}}}^{1} * \overbrace{\frac{\bar{b}}{s_b} * \frac{s_{icv}}{\overline{icv}}}^{2} < 1$$

Then, we notice two things. Firstly, that the coefficient of variation is more likely to be reduced by proportion normalization if the (positive) Pearson's correlation is above 0.5 and less likely to be reduced if it is below 0.5. Secondly, that proportion normalization is more likely to reduce the coefficient of variation (and hence the sample size needed for a t-test) for brain volumes with a small mean and a large variance, compared to brain volumes with a large mean and a small variance (given the same Pearson's correlations).

Unfortunately, Hansen *et al.*[36] did not report all the factors needed for us to tell exactly why their proportion normalization of hippocampus by eTIV implied a larger sample size needed. Either $\frac{s_{icv}}{\overline{icv}}$ must be larger in eTIV compared to the estimates from the other methods, or the Pearson's correlation between eTIV and hippocampal volume smaller (or both).

Hansen *et al.*[36] did also notice that least-squares normalization reduced the needed sample size more than proportion normalization. Is that always so? As noted above, when the change in coefficient of variation is of interest, we can rewrite the composite function for variance after proportion normalization as

$$s_b^2 + \frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2}$$

Then we notice that the component that reduces the coefficient of variation in the brain estimates is

$$+\frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2}$$

This can be compared to the component that reduces the coefficient of variation (and the variance) in brain estimates when using least-squares normalization (see Table 7 in Section 3.11), which is

$$-r_{b,icv}^2 s_b^2$$

Then, is it possible for the adjustment factor of proportion normalization to be smaller (has a larger negative value) than the adjustment factor of least-squares normalization? We can test this by examine the relationship

$$\frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2} < -r_{b,icv}^2 s_b^2$$

$$\frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2 s_b^2} < -r_{b,icv}^2$$

$$r_{b,icv}^2 + \frac{\bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2 s_b^2} < 0$$

$$\frac{\overline{icv}^2 s_b^2 r_{b,icv}^2 + \bar{b}^2 s_{icv}^2 - 2\bar{b}\overline{icv} r_{b,icv} s_b s_{icv}}{\overline{icv}^2 s_b^2} < 0$$

$$\frac{\left(\overline{icv} s_b r_{b,icv} - \bar{b} s_{icv}\right)^2}{\overline{icv}^2 s_b^2} < 0$$

As neither the denominator nor the numerator of the left-hand side of the equation can be negative, this equation is always false. Proportion normalization cannot reduce the coefficient of variation more than least-squares normalization does. This, and the assumption that normalization will not affect the relative mean difference in brain volumes between two samples, explains why Hansen *et al.*[36] found that the sample size was reduced more by least-squares normalization than by proportion normalization.

While not noticeable in the study by Hansen *et al.*[36] there will be occasions when the proportion and least-squares normalization will reduce the coefficient of variation with the same amount. This will be the case when the expression

$$\frac{\left(\overline{icv} r_{b,icv} s_b - \bar{b} s_{icv}\right)^2}{\overline{icv}^2 s_b^2} = 0$$

is true, which it will be when $\overline{icv} r_{b,icv} s_b = \bar{b} s_{icv}$.

5.4.2 REDUCED LINEAR ASSOCIATION TO INTRACRANIAL VOLUME

As expected, least-squares normalization has been shown to remove any linear association between the brain estimates and ICV[28,37]. In contrast, when using inferred least-squares normalization, the linear association was only fully removed in the sample from which the regression coefficient was calculated[28]. This is also wanted, as inferred least-squares normalization is used to avoid that variance is reduced too much by normalization. The rationale is that some of the association between the brain estimates and ICV might be related to a phenomenon of interest in some subsample and so is an association that we want to keep. The Pearson's correlation left after inferred least-squares normalization varied between –0.2 and 0.15 in the study by Voevodskaya *et al.*. The question is if this remaining association is there by chance or if it is related to a phenomenon of interest? In Paper IV, the composite function for Pearson's correlation after inferred least-squares normalization (see Table 7 in Section 3.11) gives a clue to the answer of this question. The composite function gives the confidence interval for the remaining Pearson's correlation after inferred least-squares normalization. The function assumes that the regression coefficients do not differ between controls and patients in the population(s). Still, the composite function shows that we should expect some Pearson's correlation to remain in the patient sample after applying inferred least-squares normalization (just by chance). In Paper IV, we found that 93.9% of the Pearson's correlation found by Voevdodskaya *et al.* after applying inferred least-squares normalization were within the 95% confidence interval for the predictions using the composite function (and the interval was not adjusted for multiple comparisons). This indicates that many, if not all, of the remaining Pearson's correlations that Voevodskaya *et al.* did find are there just by chance.

In Figure 13 (on the next page), I illustrate the relationship between sample size and the remaining Pearson's correlation between ICV and a brain estimate after using inferred least-squares normalization.

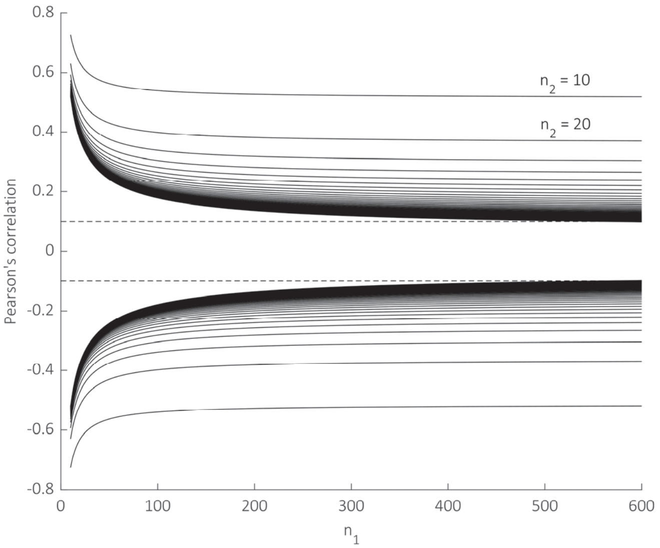


Figure 13. 95% confidence intervals of the remaining Pearson's correlation in a patient sample after applying inferred least-squares normalization. The y axis shows the confidence intervals and the x axis the sample size of the patient sample (n_1). The solid lines show the reduced Pearson's correlation at different sizes of the control sample (n_2, varies from 10 to 500). The two outermost solid lines are at $n_2$1=10 and the two next at n_2=20 and so on. The dotted lines mark a Pearson's correlation of ±0.1.

For proportion normalization, we expect that some association will remain after normalization, at least if the relationship between ICV and the brain estimate is not exactly proportional before normalization[38]. Voevodskaya *et al.*[28] also found that after proportion normalization, most brain estimates had a negative association to ICV. The few exceptions were ventricular and cerebrospinal fluid volumes that still had a positive association to ICV. The Pearson's correlation left after proportion normalization varied between –0.5 and 0.4 (in normal controls and patients with mild cognitive impairment or Alzheimer's disease). Mathalon *et al.*[37] also found the remaining Pearson's after proportion normalization to range somewhere between –0.38 and 0.42

(in normal controls). All proportion normalized effects reported by Voevodskaya *et al.* were predictable by the composite function for Pearson's correlation after proportion normalization in Paper IV (see Table 7 in Section 3.11). The function shows that the effect of proportion normalization on the linear association will depend on three factors: the coefficient of variation of the brain estimates, the coefficient of variation of the ICV, and the Pearson's correlation between the brain estimates and ICV.

Proportion normalization will never reduce the linear association more than least-squares normalization, as the latter always removes this association. However, proportion normalization will also remove any linear association if

$$\frac{r_{b,icv} C_b C_{icv} - C_{icv}^2}{C_{icv}\sqrt{C_b^2 + C_{icv}^2 - 2r_{b,icv} C_b C_{icv}}} = 0$$

For this equation to be true, the numerator must equal zero. That is

$$r_{b,icv} C_b C_{icv} - C_{icv}^2 = 0$$

$$r_{b,icv} C_b C_{icv} = C_{icv}^2$$

$$r_{b,icv} = \frac{C_{icv}}{C_b}$$

$$r_{b,icv} = \frac{\left(\frac{s_{icv}}{\overline{icv}}\right)}{\left(\frac{s_b}{\bar{b}}\right)}$$

$$r_{b,icv} = \frac{\bar{b} s_{icv}}{\overline{icv} s_b}$$

$$\overline{icv} r_{b,icv} s_b = \bar{b} s_{icv}$$

Thus, when $\overline{icv}r_{b,icv}s_b = \bar{b}s_{icv}$ is true, proportion and least-squares normalization will both remove the linear association between the brain estimates and ICV and will reduce the coefficient of variation equally much (see Section 5.1.1).

It should be noted that the reduction of the linear association when using least-squares normalization is for the data set as a whole. If one for example divides the data set after normalization into a control and a patients sample one will still detect some linear association in these two samples separately. The confidence intervals for the remaining association in the separate samples depend on the sample sizes in a way similar to that of inferred least-squares normalization. This function will also be like that illustrated in Figure 13, but with narrower confidence intervals. It is possible to adjust for the difference between samples by including a group*ICV interaction term in the regression analysis. However, the inclusion of an interaction term may lead to overfitting and potentially to type I errors (due to overly reduced variance). As estimates of the linear association in the population will vary between samples, the differences between the subsamples and the total data set are not surprising. The estimate from the total data set should be less uncertain compared to those from the subsamples, as the total data set includes more observations.

5.4.3 REDUCED ESTIMATION RELIABILITY

In 1990, Arndt *et al.*[65] raised a concern about normalization by total brain volume in neuroimaging studies. They showed that both proportion and least-squares normalization reduce the reliability of regional brain estimates. They did so by evaluating how the interrater reliability dropped for a number of regional brain estimates when normalizing them by total brain volume using either proportional or least-squares normalization. The drop in reliability was very similar for proportion and least-squares normalization. Arndt *et al.* also showed that theoretically, this reduction of reliability is expected when using proportion normalization and should become larger the larger the Pearson's correlation to the denominator is before normalization. Therefore, Arndt *et al.* question whether normalization of brain estimates should be used at all.

Mathalon *et al.*[37] replicated the work by Arndt *et al.*[65] for ICV normalization with the same results. Both proportion and least-squares normalization by ICV lowered the reliability of the brain estimates. However, Mathalon *et al.* clarify that the reduced reliability seen after normalization has two possible sources. Firstly, the reduced reliability might be due to an introduced estimation error from the ICV estimates. Secondly, the reduced reliability might be due to a reduced true score variance in the normalized brain estimates. True score variance is variance in the brain estimates that is not due to estimation error. When applying ICV normalization, the aim generally is to reduce the proportion of total true score variance that is explained by ICV. Mathalon *et al.* also point out that we can expect a reduced true score variance when using either least-squares or proportion normalization. They also show that the reduction of variance in the brain estimates (when using either of the normalization approaches) will be stronger the higher the Pearson's correlation between the brain estimates and ICV is before normalization (independently of estimation errors).

One of the statistical tests that Mathalon *et al.*[37] performed was to calculate the Pearson's correlation between seven regional brain volumes and age. This was done both before and after normalization. Through previous studies, Mathalon *et al.* expected that some regional brain volumes would be linearly associated with age. The rationale for once again evaluating such an association was the following. If the reduced reliability after ICV normalization is due only to an increased estimation error, we would expect one of two things (assuming that the estimation error is independent of age). If the sample size is large enough, we expect the Pearson's correlation between the regional brain volume and age to be unaffected by normalization. We also expect the Pearson's correlation to be lower the larger the estimation error is (relative to the sample size). In neither of these cases, we expect the Pearson's correlation to increase. However, if the reduced reliability is due to a reduced true score variance, we rather expect the Pearson's correlation to be unaffected or possibly to increase in size. Mathalon *et al.* found that the Pearson's correlation between volume and age was unaffected or reduced for most of the seven brain regions included in their study after normalization. However, for gray matter volume, the Pearson's correlation increased from about –0.48 to –0.65 after proportion normalization and to –0.66 after least-squares normalization. These results indicate that both proportion and least-squares normalization may in fact result in a reduced true score variance. The question then becomes to what extent the reduced reliability in normalized brain estimates depends on an introduced estimation error from the ICV estimates and how much it depends on the reduced true score variance.

A similar question evaluated by Sanfilipo *et al.*[66] is whether different kinds of errors in the ICV estimates will affect normalized brain estimates differently depending on the chosen normalization approach. Sanfilipo *et al.* evaluated this by systematically introducing two types of artificial errors in a set of 36 ICV estimates and belonging estimates of total brain parenchymal volume. The errors were introduced to either the ICV estimates, the estimates of total brain parenchymal volumes or both. After introducing the errors, the total brain parenchymal volumes were normalized after which their mean volume and standard deviation were calculated. The first type of error introduced was a change in mean volume in six steps from –6% to +6%. This change was done so that the standard deviation would be unaffected. The second type of error introduced was a change in the standard deviation in six steps from –45% to +45%. This change was done while keeping mean volume and the Pearson's correlation to ICV unaltered.

The introduced errors to the ICV estimates had no effect on the least-squares normalized brain estimates[66]. For proportion normalization, both kinds of errors affected both the mean and the standard deviation of the normalized total brain parenchymal volume. When an equal error was introduced to both the brain estimates and ICV, proportion normalization cancelled out the errors in the mean volume of the normalized brain estimates (while the standard deviation was slightly affected). Using least-squares normalization, the error introduced to the brain estimates remained after normalization. To summarize, least-squares normalization seemed to be unaffected by the two types of errors introduced while both affected proportion normalization. In Paper IV, we were able to predict all these results through the composite functions presented there (see Table 7 in Section 3.11).

Through the composite functions in Paper IV, we know that both least-squares and proportion normalization are related to the Pearson's correlation between the brain estimate and ICV. Errors that affect this Pearson's correlation will therefore affect both these normalization methods. As estimation errors often has a random component, they do risk affecting all three normalization methods. Sanfilipo *et al.*[66] seem to know about this possibility (see their limitations section), but left it out in their evaluation.

Using the composite functions from Paper IV, it is possible to express how different errors will affect the results of the different normalization approaches. Then, it should also be possible to express to what degree the reduced reliability after normalization is due to a reduced true score variance and to an introduced estimation error, respectively.

5.5 EFFECTS OF INTRACRANIAL VOLUME NORMALIZATION ON A THIRD FACTOR

When ICV normalizing brain estimates, the purpose often is to improve the analysis of the estimates in relation to some third factor. By reducing the proportion of total variance explained by ICV in the brain estimates, we might affect the association between the brain estimates and some third factor of interest. In some cases the association will get stronger and in others weaker. The composite functions presented in Paper IV are not enough to understand the effect of ICV normalization on the association of a brain estimate and a third factor. One also has to take into account the relationship between the brain estimate and the third factor and the relationship between the third factor and ICV. Further, one must consider the statistical tool used to evaluate the relationship between the brain estimate and the third factor. While all this is possible to do, I will just shortly speculate about the effect of ICV normalization when evaluating gender differences in the brain and when using brain estimates as a diagnostic tool. Of course, this latter example is of great interest in research about dementia diseases.

5.5.1 GENDER DIFFERENCES

When investigating gender differences in regional brain measures, it is common to try to answer the question "Is there a difference between genders with respect to these measures?". To answer this kind of question, unnormalized and ICV normalized brain estimates have been used with different results[19]. Generally, it is found that unnormalized brain volumes tend to be larger in males compared to females[18,19,28,38]. After using proportion normalization, females tend to have larger volumes than males[19,28,38]. And after least-squares normalization the difference between genders is reduced[19,38] or completely removed[28].

As mentioned in Section 5.1.1, not finding a statistically significant difference between two samples is just a sign of too small samples. This also applies to the above findings about gender differences (regardless of ICV normalization or other circumstantial differences). A better kind of question than that above therefore is: "Is there a *meaningful* difference between genders, *when calculated in this way*?". We must then state what we regard as meaningful and what the procedure is for arriving at the results. This second question is not answerable just by a statistical significance. The observed difference must also be considered. It is through this second question that we can understand the effect of different ICV normalization methods on gender differences. First, ICV normalization will affect what we can constructively state as a meaningful difference. Using unnormalized brain volumes, 10 mm^3 could be a meaningful difference, but not using proportion normalized brain estimates (where the amount of dissimilarity is better expressed as a fraction). Secondly, ICV normalization will also change the comparison from "between unnormalized brain estimates" to for example "between brain estimates of persons with the same size of ICV". As the question changes by ICV normalization, the likelihood for a statistical significant difference changes too. This is why a difference in hippocampal volume might be statistically significant when comparing unnormalized volumes, but not after using least-squares normalization.

To understand how the likelihood for a statistically significant difference changes by ICV normalization, power calculations may be done using the composite functions in Table 7 (Section 3.11). However, one must remember that different questions are asked when applying different ICV normalization methods and that a statistically significant difference is not *per se* meaningful. I will give two examples from the literature where the interpretation of the effect of ICV normalization is a bit problematic. The second of these examples are closely related to what I just discussed.

In a study by Barnes *et al.*[18], the effect of normalization on gender difference was evaluated in a sample of 78 normal controls with an age of 61 ±14 years. Before adjusting for ICV, gender explained about 17% of the variance in total brain volume compared to 5–15% in regional brain volumes. The least variance was explained in hippocampus where only 1% was explained by gender. When including age and ICV as covariates in the linear regression, the variance that was explained by gender dropped to between 0–4% for all volumes but hippocampal volume. For hippocampal volume, the variance that was explained by gender increased to 2% after the adjustment. As mentioned by Nordenskjöld *et al.*[38], gender and ICV are highly associated and multicollinearity becomes a problem when both gender and ICV are included as covariates in a regression model. Due to multicollinearity, the estimated effect of gender and ICV in the regression model risks becoming incorrect. With this in mind, the variance in the regional brain volumes that is explained by gender after including ICV should be interpreted cautiously.

In year 2014, Voevodskaya *et al.*[28] concluded that "…residually corrected data effectively removed… the differences in cerebral substructures between men and women". When testing the effect of ICV normalization, they had seen that for 51 different brain regions, none showed a statistically significant gender difference after least-squares normalization. One year later, Nordenskjöld *et al.*[38] did find that some volumes actually are significantly larger in females also after least-squares normalization (thus in contrast to what Voevodskaya *et al.* conclude). Nordenskjöld *et al.* only included five brain regions in their evaluation, two of which were also included in the study by Voevodskaya *et al.*, namely corpus callosum and hippocampus. While neither study found a statistically significant difference in the volume of corpus callosum after least-squares normalization, Nordenskjöld *et al.* did find a statistically significant difference in hippocampal volume. Thus, it could seem as if the effect of least-squares normalization differs between these two studies although both studies use almost the exact same data set from the PIVUS cohort. One difference between the two studies is that Nordenskjöld *et al.* normalize by a manual estimate of ICV while Voevodskaya *et al.* use eTIV. Another difference is that Voevodskaya *et al.* evaluate the effect of normalization on the average hippocampal volume while Nordenskjöld *et al.* evaluate the effect on hippocampal volume separately for the two hemispheres. However, the main cause of the contradicting findings is probably an adjustment for multiple comparisons. The difference in hippocampal volume between genders after least-squares normalization in the study by Nordenskjöld *et al.* is about 2.5%. While Voevodskaya *et al.* do not report the effect sizes, they should have seen the same difference (as basically the same data are being evaluated). However, as Voevodskaya *et al.* applies a correction for multiple comparison (with at least 51 comparisons being adjusted for), this difference is not statistically significant. This highlights the importance of considering effect sizes (and not just statistical significances) when evaluating the effect of ICV normalization on a third factor, such as gender differences.

5.5.2 DIAGNOSTIC ACCURACY

Using logistic regression, both Bigler *et al.*[98] and Voevodskaya *et al.*[28] evaluated the change in diagnostic accuracy of brain estimates after ICV normalization.

Bigler *et al.* did so using proportion normalization, trying to distinguish controls (237 normal controls and 120 with traumatic brain injury) from patients with dementia and patients with some other cognitive disorder (85 participants with Alzheimer's disease and 90 other). Voevodskaya *et al.* used both proportion and inferred least-squares normalization to classify healthy controls (n=223), participants with mild cognitive impairment (n = 325), and participants with Alzheimer's disease (n=175). While Bigler *et al.* evaluated the diagnostic accuracy of a number of brain estimates, Voevodskaya *et al.* did so only for hippocampal volume.

Voevodskaya *et al.*[28] found that proportion normalization and inferred least-squares normalization had only a minor impact on the diagnostic accuracy of hippocampal volume. This was true when trying to differentiate healthy controls from patients with mild cognitive impairment, but also from patients with Alzheimer's disease, and when trying two differentiate the two patient groups. Similarly, Bigler *et al.*[98] did not find any apparent effect of proportion normalization on diagnostic accuracy for hippocampal or temporal horn volume. They did however find a slight positive effect when normalizing total parenchymal brain volume and an unspecified ventricular volume.

Possibly, the positive effect of proportion normalization on total parenchymal brain volume and ventricular volume, but not on hippocampal or temporal horn volume could be that the two former have a stronger Pearson's correlation to ICV. As shown in Section 5.4.1, there is an increased chance that proportion normalization will reduce the coefficient of variation with stronger Pearson's correlation between the brain estimate and ICV. By reducing the proportion of total variance that is explained by ICV, a larger proportion of the variance in the brain estimate will be explained by the disease. Then, the diagnostic accuracy should increase. However, there could be a lot of other explanations too.

Why the inferred residual normalization only had a minor effect on the diagnostic accuracy of hippocampal volume is less easy to understand. All samples had much lower Pearson's correlation to ICV after normalization, which indicates that the coefficient of variation was lowered in all samples. Then disease should explain a larger degree of the total variance after normalization and thus the diagnostic accuracy should increase. The diagnostic accuracy does increase, but just slightly. As the Pearson's correlation between hippocampal volume and ICV in this case was reduced to almost zero in all samples, I do not think that different results would have been found if least-squares normalization had been used. Then, why did the normalization not increase the diagnostic accuracy more?

Another interesting finding was that while Bigler *et al.*[98] did not find that proportional ICV normalization increased the diagnostic accuracy of hippocampal volume, they did find that the diagnostic accuracy increased when normalizing by total brain parenchymal volume instead. Then the specificity increased from 79% to 87% and the sensitivity from 81% to 86%. One reason could be that total brain parenchymal volume still was largely unaffected by the cognitive disorder while having a stronger Pearson's correlation to hippocampal volume (than ICV had). Another reason could be that total brain parenchymal volume also had some diagnostic value that was included through the proportion normalization. Larger total brain volume is a plausible sign of normality.

Westman *et al.*[99] also evaluated the use of proportion normalization (as far as we can tell) when trying to differentiate patients with mild cognitive impairment that latter convert to Alzheimer's disease to those who do not. This was done by creating OPLS (orthogonal projections to latent structures) models based on different types of brain estimates (subcortical volumes, cortical volumes, and cortical thickness estimates). The models were created using MR images from healthy controls and patients with Alzheimer's disease and then evaluated using MR images from 287 patients with mild cognitive impairment of which 87 converted to Alzheimer's disease within 1.5 years. The model that achieved the highest diagnostic accuracy during evaluation was that in which all three types of brain estimates were included, but only the subcortical and cortical volumes were proportion normalized. This model had an increased specificity of 66.5% compared to 64.0% when only unnormalized brain estimates were used. The sensitivity, of 75.9%, was not affected by normalization. Similarly, the effect of proportion normalization on the diagnostic accuracy of all other models evaluated overall was slight or none. According to Westman *et al.*, the small effect of proportion normalization on diagnostic accuracy in the OPLS models could be that the use of multiple brain estimates for prediction is robust enough to not benefit from ICV normalization.

The effect of ICV normalization on diagnostic accuracy is hard to assess, but is of large interest not least for studies of dementia diseases. Through an understanding of what happens in different scenarios, maybe with the use of the composite functions presented in Table 7 (Section 3.11) and some more thought, a clearer picture of the effect of different ICV normalization approaches will be possible to get. Current findings do however indicate that the effect of ICV normalization on diagnostic accuracy of brain estimates in dementia classification is limited.

5.6 FINAL THOUGHTS

In Section 1.5, I described the purpose of ICV normalization by

> *ICV normalization is done to reduce the proportion of the total variance of a brain estimate that is predicted by ICV, using some statistical model that supposedly describes some true relationship between ICV and the brain region.*

Least-squares, inferred least-squares, and proportion normalization all are means to fulfill this purpose. However, with both proportion normalization and inferred least-squares normalization there is a risk that the variance explained by ICV is increased. For inferred least-squares normalization, there is a theoretical reasoning behind taking this risk, which might be sound. Still, the potential benefit should be evaluated more thoroughly than has been done up to date. For proportion normalization, the reasoning behind the choice is unclear to me and it is questionable as normalization method due to the varying effect. I recommend referring to proportion normalization as “division by ICV” to not put any misleading emphasis on normalization. If it is reasonable to expect a normalizing effect, one should show this (for example using the composite functions in Table 7). Further, to make least-squares normalization more graspable for most scientists, I think it is better referring to it as “controlling for ICV”.

When applicable, I would use multiple linear regression (or a generalized linear regression) with ICV as a covariate to control for ICV, instead of using the (least-squares) function presented in Section 1.5. I base this choice on the facts that the exact same results should be achievable through both methods (with different configurations), and that linear regression models are well known in the scientific community. By using ICV as a covariate in regression analyzes, it also becomes natural to follow existing guidelines about how to perform good regression analyzes.

When controlling for ICV in linear regression models, or when applying inferred least-squares normalization, it is enough if the ICV estimate has a strong linear association to gold standard ICV. However, if the size of the effect of ICV also is of interest, an estimate with good agreement with gold standard ICV is necessary. For the first purpose, the segmentation of two ICA as proposed in Paper II will do fine. When possible, I would avoid the use of FreeSurfer’s eTIV as it has a weaker correlation with gold standard ICV. Additionally, eTIV might be biased by total brain volume. With that said, I am not entirely against the use of eTIV. The correlation is strong enough to reduce most of the association between brain estimates and ICV and the potential bias might not be too problematic in most situations. Another automatic ICV estimation might still be preferable. Finally, when an estimate with good agreement to gold standard ICV is desirable, I would use manual segmentations of every 10th ICA.

When using proportion normalization (division by ICV), ICV estimates having strong agreement with gold standard ICV are necessary.

Two assumptions are often made when using ICV normalization. First, that the association between the phenomenon under study and the brain estimate to normalize is independent of ICV. In dementia research, such an assumption is sometimes questioned by a concept called “brain reserve”. The concept suggests that larger premorbid brain volume is a protective factor against cognitive impairment. There are a number of studies addressing the brain reserve issue, but the results differs[47,100-102]. In developmental psychiatric disorders, it is also possible that the premorbid brain volume becomes affected by the disorder during growth. For example, patients with autism seem to have a different brain growth curve than do healthy controls[103].

The second assumption made when using ICV normalization is that the intracranial vault is constant throughout adulthood (and therefore a good proxy for premorbid brain volume). This is not always the case. For example, in cases of hyperostosis frontalis interna, the intracranial vault will become smaller due to thickening of the inner surface of the frontal bone[104].

It is important that the two mentioned assumptions are considered when using ICV normalization in order to avoid misinterpretations of normalized brain estimates in cases when the assumptions might not hold.

Throughout the thesis, I have discussed ICV normalization in the perspective of cross-sectional settings. This is the most common case as longitudinal studies focus on intraindividual variability over time and ICV should in general be stable over time (in adulthood). Hence, it will not explain much of the intraindividual variability. Despite this, ICV normalization might actually be of value in longitudinal studies too. Whitwell *et al.*[39] point out that when trying to find subtle changes in brain volumes over time, the possibility to detect these differences may decrease by changes in the MR acquisition. One such change that they mention is that of voxel size due to scanner drift. They also hypothesize that such a change might be controlled for in longitudinal studies by adjusting brain estimates for ICV.

6 CONCLUSION

In the present thesis, it was shown that a frequently used method for estimation of ICV, where every 10th ICA is segmented, is valid. This supports the findings previously shown by Eritaia *et al.*[40]. We also showed that the orientation of the ICA or the interpolation method used did not affect the validity of the estimates in any practical sense (when every 10th ICA is segmented). The intra-class correlation with absolute agreement was above 0.999 compared to gold standard ICV. When considering rater estimation errors, a slightly lower agreement was found.

It was also shown that ICV estimates with very strong linear association to gold standard ICV were achievable by segmenting two selected ICAs and measuring the intracranial diameter perpendicular to the orientation of the ICAs. The Pearson’s correlation was shown to be 0.997.

eTIV from FreeSurfer was investigated and both theoretical and empirical considerations revealed a potential bias by total brain volume. In the sample, at least 1% of the variance in eTIV was explained by total brain volume after controlling for gold standard ICV, but it could be up to 85%. As the study design was cross-sectional, causality could not be tested.

In the thesis, I present composite functions that predict the variance, mean, and Pearson’s correlation to ICV after different ICV normalization procedures. These composite functions may help to improve our understanding of ICV normalization. For example, it was shown that proportion normalization increases the coefficient of variation for some brain estimates and reduces it for others. It was also shown that when proportion normalization reduces the coefficient of variation, this reduction will never exceed that of least-squares normalization. Further, after inferred least-squares normalization, some Pearson’s correlation to ICV will likely remain just by chance.

7 FUTURE PERSPECTIVES

It should be established exactly how the reduced reliability in brain estimates after ICV normalization depends on a reduced true score variance and an introduced estimation error (of ICV) respectively. This should be possible by the use of the composite functions in Table 7 (Section 3.11).

Some of the association that will remain between ICV and inferred least-squares normalized brain estimates is likely to be there by chance. To understand the use of inferred least-squares normalization, it should be evaluated if this expected random error is outweighed by the benefit of the remaining variance of interest (which would have been removed using least-squares normalization).

There are indications of that eTIV from FreeSurfer is biased by total brain volume. These indications should be followed up to confirm/reject the presence of such a bias. Similarly, other frequently used automatic estimates of ICV need to be examined with regard to the risk of biases related to how they estimate ICV.

The role of biomarkers in diagnosis of dementia diseases is becoming stronger. Therefore, it will become even more important to optimize the diagnostic accuracy of such biomarkers. There are reasons to believe that ICV normalization will increase the diagnostic accuracy of brain estimates from MR images. In previous studies, the effect of ICV normalization on diagnostic accuracy seems to be small. The causes of this should be investigated and the theoretically possible gain with ICV normalization in this context established. The composite functions in Table 7 (Section 3.11) will help in such an effort.

ACKNOWLEDGEMENTS

Anders Wallin, my main supervisor. Thank you for giving me the opportunity to do research under your supervision, for giving me room to grow as a researcher, and for all your support throughout my PhD studentship!

Helge Malmgren, my co-supervisor. Thank you for introducing me to research, for your endless support, brilliant mind, and friendship! I feel privileged to have been your student.

Carl Eckerström, my co-supervisor. Thank you for all your support throughout my PhD studentship! I always felt secure in my progress knowing that you kept an eye on it.

Erik Olsson. Thank you for your endless support and friendship! You have played a substantial role in my scientific education and always helped me more than I could have asked for.

Simon Skau. Thank you for all your help, fun discussions, and friendship!

Fellow PhD students and coffee drinkers at the Memory clinic, Mölndal. Thank you for all your support and shared experiences. I miss our coffee breaks!

Thank you to coworkers in Anders Wallin's research group and at MedTech West, and all others that have helped me during my PhD studentship! To mention a few Christel Källquist Karlsson, Daniel Ruhe, Rolf Heckemann, Eva Bringman, and Helene Lindström.

My parents and siblings. Thank you for always believing in me and for all your love!

Louise Karlsson. Thank you for finding me. For all your love and kindness.

Further thanks to Sahlgrenska University Hospital (ALF), Swedish Brain Power, Stiftelsen psykiatriska forskningsfonden, Hjalmar Svensson Foundation, The Gothenburg Foundation for Neurological Research, Wilhelm och Martina Lundgrens Vetenskapsfond, and Demensfonden for funding my research.

REFERENCES

1. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical descriptions and Diagnostic Guidelines. Geneva: World Health Organization, 1992.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing, 2013.
3. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand 2009;119:252-265.
4. Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. J Psychiatr Res 2009;43:411-431.
5. Weissberger GH, Strong JV, Stefanidis KB, Summers MJ, Bondi MW, Stricker NH. Diagnostic Accuracy of Memory Measures in Alzheimer's Dementia and Mild Cognitive Impairment: a Systematic Review and Meta-Analysis. Neuropsychol Rev 2017;27:354-388.
6. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 2010;6:67-77.
7. Olsson B, Lautner R, Andreasson U, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. Lancet Neurol 2016;15:673-684.
8. Whitwell JL. Alzheimer's disease neuroimaging. Curr Opin Neurol 2018;31:396-404.
9. Eckerström C, Olsson E, Bjerke M, et al. A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia. J Alzheimers Dis 2013;36:421-431.
10. Eckerström C, Olsson E, Klasson N, et al. Multimodal prediction of dementia with up to 10 years follow up: the Gothenburg MCI study. J Alzheimers Dis 2015;44:205-214.
11. Frölich L, Peters O, Lewczuk P, et al. Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. Alzheimers Res Ther 2017;9:84.
12. Bessi V, Mazzeo S, Padiglioni S, et al. From Subjective Cognitive Decline to Alzheimer's Disease: The Predictive Role of Neuropsychological Assessment, Personality Traits, and Cognitive Reserve. A 7-Year Follow-Up Study. J Alzheimers Dis 2018;63:1523-1535.
13. Markisz J, Aquilia M. Technical magnetic resonance imaging. Stamford, Connecticut: Appleton & Lance, 1996.
14. Wahlund L-O, Westman E, van Westen D, Wallin A, Cavallin L, Larsson EM. Strukturell hjärnavbildning kan förbättra diagnostiken vid demens. Läkartidningen 2013;110.
15. Sorensen L, Igel C, Liv Hansen N, et al. Early detection of Alzheimer's disease using MRI hippocampal texture. Hum Brain Mapp 2016;37:1148-1161.
16. Achterberg HC, van der Lijn F, den Heijer T, et al. Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Hum Brain Mapp 2014;35:2359-2371.

17. Ma D, Gulani V, Seiberlich N, et al. Magnetic resonance fingerprinting. Nature 2013;495:187-192.

18. Barnes J, Ridgway GR, Bartlett J, et al. Head size, age and gender adjustment in MRI studies: a necessary nuisance? Neuroimage 2010;53:1244-1255.

19. Pintzka CW, Hansen TI, Evensmoen HR, Haberg AK. Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a HUNT MRI study. Front Neurosci 2015;9:238.

20. Fraser MA, Shaw ME, Anstey KJ, Cherbuin N. Longitudinal Assessment of Hippocampal Atrophy in Midlife and Early Old Age: Contrasting Manual Tracing and Semi-automated Segmentation (FreeSurfer). Brain Topogr 2018;31:949-962.

21. Geuze E, Vermetten E, Bremner JD. MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders. Mol Psychiatry 2005;10:160-184.

22. Jansen AG, Mous SE, White T, Posthuma D, Polderman TJ. What twin studies tell us about the heritability of brain development, morphology, and function: a review. Neuropsychol Rev 2015;25:27-46.

23. Gianaros PJ, Jennings JR, Sheu LK, Greer PJ, Kuller LH, Matthews KA. Prospective reports of chronic life stress predict decreased grey matter volume in the hippocampus. Neuroimage 2007;35:795-803.

24. Erickson KI, Prakash RS, Voss MW, et al. Aerobic fitness is associated with hippocampal volume in elderly humans. Hippocampus 2009;19:1030-1039.

25. Hibar DP, Westlye LT, Doan NT, et al. Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol Psychiatry 2018;23:932-942.

26. Woollett K, Maguire EA. Acquiring "the Knowledge" of London's layout drives structural brain changes. Curr Biol 2011;21:2109-2114.

27. Draganski B, Gaser C, Kempermann G, et al. Temporal and spatial dynamics of brain structure changes during extensive learning. J Neurosci 2006;26:6314-6317.

28. Voevodskaya O, Simmons A, Nordenskjold R, et al. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. Front Aging Neurosci 2014;6:264.

29. Jack CR, Jr., Twomey CK, Zinsmeister AR, Sharbrough FW, Petersen RC, Cascino GD. Anterior temporal lobes and hippocampal formations: normative volumetric measurements from MR images in young adults. Radiology 1989;172:549-554.

30. Schmidt MF, Storrs JM, Freeman KB, et al. A comparison of manual tracing and FreeSurfer for estimating hippocampal volume over the adult lifespan. Hum Brain Mapp 2018;39:2500-2513.

31. Poulakis K, Pereira JB, Mecocci P, et al. Heterogeneous patterns of brain atrophy in Alzheimer's disease. Neurobiol Aging 2018;65:98-108.

32. Machts J, Vielhaber S, Kollewe K, Petri S, Kaufmann J, Schoenfeld MA. Global Hippocampal Volume Reductions and Local CA1 Shape Deformations in Amyotrophic Lateral Sclerosis. Front Neurol 2018;9:565.

33. Vernooij MW, Jasperse B, Steketee R, et al. Automatic normative quantification of brain tissue volume to support the diagnosis of dementia: A clinical evaluation of diagnostic accuracy. Neuroimage Clin 2018;20:374-379.

34. Lopez OL, Becker JT, Chang Y, et al. Amyloid deposition and brain structure as long-term predictors of MCI, dementia, and mortality. Neurology 2018;90:e1920-e1928.

35. O'Brien LM, Ziegler DA, Deutsch CK, et al. Adjustment for whole brain and cranial size in volumetric brain studies: a review of common adjustment factors and statistical methods. Harv Rev Psychiatry 2006;14:141-151.

36. Hansen TI, Brezova V, Eikenes L, Haberg A, Vangberg TR. How Does the Accuracy of Intracranial Volume Measurements Affect Normalized Brain Volumes? Sample Size Estimates Based on 966 Subjects from the HUNT MRI Cohort. AJNR Am J Neuroradiol 2015;36:1450-1456.

37. Mathalon DH, Sullivan EV, Rawles JM, Pfefferbaum A. Correction for head size in brain-imaging measurements. Psychiatry Res 1993;50:121-139.

38. Nordenskjold R, Malmberg F, Larsson EM, et al. Intracranial volume normalization methods: considerations when investigating gender differences in regional brain volume. Psychiatry Res 2015;231:227-235.

39. Whitwell JL, Crum WR, Watt HC, Fox NC. Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. AJNR Am J Neuroradiol 2001;22:1483-1489.

40. Eritaia J, Wood SJ, Stuart GW, et al. An optimized method for estimating intracranial volume from magnetic resonance images. Magn Reson Med 2000;44:973-977.

41. Eckerström C, Olsson E, Borga M, et al. Small baseline volume of left hippocampus is associated with subsequent conversion of MCI into dementia: the Goteborg MCI study. J Neurol Sci 2008;272:48-59.

42. Ferguson KJ, Wardlaw JM, Edmond CL, Deary IJ, Maclullich AM. Intracranial area: a validated method for estimating intracranial volume. J Neuroimaging 2005;15:76-78.

43. MacLullich AM, Ferguson KJ, Deary IJ, Seckl JR, Starr JM, Wardlaw JM. Intracranial capacity and brain volumes are associated with cognition in healthy elderly men. Neurology 2002;59:169-174.

44. Nandigam RN, Chen YW, Gurol ME, Rosand J, Greenberg SM, Smith EE. Validation of intracranial area as a surrogate measure of intracranial volume when using clinical MRI. J Neuroimaging 2007;17:74-77.

45. Schofield PW, Logroscino G, Andrews HF, Albert S, Stern Y. An association between head circumference and Alzheimer's disease in a population-based study of aging and dementia. Neurology 1997;49:30-37.

46. Perneczky R, Wagenpfeil S, Lunetta KL, et al. Head circumference, atrophy, and cognition: implications for brain reserve in Alzheimer disease. Neurology 2010;75:137-142.

47. Jenkins R, Fox NC, Rossor AM, Harvey RJ, Rossor MN. Intracranial volume and Alzheimer disease: evidence against the cerebral reserve hypothesis. Arch Neurol 2000;57:220-224.

48. Nordenskjold R, Malmberg F, Larsson EM, et al. Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements. Neuroimage 2013;83:355-360.

49. Aghamohammadi-Sereshki A, Huang Y, Olsen F, Malykhin NV. In vivo quantification of amygdala subnuclei using 4.7 T fast spin echo imaging. Neuroimage 2018;170:151-163.

50. Luo Y, Cao Z, Liu Y, et al. T2 signal intensity and volume abnormalities of hippocampal subregions in patients with amnestic mild cognitive impairment by magnetic resonance imaging. Int J Neurosci 2016;126:904-911.

51. Stening E, Persson J, Eriksson E, Wahlund LO, Zetterberg H, Soderlund H. Apolipoprotein E 4 is positively related to spatial performance but unrelated to hippocampal volume in healthy young adults. Behav Brain Res 2016;299:11-18.

52. Buckner RL, Head D, Parker J, et al. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage 2004;23:724-738.

53. Keihaninejad S, Heckemann RA, Fagiolo G, Symms MR, Hajnal JV, Hammers A. A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T). Neuroimage 2010;50:1427-1437.

54. Daugherty AM, Yu Q, Flinn R, Ofen N. A reliable and valid method for manual demarcation of hippocampal head, body, and tail. Int J Dev Neurosci 2015;41:115-122.

55. Raz N, Rodrigue KM, Head D, Kennedy KM, Acker JD. Differential aging of the medial temporal lobe: a study of a five-year change. Neurology 2004;62:433-438.

56. Pengas G, Pereira JM, Williams GB, Nestor PJ. Comparative reliability of total intracranial volume estimation methods and the influence of atrophy in a longitudinal semantic dementia cohort. J Neuroimaging 2009;19:37-46.

57. Malone IB, Leung KK, Clegg S, et al. Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. Neuroimage 2015;104:366-372.

58. Raz N, Ghisletta P, Rodrigue KM, Kennedy KM, Lindenberger U. Trajectories of brain aging in middle-aged and older adults: regional and individual differences. Neuroimage 2010;51:501-511.

59. Crowley SJ, Tanner JJ, Ramon D, Schwab NA, Hizel LP, Price CC. Reliability and Utility of Manual and Automated Estimates of Total Intracranial Volume. J Int Neuropsychol Soc 2018;24:206-211.

60. Smith SM, Jenkinson M, Woolrich MW, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 2004;23 Suppl 1:S208-219.

61. Ashburner J. Computational anatomy with the SPM software. Magn Reson Imaging 2009;27:1163-1174.

62. FreeSurfer website. eTIV [online]. Available at: http://www.freesurfer.net/fswiki/eTIV. Accessed 18 Jan.

63. Katuwal GJ, Baum SA, Cahill ND, et al. Inter-Method Discrepancies in Brain Volume Estimation May Drive Inconsistent Findings in Autism. Front Neurosci 2016;10:439.

64. Huo Y, Asman AJ, Plassard AJ, Landman BA. Simultaneous total intracranial volume and posterior fossa volume estimation using multi-atlas label fusion. Hum Brain Mapp 2017;38:599-616.

65. Arndt S, Cohen G, Alliger RJ, Swayze VW, 2nd, Andreasen NC. Problems with ratio and proportion measures of imaged cerebral structures. Psychiatry Res 1991;40:79-89.

66. Sanfilipo MP, Benedict RH, Zivadinov R, Bakshi R. Correction for intracranial volume in analysis of whole brain atrophy in multiple sclerosis: the proportion vs. residual method. Neuroimage 2004;22:1732-1743.

67. Wallin A, Nordlund A, Jonsson M, et al. The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. J Cereb Blood Flow Metab 2015.

68. Wallin A, Nordlund A, Jonsson M, et al. Alzheimer's disease-subcortical vascular disease spectrum in a hospital-based setting: overview of results from the Gothenburg MCI and dementia studies. J Cereb Blood Flow Metab 2015.

69. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189-198.

70. Reisberg B, Ferris SH. Diagnosis and assessment of the older patient. Hosp Community Psychiatry 1982;33:104-110.

71. Reisberg B, Ferris SH, Shulman E, et al. Longitudinal course of normal aging and progressive dementia of the Alzheimer's type: a prospective study of 106 subjects over a 3.6 year mean interval. Prog Neuropsychopharmacol Biol Psychiatry 1986;10:571-578.

72. Royall DR, Mahurin RK, Gray KF. Bedside assessment of executive cognitive impairment: the executive interview. J Am Geriatr Soc 1992;40:1221-1226.

73. Wallin A, Edman A, Blennow K, et al. Stepwise comparative status analysis (STEP): a tool for identification of regional brain syndromes in dementia. J Geriatr Psychiatry Neurol 1996;9:185-199.

74. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry 1982;140:566-572.

75. Berg L, Miller JP, Storandt M, et al. Mild senile dementia of the Alzheimer type: 2. Longitudinal assessment. Ann Neurol 1988;23:477-484.

76. Morris JC. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. Int Psychogeriatr 1997;9 Suppl 1:173-176; discussion 177-178.

77. Tversky A, Kahneman D. Belief in the law of small numbers. Psychol Bull 1971;76:105-110.

78. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14:365-376.

79. Bunketorp Kall L, Malmgren H, Olsson E, Linden T, Nilsson M. Effects of a Curricular Physical Activity Intervention on Children's School Performance, Wellness, and Brain Development. J Sch Health 2015;85:704-713.

80. Olsson E, Eckerstrom C, Berg G, et al. Hippocampal volumes in patients exposed to low-dose radiation to the basal brain. A case-control study in long-term survivors from cancer in the head and neck region. Radiat Oncol 2012;7:202.

81. Fischl B. FreeSurfer. Neuroimage 2012;62:774-781.

82. FreeSurfer website. What is included in the ICV value [online]. Available at: https://mail.nmr.mgh.harvard.edu/pipermail/freesurfer/2008-September/008370.html. Accessed 18 Jan.

83. FreeSurfer website. MorphometryStats [online]. Available at: http://www.freesurfer.net/fswiki/MorphometryStats. Accessed 18 Jan.

84. Seltman H. Approximations for E(R/S) and V(R/S) for any random variables R and S [online]. Available at: http://www.stat.cmu.edu/~hseltman/files/ratio.pdf. Accessed 29 Nov.

85. Pearson K. Mathematical contributions to the theory of evolution - On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 1897;60:489-498.

86. Fuguitt GV, Lieberson S. Correlation of ratios or difference scores having common terms. Sociological Methodology 1973-1974;5:128-144.

87. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-163.

88. Sepkoski JJ. Quantified coefficients of association and measurement of similarity. Math Geol 1974;6.

89. Zou GY. Toward using confidence intervals to compare correlations. Psychol Methods 2007;12:399-413.

90. Pigeot I. The jackknife and bootstrap in biomedical research - Common principles and possible pitfalls. Drug Inf J 2001;35:1431-1443.

91. Shao J, Wu CFJ. A General-Theory for Jackknife Variance-Estimation. Ann Stat 1989;17:1176-1197.

92. Blatter DD, Bigler ED, Gale SD, et al. Quantitative volumetric analysis of brain MR: normative database spanning 5 decades of life. AJNR Am J Neuroradiol 1995;16:241-251.

93. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet 1986;1:307-310.

94. Meehl PE. Theory-Testing in Psychology and Physics - Methodological Paradox. Philos of Sci 1967;34:103-115.

95. Cortinhas C, Black K. Statistics for business and economics: John Wiley & Sons, 2012.

96. Altman DG. Practical statistics for medical research. London: Chapmann and Hall, 1991.

97. Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM. 3d Statistical Neuroanatomical Models from 305 Mri Volumes. Nuclear Science Symposium & Medical Imaging Conference, Vols 1-3 1993:1813-1817.

98. Bigler ED, Tate DF. Brain volume, intracranial volume, and dementia. Invest Radiol 2001;36:539-546.

99. Westman E, Aguilar C, Muehlboeck JS, Simmons A. Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. Brain Topogr 2013;26:9-23.

100. Wolf H, Julin P, Gertz HJ, Winblad B, Wahlund LO. Intracranial volume in mild cognitive impairment, Alzheimer's disease and vascular dementia: evidence for brain reserve? Int J Geriatr Psychiatry 2004;19:995-1007.

101. Edland SD, Xu Y, Plevak M, et al. Total intracranial volume: normative values and lack of association with Alzheimer's disease. Neurology 2002;59:272-274.

102. Sumowski JF, Rocca MA, Leavitt VM, et al. Brain reserve and cognitive reserve protect against cognitive decline over 4.5 years in MS. Neurology 2014;82:1776-1783.

103. Courchesne E, Campbell K, Solso S. Brain growth across the life span in autism: age-specific changes in anatomical pathology. Brain Res 2011;1380:138-145.

104. May H, Mali Y, Dar G, Abbas J, Hershkovitz I, Peled N. Intracranial volume, cranial thickness, and hyperostosis frontalis interna in the elderly. Am J Hum Biol 2012;24:812-819.

APPENDIX

A PICTORIAL GUIDE FOR SEGMENTATION OF INTRACRANIAL VOLUME

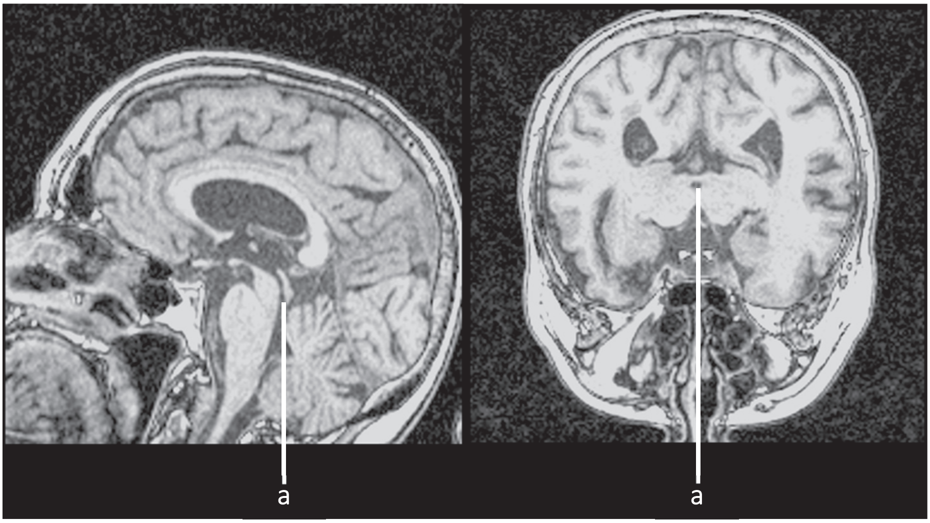
To make it easier for new raters to segment the intracranial volume and to keep a high intra- and interrater reliability throughout my work, I made a pictorial guide for segmentation of the intracranial vault. The guide is based on the landmarks described by Eritaia *et al.* [2000] (see Section 3.6.2), but includes a few more landmarks. The additional landmarks are the cerebral aqueduct, epidermis, foramen lacerum, jugum sphenoidale, maxillary sinus, pituitary gland, and sphenoid sinus.

In the guide, I begin with describing and illustrating most of the landmarks. The guide then continues with an example segmentation of about half an intracranial vault. The intensity setting in the images is the one described in Section 3.5 except that the images are made a bit darker. The pictorial guide is made for segmentation in sagittal orientation, preferably guided by the other orientations too.

Cerebral Aqueduct

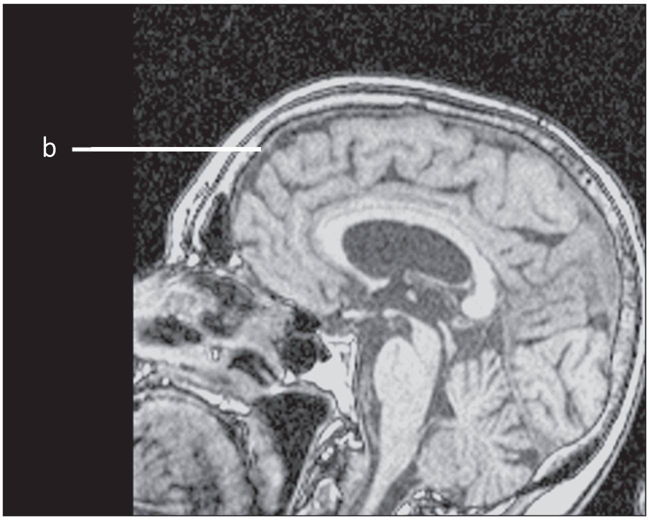
The demarcation is easiest to carry out starting near the longitudinal fissure, where the cerebral hemispheres meet each other, and continue laterally in both directions until the last traces of the meninges disappear. At the lateral ends, it can be hard to distinguish the meninges from the skull without considering the surrounding slices. Starting at one of the lateral ends makes it easy to miss a slice where the meninges are visible or to demarcate areas that should not be included in the intracranial volume.

The cerebral aqueduct (a) is used as a landmark for the slice where to begin the segmentation. To find this slice, scroll through the sagittal slices to the slice where the cerebral aqueduct is most pronounced. The cerebral aqueduct is a passage between the third and the fourth ventricles and is filled with cerebrospinal fluid. In T1-weighted images, it appears as a dark line in sagittal orientation.



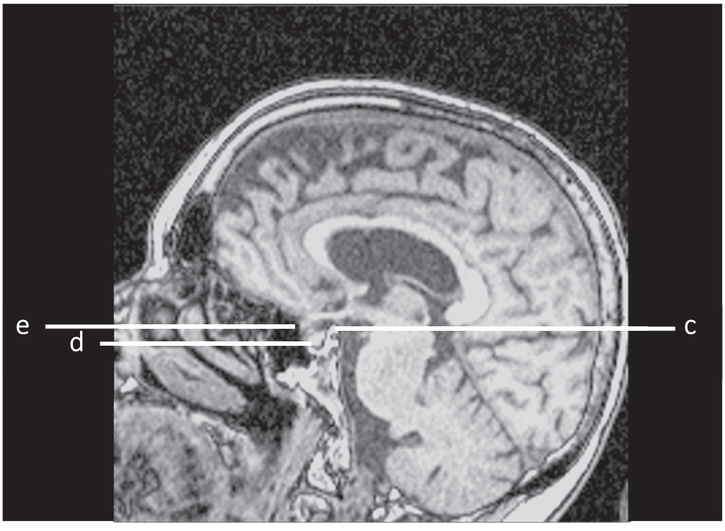
Dura Mater

The dura mater (b) is a thick layer of collagenous connective tissue and is the outermost layer of the meninges. The dura mater is closely attached to the inner surface of the skull and is seen in T1-weighted images as a bright line separated from the brain by cerebrospinal fluid. When the brain lies against the dura mater, they are hard to distinguish from each other. In the image below, it is for example hard to separate the dura mater from the brain posteriorly.



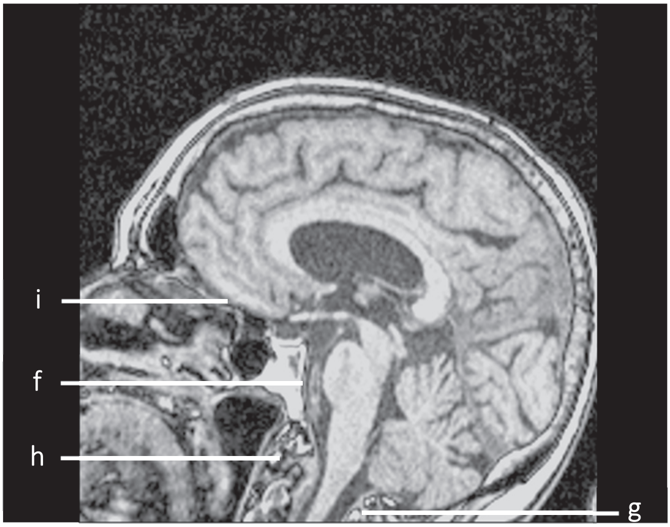
Dorsum Sellae

Dorsum sellae (c) is a part of the sphenoid bone and lies posterior to the pituitary gland (d) that is located in a small depression formed by the sella turcica. During the segmentation, the pituitary gland is excluded from the intracranial volume by drawing across the sella turcica from the dorsum sellae to the jugum sphenoidale (e) that connects the two lesser wings of the sphenoid bone.



Clivus

Clivus (f) is a part of the posterior cranial fossa and slopes from dorsum sellae to foramen magnum. The clivus is demarcated all the way from foramen magnum up to dorsum sellae.



Foramen Magnum

At foramen magnum, the large hole in the occipital bone, the demarcation line is drawn across the spinal cord at the position of the posterior arch of atlas (g) to the position of the anterior arch of atlas (h). Atlas is the superior cervical vertebra and supports the skull.

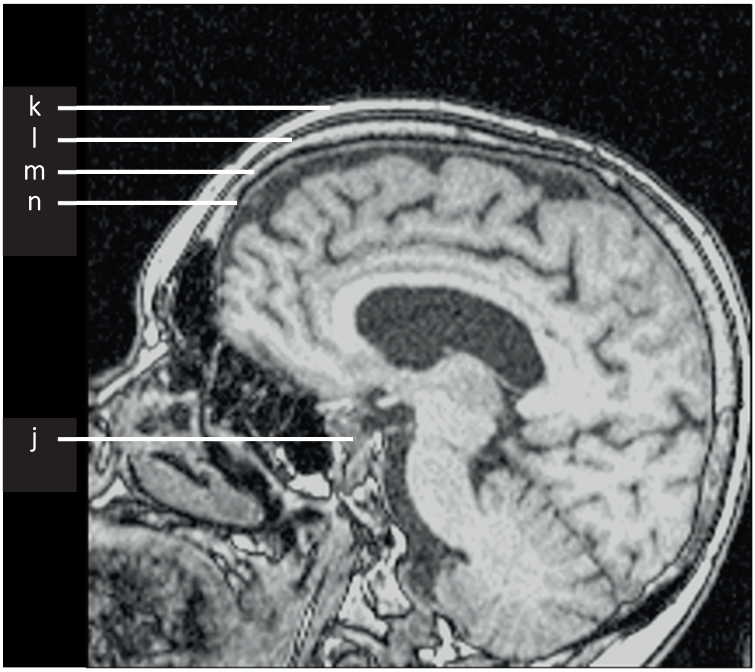
Undersurface of the frontal lobe

The undersurface of the frontal lobe (i) lies above the anterior cranial fossa and the sphenoid bone. It is outlined, as the rest of the brain contour, when the dura mater is not visible.

Foramen lacerum

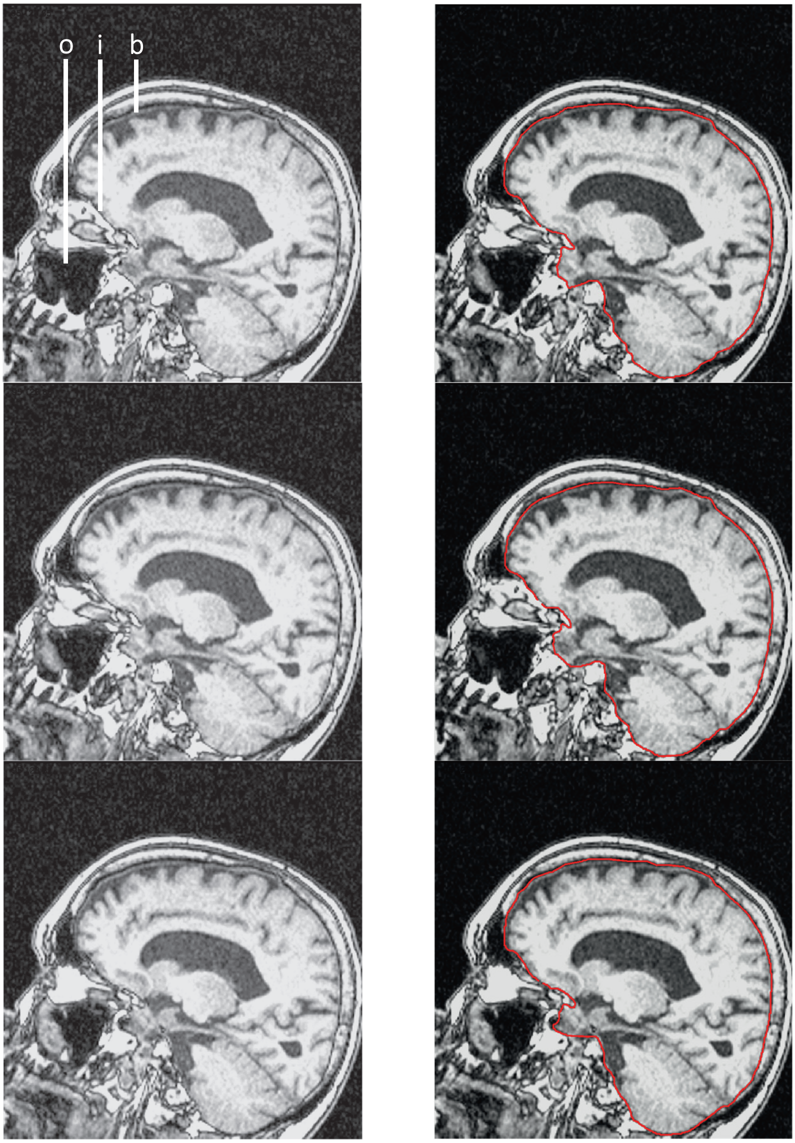
Foramen lacerum (j) is an opening between the sphenoid and temporal bone lateral to the sella turcica and through which for instance the internal carotid artery runs. Superior to the foramen lacerum is the cavernous sinus that creates a small cavity of veins. Foramen lacerum is used as a landmark to end the narrow demarcation used to exclude the pituitary gland. The demarcation line is drawn across the foramen lacerum at its most superior part and then follows the structure of the sphenoid bone.

In the image below epidermis (k), the outer table of the skull (l), the diploë (m) and the inner table of the skull (n) are marked besides foramen lacerum (j).

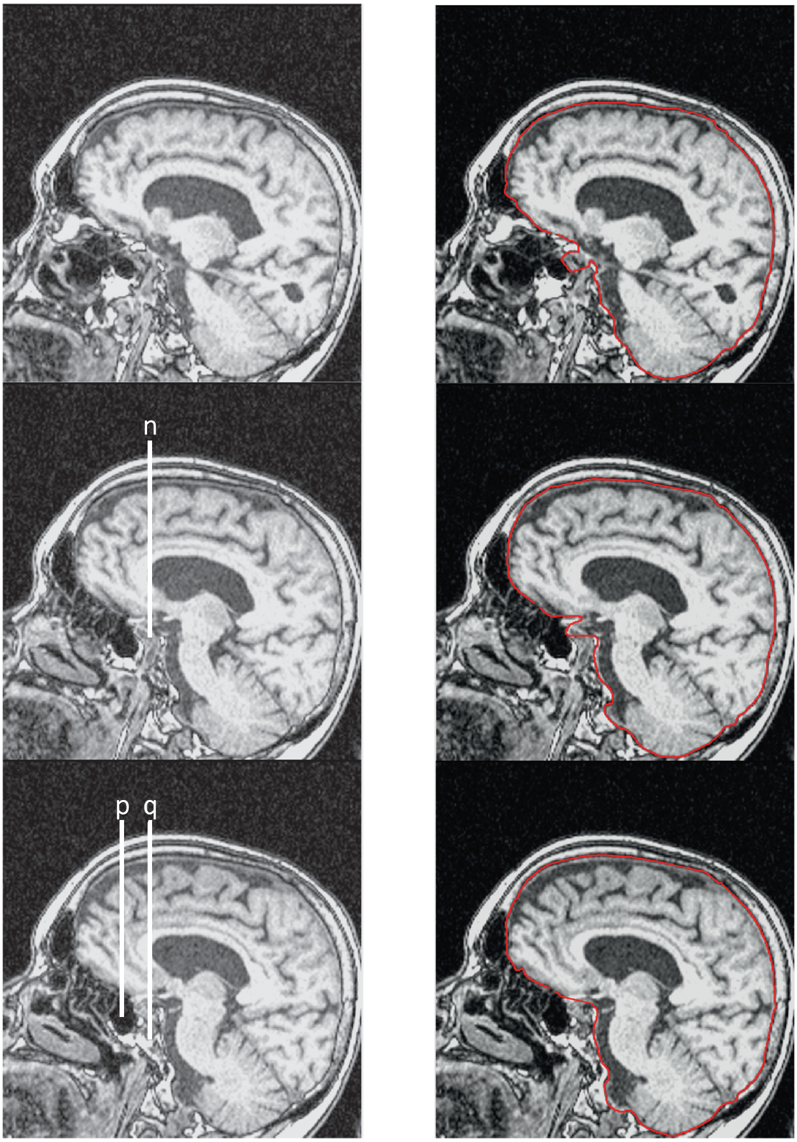


Example Segmentation

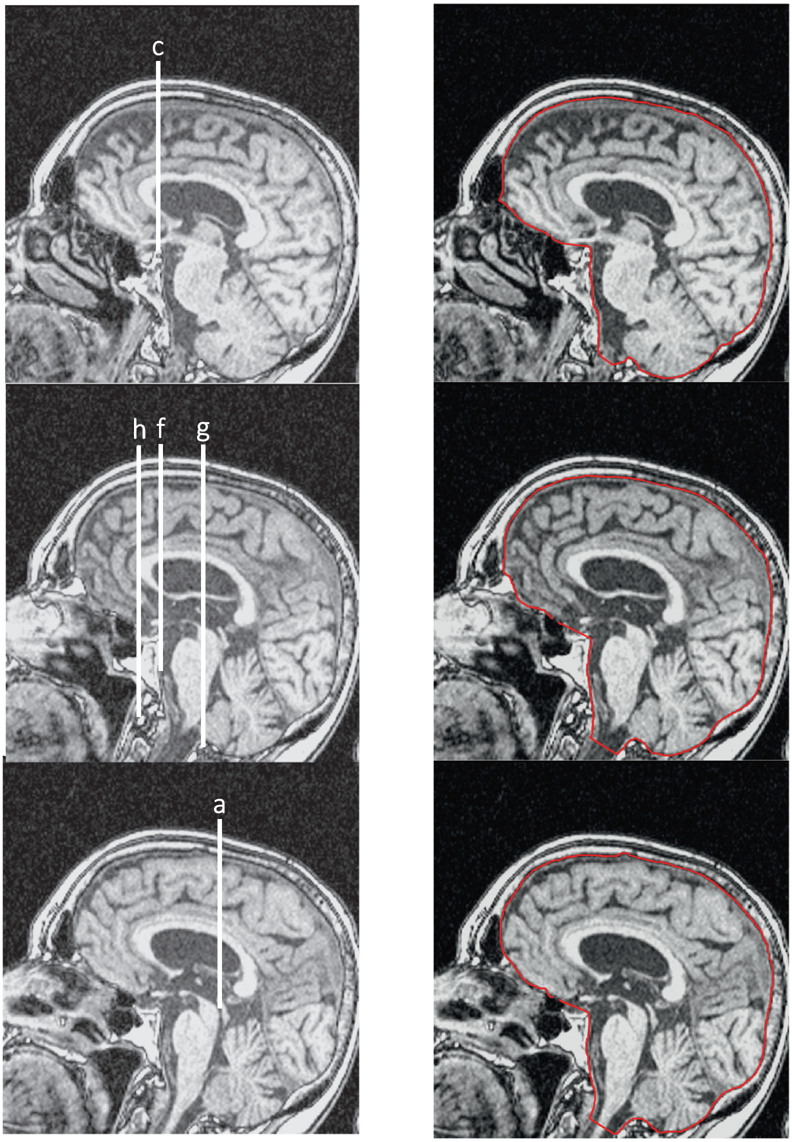
In the following pages, an example segmentation is illustrated for a bit more than half an intracranial vault. The left column shows the MR images before the segmentation and the right column after the segmentation.



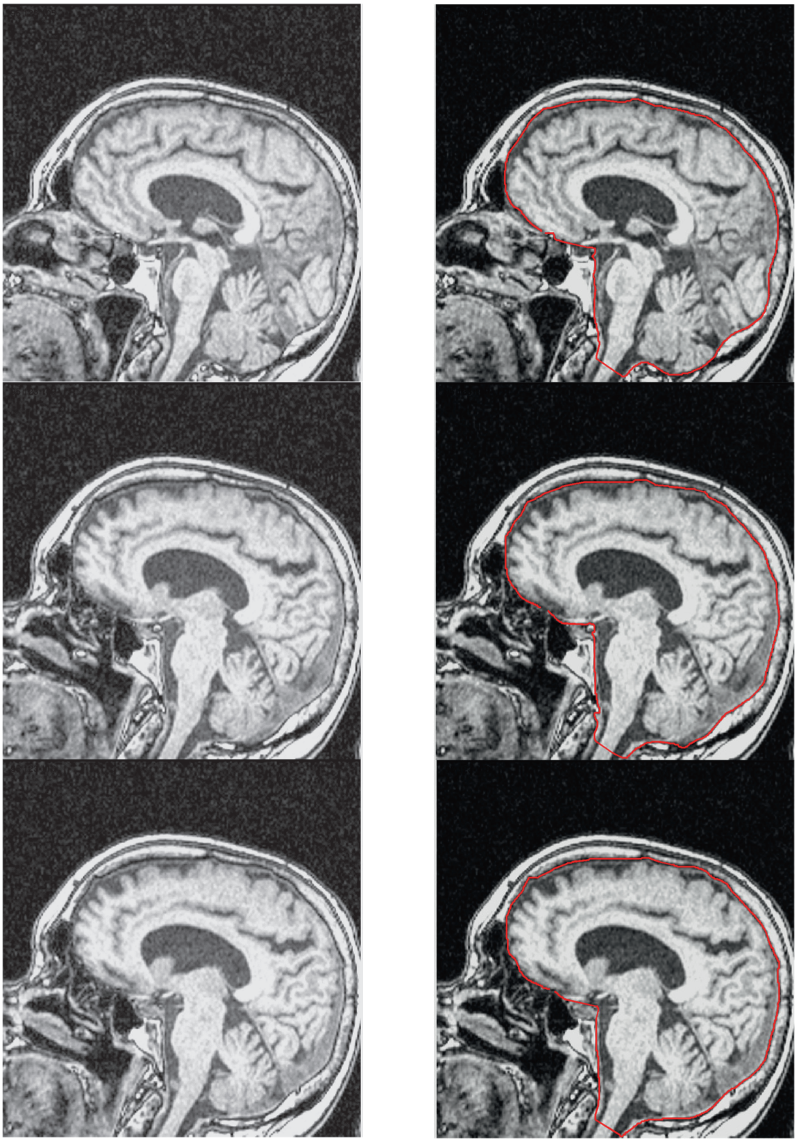
b) dura mater, i) undersurface of the frontal lobe, o) maxillary sinus

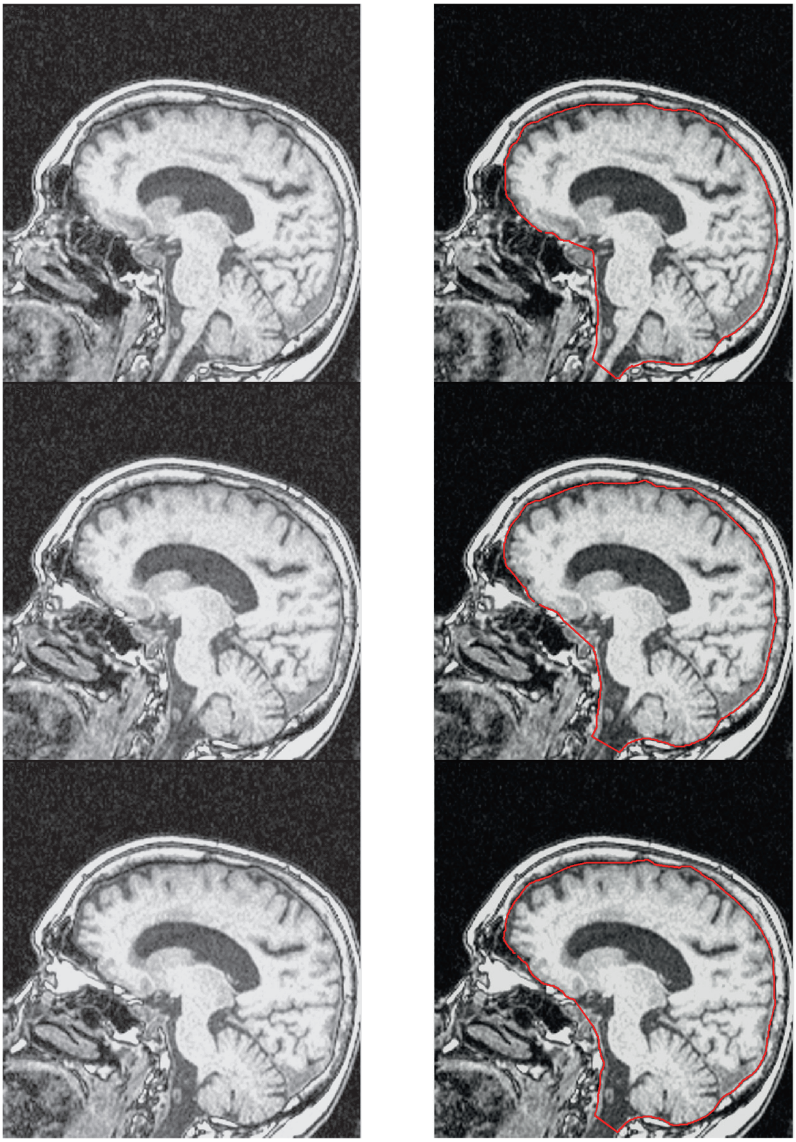


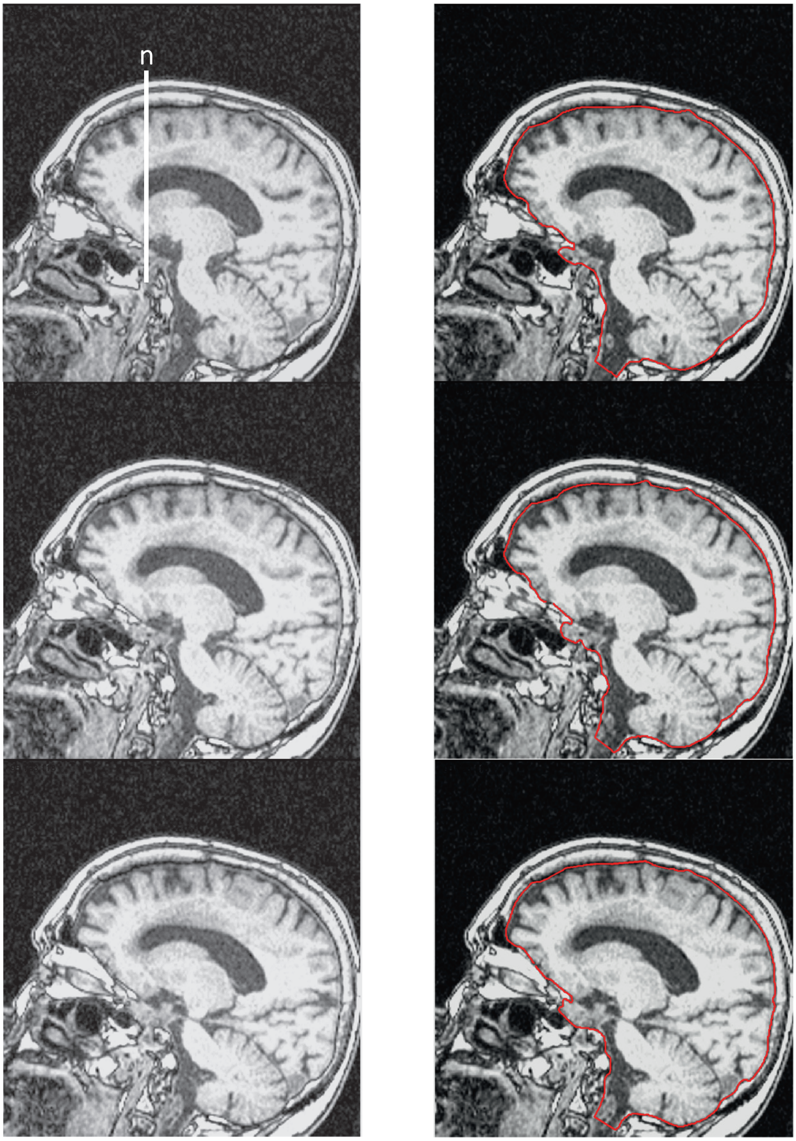
n) foramen lacerum “open”, p) sphenoid sinus, q) foramen lacerum “closed”



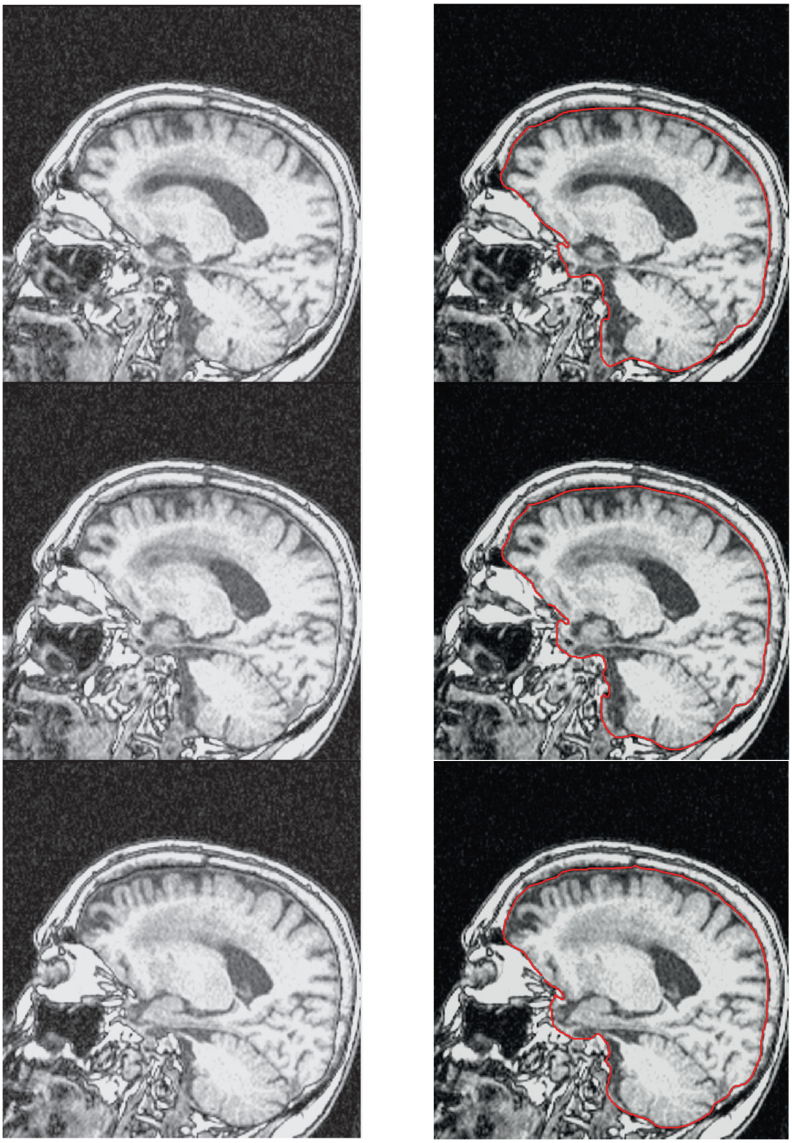
a) cerebral aqueduct, c) dorsum sella, f) clivus, g) posterior arch of Atlas, h) anterior arch of Atlas

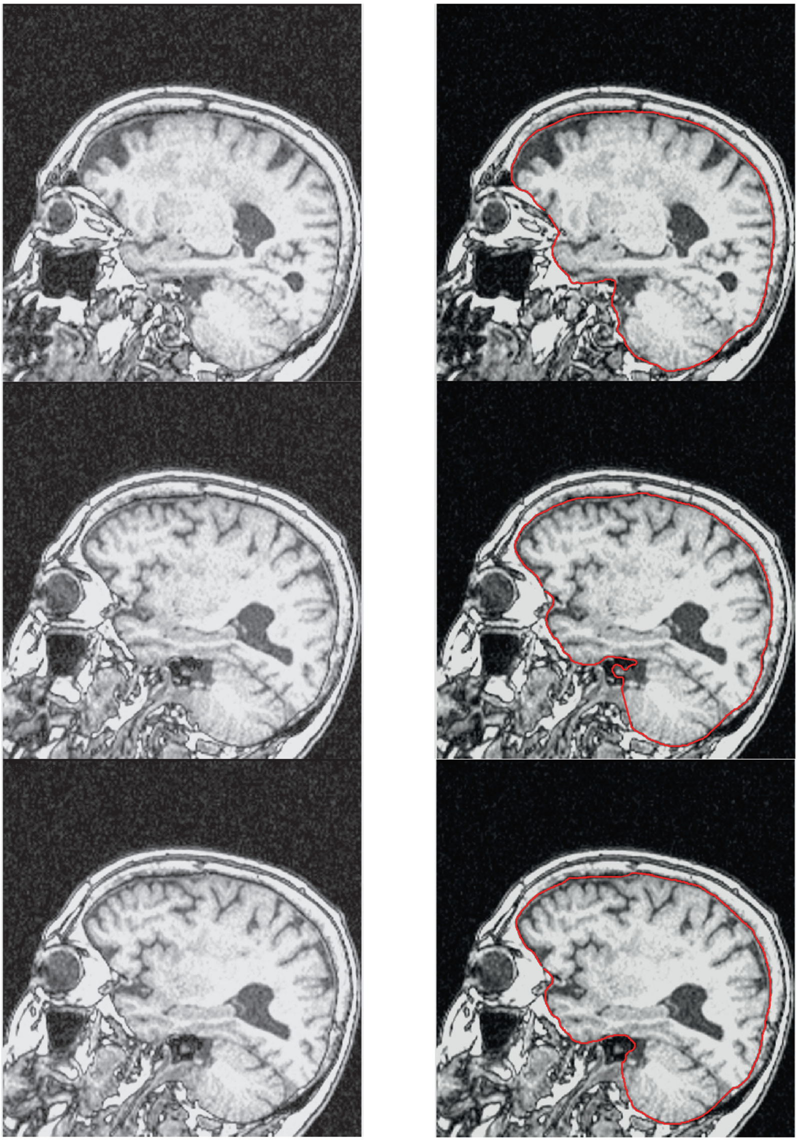


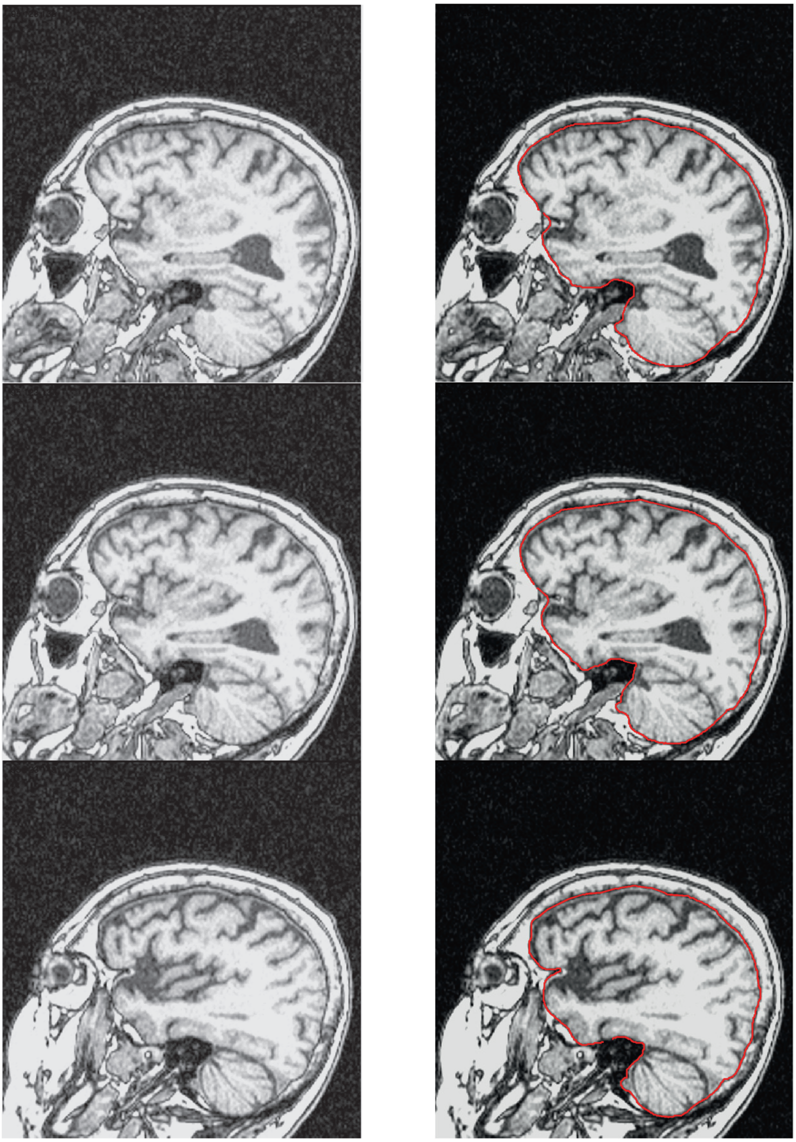


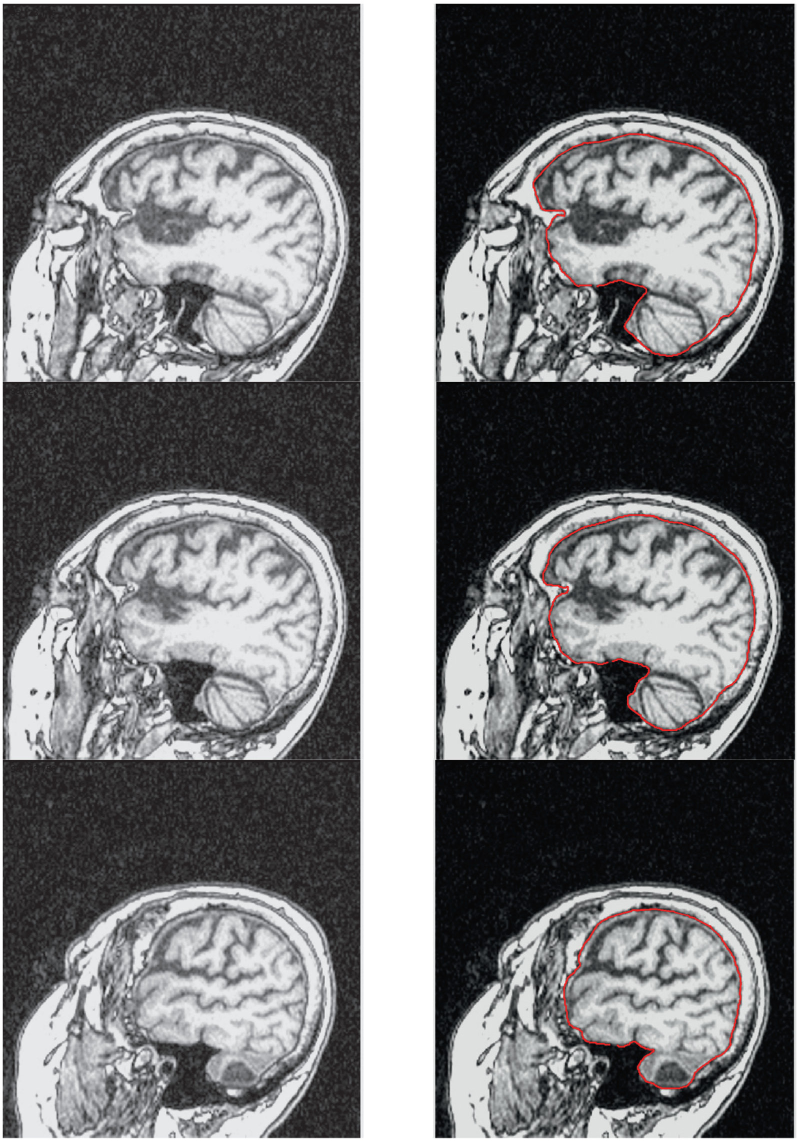


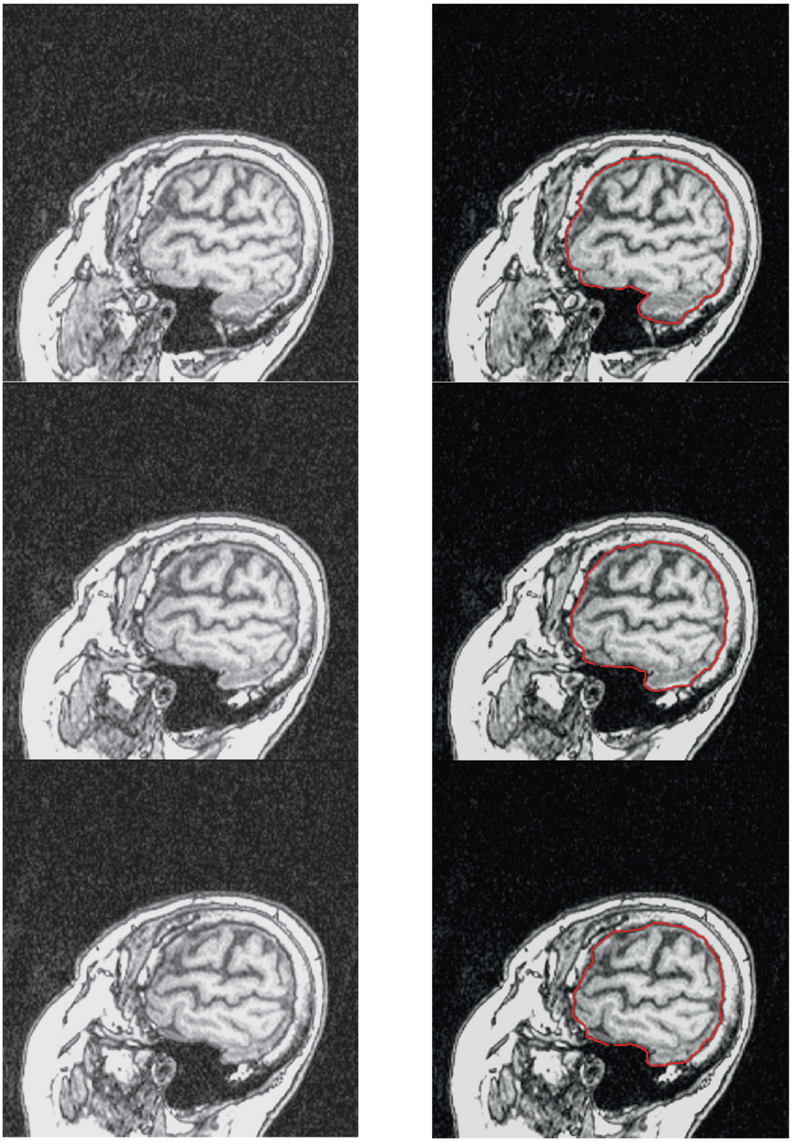
n) foramen lacerum “open”

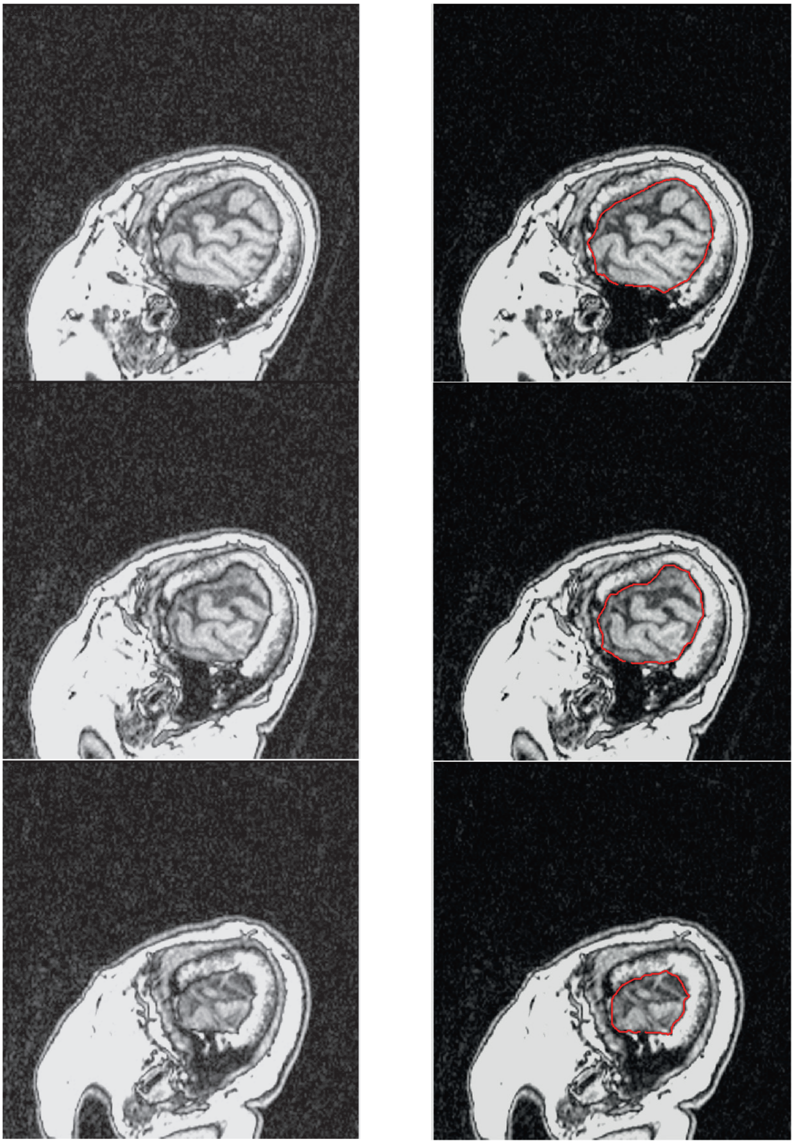


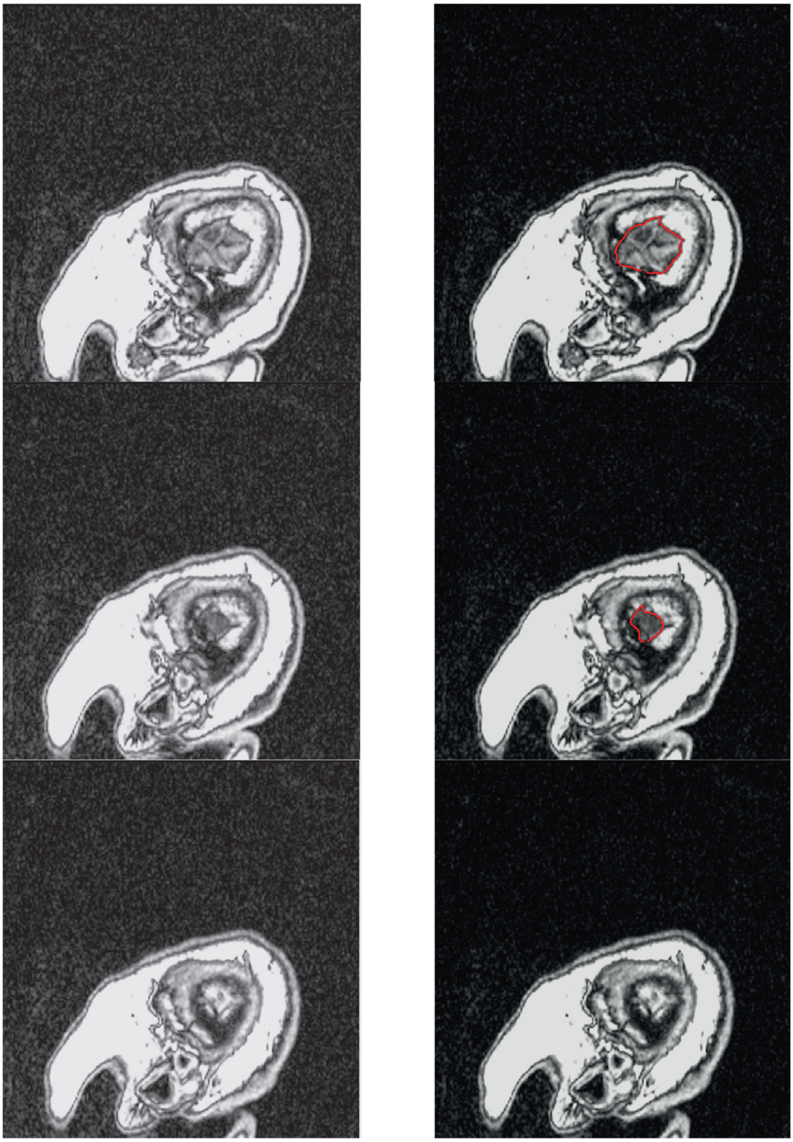












B MATHEMATICAL PROOFS

Here, I will show mathematical proofs for the composite functions in Table 7 (Section 3.11) that give point estimates in the cases of least-squares and inferred least-squares normalization. The proofs relies on the following properties:

Properties of variance (s^2) where X and Y are variables and α and β constants

$$s^2_{(\alpha X+\beta Y)} = \alpha s^2_X + \beta s^2_Y + 2\alpha\beta cov(X,Y)$$

$$s^2_{(\alpha X-\beta Y)} = \alpha s^2_X + \beta s^2_Y - 2\alpha\beta cov(X,Y)$$

$$s^2_\alpha = 0$$

Properties of covariance (cov) where X and Y are variables and α and β constants

$$cov(\alpha X, \beta Y) = \alpha\beta cov(X,Y)$$

$$cov(X,X) = s^2_X$$

$$cov(X,\alpha) = 0$$

Properties of the regression coefficient (k) from a simple linear regression and the Pearson's correlation ($r_{X,Y}$) from the same regression analysis. X and Y are variables.

$$k = \frac{cov(X,Y)}{s^2_X} = \frac{r_{X,Y}s_Y}{s_X}$$

$$r_{X,Y} = \frac{cov(X,Y)}{s_X s_Y} = \frac{ks_X}{s_Y}$$

Proofs for the least-squares normalization composite functions

When using least-squares normalization, the values of the brain estimates in the sample (b) after normalizing by intracranial volume (icv) will be

$$b_{norm} = b - k(icv - \overline{\imath cv})$$

Here b_{norm} is the normalized brain estimates and $\overline{\imath cv}$ the mean icv in the sample. Then the mean of the normalized brain estimates is

$$\overline{b_{norm}} = \overline{b - k(\imath cv - \overline{\imath cv})}$$

$$\overline{b_{norm}} = \bar{b} - \overline{k(\imath cv - \overline{\imath cv})}$$

$$\overline{b_{norm}} = \bar{b} - k(\overline{\imath cv} - \overline{\imath cv})$$

$$\overline{b_{norm}} = \bar{b} - k(0)$$

$$\overline{b_{norm}} = \bar{b}$$

The variance (s^2) of the normalized brain estimates is

$$s^2_{b_{norm}} = s^2_{b-k(icv-\overline{\imath cv})}$$

$$s^2_{b_{norm}} = s^2_b + s^2_{k(icv-\overline{\imath cv})} - 2cov(b, k(icv - \overline{\imath cv}))$$

$$s^2_{b_{norm}} = s^2_b + k^2 s^2_{(icv-\overline{\imath cv})} - 2kcov(b, (icv - \overline{\imath cv}))$$

$$\begin{aligned} s^2_{b_{norm}} = s^2_b + k^2\big(s^2_{icv} + s^2_{\overline{\imath cv}} - 2cov(icv, \overline{\imath cv})\big) \\ - 2k(cov(b, icv) - cov(b, \overline{\imath cv})) \end{aligned}$$

$$s^2_{b_{norm}} = s^2_b + k^2\big(s^2_{icv} + 0 - 0\big) - 2k(cov(b, icv) - 0)$$

$$s_{b_{norm}}^2 = s_b^2 + k^2 s_{icv}^2 - 2kcov(b, icv)$$

$$s_{b_{norm}}^2 = s_b^2 + \left(\frac{r_{b,icv} s_b}{s_{icv}}\right)^2 s_{icv}^2 - 2\left(\frac{r_{b,icv} s_b}{s_{icv}}\right) cov(b, icv)$$

$$s_{b_{norm}}^2 = s_b^2 + r_{b,icv}^2 s_b^2 - 2\left(\frac{r_{b,icv} s_b}{s_{icv}}\right) r_{b,icv} s_b s_{icv}$$

$$s_{b_{norm}}^2 = s_b^2 + r_{b,icv}^2 s_b^2 - 2r_{b,icv}^2 s_b^2$$

$$s_{b_{norm}}^2 = s_b^2 - r_{b,icv}^2 s_b^2$$

The Pearson’s correlation ($r_{X,Y}$) between the normalized brain estimates and icv is

$$r_{b_{norm},icv} = \frac{cov(b_{norm}, icv)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b - k(icv - \overline{icv}), icv)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b, icv) - cov(k(icv - \overline{icv}), icv)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b, icv) - kcov((icv - \overline{icv}), icv)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b, icv) - k(cov(icv, icv) - cov(\overline{icv}, icv))}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b, icv) - k\left(s_{icv}^2 - 0\right)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b, icv) - ks_{icv}^2}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b,icv) - \left(\frac{cov(b,icv)}{s_{icv}^2}\right) s_{icv}^2}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = \frac{cov(b,icv) - cov(b,icv)}{s_{b_{norm}} s_{icv}}$$

$$r_{b_{norm},icv} = 0$$

Proofs for the inferred least-squares normalization composite functions

For inferred least-squares normalization, the regression coefficient (k) is calculated from one subsample and then used to normalize a second subsample. It is not certain that the regression coefficients between the two subsamples are equal to each other and we therefore denote them as k_1 and k_2 in the following equations. Then the values of the normalized brain estimates of the second subsample will be

$$b_{norm} = b_2 - k_1(icv_2 - \overline{\imath cv_1})$$

Here b_2 are the unnormalized brain estimates from the second sample, k_1 is the regression coefficient calculated from the first sample, icv_2 is the intracranial volume estimates from the second subsample, and $\overline{\imath cv_1}$ is the mean intracranial volume from the first sample (from which k_1 was calculated). Then, the mean of the inferred least-squares normalized brain estimates is

$$\overline{b_{norm}} = \overline{b_2 - k_1(\imath cv_2 - \overline{\imath cv_1})}$$

$$\overline{b_{norm}} = \overline{b_2} - k_1(\overline{\imath cv_2} - \overline{\imath cv_1})$$

$$\overline{b_{norm}} = \overline{b_2} - \left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{ICV_1}}\right)(\overline{\imath cv_2} - \overline{\imath cv_1})$$

The variance of the inferred least-squares normalized brain estimates is

$$s^2_{b_{norm}} = s^2_{b_2 - k_1(icv_2 - \overline{icv_1})}$$

$$s^2_{b_{norm}} = s^2_{b_2} + s^2_{k_1(icv_2 - \overline{icv_1})} - 2cov(b_2, k_1(icv_2 - \overline{icv_1}))$$

$$s^2_{b_{norm}} = s^2_{b_2} + k_1^2 s^2_{icv_2 - \overline{icv_1})} - 2k_1 cov(b_2, (icv_2 - \overline{icv_1}))$$

$$s^2_{b_{norm}} = s^2_{b_2} + k_1^2\left(s^2_{icv_2} + s^2_{\overline{icv_1}} - 2cov(icv_2, \overline{icv_1})\right) \\ - 2k_1(cov(b_2, icv_2) - cov(b_2, \overline{icv_1}))$$

$$s^2_{b_{norm}} = s^2_{b_2} + k_1^2\left(s^2_{icv_2} + 0 - 0\right) - 2k_1(cov(b_2, icv_2) - 0)$$

$$s^2_{b_{norm}} = s^2_{b_2} + k_1^2 s^2_{icv_2} - 2k_1 cov(b_2, icv_2)$$

$$s^2_{b_{norm}} = s^2_{b_2} + k_1^2 s^2_{icv_2} - 2k_1 cov(b_2, icv_2)$$

$$s^2_{b_{norm}} = s^2_{b_2} + \left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right)^2 s^2_{icv_2} - 2\left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right) r_{b_2,icv_2} s_{b_2} s_{icv_2}$$

$$s^2_{b_{norm}} = s^2_{b_2} + \left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right)^2 s^2_{icv_2} - 2\left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right) r_{b_2,icv_2} s_{b_2} s_{icv_2}$$

$$s^2_{b_{norm}} = s^2_{b_2} + \frac{r^2_{b_1,icv_1} s^2_{b_1} s^2_{icv_2}}{s^2_{icv_1}} - \frac{2r_{b_1,icv_1} r_{b_2,icv_2} s_{b_1} s_{b_2} s_{icv_2}}{s_{icv_1}}$$

And the Pearson's correlation between the normalized brain estimates and icv_2 is

$$r_{b_{norm},icv_2} = \frac{cov(b_{norm}, icv_2)}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2 - k_1(icv_2 - \overline{icv_1}), icv_2)}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2, icv_2) - cov(k_1(icv_2 - \overline{icv_1}), icv_2)}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2, icv_2) - k_1 cov((icv_2 - \overline{icv_1}), icv_2)}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2, icv_2) - k_1(cov(icv_2, icv_2) - cov(\overline{icv_1}, icv_2))}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2, icv_2) - k_1\left(s_{icv_2}^2 - 0\right)}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{cov(b_2, icv_2) - k_1 s_{icv_2}^2}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{r_{b_2,icv_2} s_{b_2} s_{icv_2} - \left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right) s_{icv_2}^2}{s_{b_{norm}} s_{icv_2}}$$

$$r_{b_{norm},icv_2} = \frac{r_{b_2,icv_2} s_{b_2} - \left(\frac{r_{b_1,icv_1} s_{b_1}}{s_{icv_1}}\right) s_{icv_2}}{s_{b_{norm}}}$$

$$r_{b_{norm},icv_2} = \frac{\left(\frac{r_{b_2,icv_2} s_{b_2} s_{icv_1}}{s_{icv_1}}\right) - \left(\frac{r_{b_1,icv_1} s_{b_1} s_{icv_2}}{s_{icv_1}}\right)}{s_{b_{norm}}}$$

$$r_{b_{norm},icv_2} = \frac{r_{b_2,icv_2} s_{b_2} s_{icv_1} - r_{b_1,icv_1} s_{b_1} s_{icv_2}}{s_{icv_1} s_{b_{norm}}}$$

As we know $s_{b_{norm}}^2$ from the above proof, we can replace $s_{b_{norm}}$ and get

$$r_{b_{norm},icv_2} = \frac{r_{b_2,icv_2} s_{b_2} s_{icv_1} - r_{b_1,icv_1} s_{b_1} s_{icv_2}}{s_{icv_1} \sqrt{s_{b_2}^2 + \frac{r_{b_1,icv_1}^2 s_{b_1}^2 s_{icv_2}^2}{s_{icv_1}^2} - \frac{2 r_{b_1,icv_1} r_{b_2,icv_2} s_{b_1} s_{b_2} s_{icv_2}}{s_{icv_1}}}}$$