Author's response to reviews

Title: Valid and efficient manual estimates of intracranial volume from magnetic resonance images

Authors:

Niklas Klasson (niklas.klasson@neuro.gu.se)
Erik Olsson (erik.olsson@neuro.gu.se)
Mats Rudemo (rudemo@chalmers.se)
Carl Eckerström (carl.eckerstrom@neuro.gu.se)
Helge Malmgren (helge.malmgren@filosofi.gu.se)
Anders Wallin (anders.wallin@neuro.gu.se)

Version: 4  Date: 8 September 2014

Author's response to reviews: see over
September 8, 2014

Dear editor and reviewers,

We are very thankful for the effort put down in reviewing our manuscript and think that the comments have helped us to improve the manuscript. Below is our response to the comments of reviewer 1, and then of reviewer 2.

Best regards,

Niklas Klasson

Response to reviewer 1:

Major

1. We agree that age should be given in the paper, but your comment also made us consider if not the most common demographic variables of the participants should be given as well. To fully appreciate the material of the paper, and maybe when comparing it to other studies, this could be of importance. For this reason not only age was added, but also education and mini-mental state score. The following changes were made:

   Sentence a) was added at section 2.1 at the end of paragraph 2:

   a) The complete demographics of the remaining participants can be viewed in Table 1.

   Table 1 (see the manuscript) with demographic data was added at the end of the manuscript.

2. The text “, Spacing Between Slices = 1” was removed from section 2.2.

3. In section 2.3 sentence a) was changed to b):

   a) The MRI scans were reformatted into 1 mm cubic voxels and were sagittally oriented before segmentations were performed.

   b) The MRI scans were reformatted into 1 mm cubic voxels using linear interpolation, but no realignment of the images was done to correct for head tilt.
4. Section 2.4 was changed so it now hopefully is easier to see why \( n^x \) combinations of estimates are evaluated when there really is just \( n \) times \( x \) unique estimates. In section 2.4 the sentences in a) was changed to the sentences in b):

a) Thus, using a linear spacing of \( n \) results in \( n \) possible subsamples of ICAs and therefore \( n \) possible estimates for each ICV. As there were 62 ICVs in the present study, all the possible estimates from each ICV can be combined in \( n^{62} \) (e.g. \( 10^{62} \) if \( n = 10 \)) different ways. It is practically impossible to evaluate all of these possible combinations and therefore 2000 combinations were chosen randomly for each linear spacing. Practically a combination of ICA subsamples was constructed by randomly choosing the position of the first ICA for each of the 62 ICVs. By repeating this process 2000 times, 2000 combinations of ICA subsamples were constructed.

b) Thus, using a linear spacing of \( n \) results in \( n \) possible subsamples of ICAs and therefore \( n \) possible estimates for each of the 62 ICV segmentations. To calculate the validity of the estimates, combinations containing one estimate for each ICV segmentation were used. Such a combination can be derived in \( n_{ICV1} \cdot n_{ICV2} \cdot n_{ICV3} \ldots \cdot n_{ICV62} \) different ways (the multiplication principle). The validity of each of the combinations depends on which estimates were chosen, and then not only on the linear spacing, but also on which positions were used for the first ICAs. Therefore, to describe the validity of estimates only due to linear spacing, independently of the position of the first ICA, the validity of all possible combinations must be described. As it is practically impossible to evaluate all of these \( n^{62} \) combinations of estimates, 2000 combinations were chosen randomly for each linear spacing. The randomization was done by randomly choosing one estimate out of the \( n \) possible for each ICV segmentation to construct one combination, and then this procedure was repeated 2000 times.

5. We do not think that similarity/overlap analyses would add enough value to this particular study to be included. The correlation is important for normalization purposes and the actual volume at least for descriptive purposes, but the similarity index is not useful as such. If the aim had been to create estimates as similar as possible to the manual segmentations of the intracranial volumes, then similarity analyses would have been a good approach.

Still, as you point out, a similarity analysis could be used to highlight the differences between the interpolation methods, and to partly describe why the cubic spline interpolation results in better validities than the piecewise constant interpolation. However, this is to some extent already done by Figure 6 (previously Figure 4). While not as thorough as similarity analyses we think Figure 6 is good enough; it is relatively easy to grasp and does not take too much attention from the other findings.

6. An illustration of the sagittal segmentation along with the coronal and transversal reconstructions was added (as Figure 2). Sentence a) was added at section 2.3 at the end of paragraph 4:

a) An illustration of the initial sagittal segmentation and its coronal and transversal reconstructions can be viewed in Figure 2.
Further, Figure 5 was added where intracranial volumes are shown in mm$^3$ for each participant (plotted against age), for males and females, and for controls and patients.

The suggested age and gender test was also added to the manuscript with an addition of a test of relation between intracranial volume and GDS score (atrophy should not affect the intracranial volume). The following changes were made:

Paragraph b) was added at section 2.5:

b) As a way of evaluating the ICV segmentations, the association of ICV to age, gender and group belonging (controls, GDS 2, 3 or 4) was analyzed. The possible correlation between ICV and age was evaluated using Pearson’s linear correlation, gender difference was evaluated using an independent-samples t-test, and the difference due to group belonging by a Kruskal-Wallis test. As the male to female ratio differed between controls and patients, the Kruskal-Wallis test was also performed for males and females separately.

Paragraph c) was added at section 3 (along with Figure 5):

c) No correlation between the ICV and age could be seen ($p = 0.376$). Further, the distributions of the ICVs seem to be the same regardless of group belonging ($p_{all} = 0.977$, $p_{females} = 0.458$, $p_{males} = 0.672$), while a difference in mean ICV could be seen by gender ($p < 0.001$). In Figure 5 the ICVs are displayed in relation to age, group belonging and gender.

Paragraph d) was added at section 4:

d) One of the most important aspects of measures of intracranial volume is that they should not be affected by atrophy. The ICVs in the present study were segmented following the dura mater to make sure not to include a bias due to atrophy in the findings. There was no difference in ICV due to group belonging (controls, GDS 2, 3 or 4) or any correlation to age, which speaks against a possible bias in the manual segmentations. The highly similar results for the piecewise constant interpolation to those of Eritaia et al. [4], that only included normal controls, also indicate that atrophy has not affected the present results.

Minor

1. It might not have been clear, but the present study is just a part of the Gothenburg MCI study. Sentence a) was changed to b) in the first paragraph of section 2.1 to make this clearer:

a) The present study is a substudy of the Gothenburg MCI (mild cognitive impairment) study from which 38 patients and 32 controls with 1.5 T MRI scans were included.

b) The present study is part of the Gothenburg MCI (mild cognitive impairment) study from which a subsample of 38 patients and 32 controls with 1.5 T MRI scans was included.
The sample of participants in the present study is not representative for the larger Gothenburg MCI study with over 700 participants. Which participants were included in the present study should not matter for the purpose of it. The manual method should be equally good regardless of whether the cranial vault belongs to a healthy control, someone with mild cognitive impairment or possible dementia, as the dura mater was followed during the segmentations of the intracranial vaults.

2. The suggested change was made.

3. The use of capitals was changed.

4. Sentence a) was changed to the sentences in b) at section 2.3 in the second paragraph:

   a) Further, a gamma correction with gamma set to 0.8 was performed.

   b) Further, the images were made brighter by a gamma correction with gamma set to 0.8. The gamma correction is performed using the function $y = b \times (x/b)^\gamma$ where $y =$ output data, $b =$ bit depth, $x =$ input data, and $\gamma =$ gamma value.

5. See answers to comment 3 and 4 of the second reviewer.

6. The suggested change was made.

7. In the third paragraph of section 2.3, sentence a) was changed to b):

   a) The mean number of sagittal slices was 135.9 per MRI scan and the segmentations took on average a little over two and a half hours per scan. N. Klasson performed the segmentations.

   b) The mean number of sagittal slices was 136 per MRI scan with a standard deviation of 5 slices, and the segmentations took on average a little over two and a half hours per scan. N. Klasson performed the segmentations.

8. In the fifth paragraph of section 2.3, sentence a) was changed to b):

   a) Both raters were blinded to participant age, gender, and cognitive status during all ICV segmentations.

   b) Both raters were blinded to participant age, gender, and cognitive status as well as to previous segmentations during all ICV segmentations.

9. Figure 3 (new figure number) shows the percentage errors referred to in the first paragraph of section 2.5. In the second paragraph the mean absolute percentage errors used to compare the piecewise constant interpolation to the cubic spline interpolation in Figure 4 (new figure number) are referred to. As the piecewise linear interpolation was not compared to the piecewise constant interpolation the mean absolute percentage errors were not calculated for the estimates of the linear interpolation. As we felt that the second paragraph in
section 2.5 might be hard to follow, we tried to improve it by changing the whole section from a) to b).

a) For each linear spacing and orientation the mean absolute percentage errors were calculated for all possible estimates (= n) at each ICV, using both piecewise constant interpolation and cubic spline interpolation. This resulted in 62 mean absolute percentage errors, one for each ICV, at each linear spacing, orientation and interpolation method. For each linear spacing and orientation these mean absolute percentage errors were then compared for the two interpolation methods using Student’s t-test for paired data.

b) To evaluate if the percentage errors of the cubic spline interpolation differed significantly from those of the piecewise constant interpolation, the mean absolute percentage error of the n estimates of each ICV segmentation was calculated for each linear spacing and orientation. These mean absolute percentage errors were then compared between the two interpolation methods for each linear spacing and orientation using Student’s t-tests for paired data.

10. The suggested change was made.

11. To emphasize the use of reconstructed intracranial areas, sentence a) was changed to b) in the conclusions:

a) At larger linear spacings cubic spline interpolation on coronal or sagittal ICAs resulted in the most valid estimates.

b) At larger linear spacings cubic spline interpolation on sagittal or reconstructed coronal ICAs resulted in the most valid estimates.

12. Sentence a) was changed to b) to clarify the kind of correlation:

a) Nordenskjöld et al. found that FreeSurfer had a correlation of 0.937 with manual ICV measures while SPM had a correlation of 0.856.

b) Nordenskjöld et al. found that FreeSurfer had a Pearson’s correlation of 0.937 to manual ICV segmentations while SPM had a Pearson’s correlation of 0.856 (correlations calculated by the square root of $R^2$ from univariate linear regression models).

13. We do not agree that Pearson’s correlation is suboptimal as the correlation is one of the most important aspects of estimates of intracranial volume due to their frequent use for normalization purposes.

The study by Nordenskjöld et al. does not contain ICC nor similarity indexes why we could not compare ICCs. The reason why we ourselves do not use similarity indexes is clarified in comment 5 above.

14. We agree with your point. While there are other automatic methods for ICV assessment methods than the one suggested, the study by Keihaninejad et al. is important to discuss as the ICCs (of 0.98-0.99) indicate that those methods are reasonable alternatives to the manual methods evaluated in the present
manuscript. We therefore added paragraph a) in section 4, while sentence b) was removed from the same section:

a) Beside the commonly used FreeSurfer and SPM, there might exist automatic methods for ICV assessment that perform as well as manual segmentations of every 50th mm. For example, Keihaninejad et al. showed promising results in that two out of four evaluated automatic methods had ICCs of 0.98-0.99 to manual estimates of ICV [12]. While the evaluations were done for T1-weighted images from both 1.5 T and 3 T scanners, the comparisons for each scanner were done using manual segmentations from only five healthy participants. As for many automatic ICV estimation methods, more thorough evaluations are needed, including the important aspect that they should not be biased by atrophy.

b) Further, a degree of precision not achievable with automatic methods might sometimes be highly desirable and manual segmentation with small linear spacing ought then to be considered.

15. The suggested change was performed.

16. The use of abbreviations in the reference list was adjusted.

17. Units were added to all y-axes.

Discretionary revisions

18. We do not think that adding the complicated formula of the ICC to the manuscript would be a good idea. If the reader is interested in the ICC formula they can follow the used nomenclature in the referenced paper by McGraw and Wong. The nomenclature used is also consistent with the commonly used software package SPSS. Further, the reader can get the Matlab code of the ICC function used in the manuscript by following the homepage link at the end of the section 2.5.

19. The suggested change was made.

20. The suggested change was made.

21. While the contrast is not optimal, we do find it acceptable. One of the benefits with the color theme chosen is that the colors are separable also when the figures are printed in black and white (besides being restful for the eye).

22. We are thankful for your honesty and thoughts regarding the subject. Skau has put a lot of effort into learning the segmentation procedure and we really appreciate his work, but at the same time we do not think that the segmentations alone are enough of a scientific contribution to add Skau as an author.

Response to reviewer 2:
1. The sample of participants from the Gothenburg MCI study was used because the main author is a PhD student at the research group conducting the Gothenburg MCI study.

The Gothenburg MCI study has quite a large number of publications and we do not think that a review of this research should be added to the current manuscript. Sadly there is yet no review of the Gothenburg MCI study that we could refer to either. Reference 14 in the first paragraph of section 2.1 in the manuscript is the best description of the study at the moment. A review of the Gothenburg MCI study is however in progress and will hopefully be published within a year.

2. It is true that the different GDS stages differ in the neuroradiological findings, and that GDS 4 stage might vary a lot depending on what kind of patients that are present. While atrophy, white matter lesions and such should not affect the intracranial volumes, this is still, as pointed out, important to establish in the present study. The following changes were done in the manuscript:

Paragraph a) was added to section 2.5:

a) As a way of evaluating the ICV segmentations, the association of ICV to age, gender and group belonging (controls, GDS 2, 3 or 4) was analyzed. The possible correlation between ICV and age was evaluated using Pearson’s linear correlation, gender difference was evaluated using an independent-samples t-test, and the difference due to group belonging by a Kruskal-Wallis test. As the male to female ratio differed between controls and patients, the Kruskal-Wallis test was also performed for males and females separately.

Paragraph b) was added to section 3, and Figure 5 was included in the manuscript:

b) No correlation between ICV and age could be seen ($p = 0.376$). Further, the distributions of the ICVs seem to be the same regardless of group belonging ($p_{all} = 0.977$, $p_{females} = 0.458$, $p_{males} = 0.672$), while a difference in mean ICV could be seen by gender ($p < 0.001$). In Figure 5 the ICVs are displayed in relation to age, group belonging and gender.

Paragraph c) was added to section 4:

c) One of the most important aspects of measures of intracranial volume is that they should not be affected by atrophy. The ICVs in the present study were segmented following the dura mater to make sure not to include a bias due to atrophy in the findings. There was no difference in ICV due to group belonging (controls, GDS 2, 3 or 4) or any correlation to age, which speaks against a possible bias in the manual segmentations. The highly similar results for the piecewise constant interpolation to those of Eritaia et al. [4], that only included normal controls, also indicate that atrophy has not affected the present results.

As it is very unlikely that the manual segmentations of intracranial volumes are affected by atrophy the interpolations won’t be either. E.g. think of the piecewise
constant interpolation where the intracranial areas are just multiplied by the linear spacing between them. This interpolation method thus just multiplies the input with a constant. The constant won’t be affected by atrophy (and therefore neither the interpolation method), but if the input would be biased by atrophy then so would the output. Even if slightly more complicated, the same is true for the other interpolation methods. Therefore, we do not think it is necessary to evaluate the different interpolation methods for the different GDS stages (as the input should not be, and does not seem to be, affected by atrophy).

3. No realignment to correct for head tilt was done. The following changes was done in the manuscript regarding this:

   In section 2.2, sentence a) was added:
   
a) The coronal planes were aligned perpendicularly to the longitudinal axis of hippocampus during the MRI examinations.
   
   In the second paragraph of section 2.3, sentence b) was changed to c):
   
b) The MRI scans were reformatted into 1 mm cubic voxels and were sagittally oriented before segmentations were performed.
   
c) The MRI scans were reformatted into 1 mm cubic voxels using linear interpolation, but no realignment of the images was done to correct for head tilt.
   
   At the fifth paragraph in section 4, d) was added:
   
d) Further, the MRIs were not realigned to correct for head tilt. While there barely were any lateral tilts of the heads, the coronal plane had been aligned during the MRI examinations so that it was perpendicular to the longitudinal axis of hippocampus (a slight backward tilt of the head). This alignment of the coronal plane makes the result for the transversal and coronal orientation less generalizable for studies where a similar alignment is not used. The generalizability should still be high, but less so for large linear spacings in these orientations.

4. In the third paragraph of section 2.3, sentence a) was changed to b):
   
a) The dural margins were traced using the landmarks described by Eritaia et al [4].
   
b) The segmentations were performed in sagittal orientation by tracing the dural margins using the landmarks described by Eritaia et al [4].

5. Matlab was used, as the first author is an experienced Matlab user, and as a lot of statistical analyses had to be done, e.g. 882000 ICCs and Pearson’s correlations (3 interpolation methods * 3 ICA orientations * 49 linear spacings * 2000 combinations). All different statistical tests were at least once performed in both Matlab and SPSS (version 19) to make sure that the Matlab functions produced the same results as SPSS.

Regarding the specific statistical tests, ICCs with absolute agreement were used as such are used in the study by Eritaia et al.. Pearson’s correlations are often
used in other studies and were included partially due to this reason. Further, the Pearson's correlations also give information about the correlations alone, which often is of more interest when ICVs only should be used for normalization purposes. Why similarity analyses not were included, see answer to comment 5 (in the Major section) of the first reviewer.