Statistical Impact of Referencing on Quantitative Susceptibility Mapping

Patrick Fuchs* P.FUCHS@UCL.AC.UK
Department of Medical Physics and Biomedical Engineering
University College London, London, UK

Oliver C. Kiersnowski O.KIERSNOWSKI@UCL.AC.UK
Department of Medical Physics and Biomedical Engineering
University College London, London, UK

Carlos Milovic CARLOS.MILOVIC@PUCV.CL
School of Electrical Engineering
Pontificia Universidad Catolica de Valparaiso, Valparaiso, Chile

Karin Shmueli K.SHMUELI@UCL.AC.UK
Department of Medical Physics and Biomedical Engineering
University College London, London, UK

Abstract

Quantitative magnetic susceptibility mapping is in practice actually relative susceptibility mapping. Due to the nature of the propagation of susceptibility effects into the measured signal (through convolution with a dipole kernel) the average susceptibility of tissue is lost. One approach to deal with this is to reconstruct the susceptibility with respect to the average susceptibility (i.e. setting the mean to zero). Another way of dealing with this unknown baseline is to set the susceptibility within a specific region of interest (persumably of known susceptibility) to zero, thereby referencing the rest of the susceptibility map to this tissue type. In this study we take a look at a clinical epilepsy study and evaluate its results with and without explicit referencing. We show that referencing explicit anatomy may introduce a bias in the presence of pathology, which is not as apparent when referencing more general regions or implicitly referencing the whole brain.

Keywords: QSM; Referencing, Clinical studies, Statistical Analysis, Regions of Interest

1 Introduction

The magnetic susceptibility (χ) is related to the measured gradient echo phase through a convolution with a magnetic dipole kernel . Inverting this signal model is known as quantitative susceptibility mapping (QSM), and, over the years, many different approaches have been published to achieve this. QSM reconstructs a relative map of bulk magnetic susceptibility of the tissue.

It is not straightforward to invert the signal model due to the unit magnetic dipole kernel containing a double cone-shaped zero-valued surface, which leads to an undefined inverse. This ill-posed problem is typically solved through the application of regularisation, both implicit as well as explicit. In 2019 a QSM reconstruction challenge [1] was launched to compare various reconstruction approaches to a ground truth (simulated) dataset, which enabled quantitative metrics to be applied. The metrics applied in the challenge are predominantly global image quality metrics, such as normalised root mean squared error (nRMSE), or the susceptibility tuned similarity metric XSIM [2]. Pathological changes in magnetic susceptibility are, however, typically constrained to a local region (e.g. specific anatomy in the brain, or regions of interest (ROI)). We have recently investigated the impact of reconstruction algorithm choice on isolated regions of interest within the brain [3] , such as e.g., deep brain gray matter (substantia nigra, red nucleus, thalamus), using the reconstruction challenge results. Since clinical applications of QSM would typically use a t-test or similar metric to determine whether pathological changes are significant we also apply this test on this challenge analysis.

In QSM, due to the dipole kernel containing a zero at the origin in k-space, the scale of the DC offset or zeroeth frequency component is unknown. This is equivalent to having an unknown or undetermined mean value. Therefore, the reconstruction is equally valid for any scalar offset, which would in turn be a value to which the reconstructed susceptibilities are referenced. Choosing the right reference can be challenging, and a number of studies investigated potential reference regions [4, 5, 6]. There are two commonly accepted approaches to perform this referencing [7, 8, 5]. The first is to reference the mean of the reconstructed volume to 0. The second involves computing the susceptibility in a region which is of known uniform susceptibility and referencing to this region. The region most often considered is the cerebrospinal fluid (CSF), which is expected to have homogeneous isotropic susceptibility like water due to its (mostly) fluid nature.

In our analysis of the challenge phantom, we have found that retrospectively referencing to the (perceived) CSF value can lead to different statistical results , improving some methods while others score markedly worse. Following this observation, we present here a statistical analysis of referencing, and investigate the influence referencing has on a clinical study. This is substantiated by presenting a clinical study on temporal lobe epilepsy [9] and evaluating the effects of referencing on this dataset. Finally, we present a novel referencing approach based not on an anatomical ROI but rather on the R_2^* values of the tissue.

2 Theory

For this analysis we assume our samples are mean values accross an ROI, and a reference ROI for n healthy controls and (m) patients, i.e.,

$$X_1, X_2, \ldots, X_n \quad \text{random sample from } N(\mu_{\mathrm{HC}}, \sigma^2_{\mathrm{HC}})$$

$$R^x_1, R^x_2, \ldots, R^x_n \quad \text{random sample from } N(\mu_{\text{ref;HC}}, \sigma^2_{\text{ref;HC}})$$
$$Y_1, Y_2, \ldots, Y_m \quad \text{random sample from } N(\mu_{\text{patient}}, \sigma^2_{\text{patient}})$$
$$R^y_1, R^y_2, \ldots, R^y_m \quad \text{random sample from } N(\mu_{\text{ref;patient}}, \sigma^2_{\text{ref;patient}})$$

With $N(\text{mean}, \text{variance})$ a normal distribution. Here μ denotes the mean of the underlying distributions ("HC" for healthy control and "patient" for the patient cohort), and σ^2 the variance. If we now want to test whether the susceptibility values found in the patient cohort are statistically different from the healthy controls, our hypothesis without referencing would be

$$H_0 : \mu_{\text{HC}} = \mu_{\text{patient}},$$

whereas in the referenced case it would read as

$$H^{\text{ref}}_0 : \mu_{\text{HC}} - \mu_{\text{ref;HC}} = \mu_{\text{patient}} - \mu_{\text{ref;patient}}.$$

An important question already arises from this formulation. Are these two tests equivalent (and should they be)? Two cases can readily be identified, one where the reference volume is (across subjects) uncorrelated with disease, i.e. $\mu_{\text{ref;HC}} = \mu_{\text{ref;patient}}$. In this case the two hypothesis tests are equal, and the same question is asked. In the second case, when the reference mean is unequal between healthy control and patient datasets the test is a different one, and a different statistical question is answered by it.

If $\mu_{\text{ref;HC}} \neq \mu_{\text{ref;patient}}$ that means that there is a disease related change to the reference volume's magnetic susceptibility. This would bias the resulting susceptibility maps, or in other words, it changes the hypothesis test you are performing.

Even though the test hypothesis may be equivalend provided $\mu_{\text{ref;HC}} = \mu_{\text{ref;patient}} \equiv \mu_{\text{ref}}$, the test statistic can change due to referencing. A t-test test statistic for two samples with unequal variance is

$$T_d = \frac{\bar{x}_n - \bar{y}_m}{S_d},$$

where $\bar{x}_n$ and $\bar{y}_m$ are the sample means of the healthy controls and patient cohort respectively, and S_d is the non-pooled variance (unbiased estimator for $\text{Var}(\bar{X}_n - \bar{Y}_m)$ for two samples with unequal variances).

$$S^2_d = \frac{S^2_X}{n} + \frac{S^2_Y}{m}.$$

Here S^2_X and S^2_Y are unbiased estimators for σ^2_X and σ^2_Y respectively.

The sample mean after referencing is the difference in sample means between the original values and the references, that is

$$\bar{x}^r_n = \bar{x}_n - \bar{r}_n.$$

The sample variance however changes according to Bienaymé's identity

$$\text{Var}\left(X - R^x\right) = \text{Var}(X) + \text{Var}(R^x) + 2\text{Cov}(X, -R^x).$$

If the reference regions are uncorrelated with the susceptibility ROI of interest ($\text{Cov}(X, -R^x) = 0$) the variance of the referenced test is guaranteed to be larger or equal to the original test. This holds for the variance of the referenced patient samples in

the same fashion. In other words, referencing with an uncorrelated reference region increases the variance (S_d) used in our test statistic. Note that a larger variance decreases the test statistic, thus decreasing the effect size and statistical strength of the test.

Important to note here is that the correlation we are looking at here is not between the groups (i.e. between the healthy and disease cohort), but rather within the group. For example, the CSF is often used as a reference because it is a relatively large and uniform region with little variance inside a volume reconstruction. The CSF between different subjects however can have much more pronounced variance.

An important reason for referencing is that it provides a baseline across subjects. It is therefore natural to expect a certain amount of correlation between the reference region and unreferenced susceptibility values. Strictly speaking the covariance is bounded by

$$-\sqrt{\mathrm{Var}(X)\mathrm{Var}(R^x)} \leq \mathrm{Cov}(X, R) \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(R^x)},$$

which we can simplify if we assume equal variance of test ROI and reference ROI ($\sigma^2_{\mathrm{HC}} = \sigma^2_{\mathrm{ref;HC}} = \sigma^2$) to

$$-\sigma^2 \leq \mathrm{Cov}(X, -R^x) \leq \sigma^2,$$

which leads to a referenced variance of

$$\mathrm{Var}\left(X - R^x\right) = \sigma^2_{\mathrm{HC}} + \sigma^2_{\mathrm{ref;HC}} + 2\mathrm{Cov}(X, -R^x)$$

$$\mathrm{Var}\left(X - R^x\right) = 2\sigma^2 + 2\mathrm{Cov}(X, -R^x).$$

This bounds the referenced variance to

$$0 \leq \mathrm{Var}\left(X - R^x\right) \leq 4\sigma^2,$$

that is, to be precise, the variance could increase fourfold or decrease to 0 (in which case the reference region can be shown to be equal to the region of interest, resulting in a constant test value). As previously mentioned, a certain minor amount of positive correlation between reference value and ROI of interest can be expected, especially since the referencing is supposed to provide a more stable baseline than no reference, that is if there is no correlation it would also not be a reasonable baseline. This does mean that final variance is expected to be less than the sum of the separate variances, though in our experiments we have found it to generally be slightly higher than the original variance.

3 Methods

We will investigate referencing on a synthetic dataset (from the 2019 QSM challenge [1]) that has a ground truth, as well as on a cohort of healthy controls and patients with temporal lobe epilepsy originally investigated by Kiersnowski et al. [9]. In the QSM challenge dataset we are working with a single subject and many different reconstructions. This means that we cannot compare in a groupwise fashion but are rather limited to between subject (reconstruction) comparisons. The epilepsy cohort on the other hand contains multiple subjects with both left and right temporal lobe epilepsy (LTLE and RTLE respectively) subjects, as well as healthy controls (HC).

A drawback of being limited to a single subject in the synthetic dataset is that with all the reconstructions being based on the same data they cannot be treated as independent measurements. This means that any groupwise statistics would be flawed. It is however possible to compare the reconstructions to the ground truth, performing the same t-test as described above, considering the voxels within an ROI to be independent draws from an ROI specific distribution. In turn, for the epilepsy dataset the statistics are analysis of variance (anova [10]) with post-hoc Tukey-Kramer multiple comparisons of the subject (ROI) means compared across the three groups, i.e. the distributions of the groups are compared as opposed to the distributions within the subject.

It should be noted especially that the referencing we are doing has a different impact on our statistical tests between the subject wise and groupwise comparisons. In the subject wise comparisons we subtract a scalar from all the values, essentially only moving the means closer or further apart, but not impacting the overall distribution. In the groupwise comparisons the referencing could in fact change the distributions (as will be shown on the epilepsy dataset later). Therefore, we will not present any distribution-related statistics (i.e. box or violin plots) for the challenge dataset, which we will use to present some of the results on the epilepsy dataset.

3.1 Reference Regions

We consider five reference regions from literature and one novel referencing approach. The first three:

1. Cerebrospinal fluid (CSF),
2. Corpus callosum (CC), and
3. Whole (masked) brain

are common reference regions based on brain anatomy [5, 11]. These are obtained from segmenting the brain, which is described in detail for the specific datasets below. The final two reference regions are based on global maps of the R_2^* relaxation rate and the relative variance of the susceptibility maps across subjects.

4. *Thresholding the relative variance map* [4] has been used as a semiautomated method that does not directly select based on anatomy.
5. *Thresholding the subject specific R_2^* map* is our novel approach to obtain a semiautomatic reference region without basing it on anatomical segmentation.

In Figure 1 an example R_2^* based reference region (thresholded at 5 Hz) can be seen overlayed a susceptibility map. It largely consists of voxels containing cerebrospinal fluid, but it encompasses a larger region of the brain, often also selecting small pockets of CSF that would not usually be part of the reference (constrained to the ventricles) [12]. Additionally, this map does not contain the CSF as a contiguous region but rather selects "specks" of CSF with low R_2^* values.

The relative variance map is computed according to

$$\text{Relative Variance} = \frac{\text{Var(Study-specific susceptibility template)}}{\text{Mean(Study-specific susceptibility template)}}$$

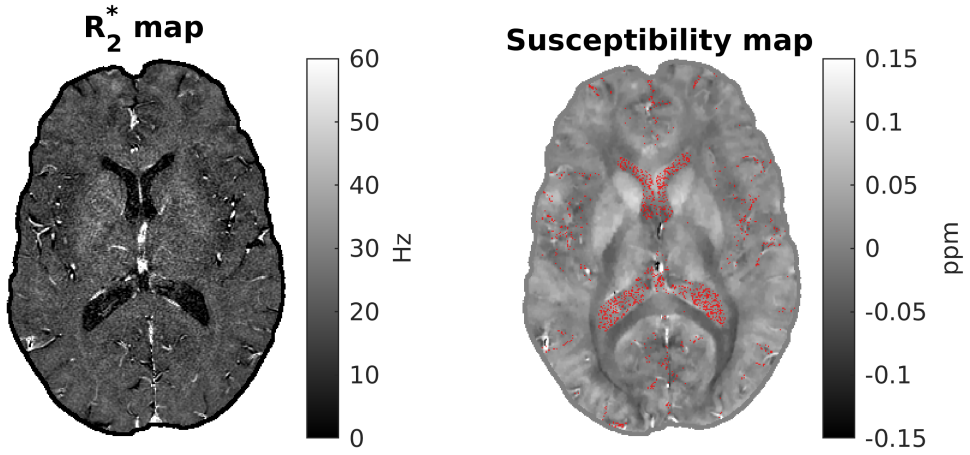

Figure 1: Example R_2^* map (left) with corresponding susceptibility reconstruction (right). Overlayed on the susceptibility map in red is the thresholded $R_2^* < 5Hz$ reference region.

where the study-specific susceptibility template is generated by registering the subject-wise susceptibilities to a study specific T1-weighted atlas to provide a single study-specific QSM template, according to Acosta-Cabronero et al. [4]. Voxels with a low relative variance ($< 3^{\text{rd}}$ percentile) are found in for example the posterior white matter as can be seen in Figure 2.

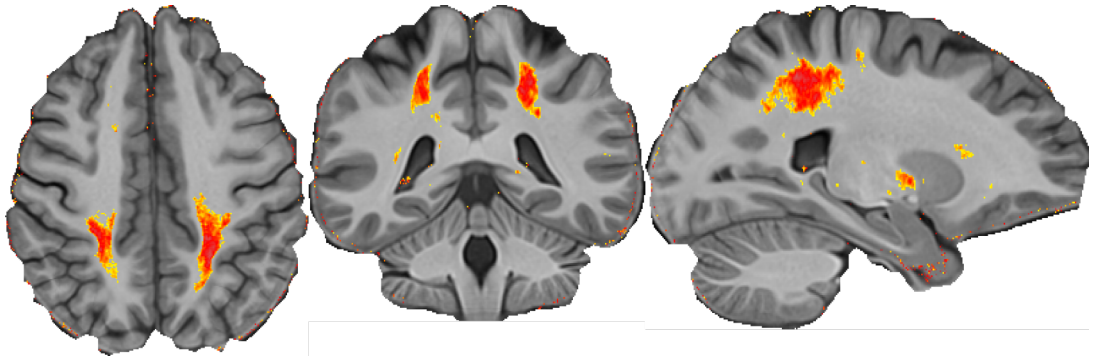

Figure 2: Study Specific T1 atlas with overlayed in yellow to red heatmap the relative regions with variance $< 3^{\text{rd}}$ percentile. Yellow being lower variance regions and red indicating the threshold of the third percentile.

3.2 QSM Reconstruction challenge 2.0 Dataset

This is a synthetic dataset, which was used to compare and evaluate QSM reconstruction methods. More than 100 different reconstructions were submitted and compared in the original publication [1] and is used frequently as a ground truth dataset for developing new reconstruction methods. In this dataset we used pairwise t-tests to compare the reconstructed

susceptibility values to the ground truth susceptibility values with and without referencing. An essential difference between the analysis on this dataset and the epilepsy dataset is that the comparisons here are between "subjects" being individual (essentially non-independent) reconstructions and the ground truth, whereas the comparisons in the epilepsy dataset are between *groups* of subjects. Therefore, any inaccuracy in the reconstruction of the reference region directly leads to an error in the value of the compared regions of interest (ROIs). We try to find whether this error is significant, and in this dataset, we can also look at which reconstruction methods are susceptible to this.

3.3 Epilepsy Dataset

To investigate pathology, we use a dataset of left and right temporal lobe epilepsy patients (denoted by LTLE and RTLE, respectively). The original analysis of this dataset can be found in Kiersnowski et al. [9]. The dataset consists of 27 healthy controls (HC), 19 patients with LTLE, and 17with RTLE, with ages ranging from 16 to 67 years old.

The original results are not (explicitly) referenced, and we compare them here with results explicitly referenced to the whole brain, corpus callosum, and cerebrospinal fluid. The referencing is in each case performed on the "raw" susceptibility maps before age correction is computed and applied. After this, the same analysis as on the original data is done, that is, the results are age corrected. Then, an analysis of variance (anova) is performed to check for groupwise differences, after which a Tukey-Kramer multiple comparison is used to compute the statistical significance of the between group differences.

GIF [13, 14, 15] was used to segment the T1 weighted images, HippoSeg [16] was used for the hippocampus specifically, and the cerebrospinal fluid was segmented using SPM12 [17] (for use in referencing). All segmented ROIs were eroded by applying a spherical kernel of radius 1 (voxel) three times to the binary ROI mask after which outliers (values outside the 1st and 99th percentile) were removed before computing the statistics (mean and standard deviation) as was done in the pipeline of the original analysis [9].

4 Results

The t-test results for the challenge dataset can be found in Figure 3. The color in these plots is used to reflect the type of reconstruction algorithm, blue for iterative optimization methods, orange for deep learning methods and green for direct dipole inversion methods. A summary of number of "correctly" reconstructed ROIs based on reference region is presented in Table 1.

Table 1: Table of number of ROIs that are reconstructed similar to the ground truth depending on the reference region chose (summed for L/R regions to summarise the information). It should be noted that the CSF and WM regions are correct in all reconstructions as expected when these regions are explicitly referenced hence they are omitted from the totals (–).

ROI	No Reference	R2s Reference	Relative Variance Reference	Explicit Whole Brain Reference	CSF Reference	White Matter Reference
Caudate	4	3	0	4	1	3
GP inter	14	14	11	14	14	14
GP ex	7	11	4	10	12	11
Putamen	5	5	4	2	5	4
Thalamus	1	2	0	2	3	1
Pulvinar	1	2	6	2	2	0
SN	19	20	20	19	20	15
STN	3	8	11	6	7	3
RN	12	14	9	10	10	14
DN	13	11	7	8	11	10
WM	0	0	0	0	0	–
GM	0	0	0	0	1	0
CSF	1	1	0	1	–	0
Blood	0	0	0	0	0	0
Fat	0	0	0	0	0	0
Bone	0	0	0	0	0	0
Air	9	0	0	0	0	0
Muscle	0	0	0	0	0	0
Calcification	3	3	3	3	3	3
Totals	92	94	75	81	89	78

In Figures 4 and 5 the groupwise (anova) p-values for the different reference regions, and the post-hoc t-test p-values for those reference regions that were statistically significant (or close to it) in the original study are given. Figure 6 gives a more in depth look at the changes in distributions for the significant ROIs in the form of box-plots and the left putamen in particular is highlighted in Figure 7 using a violin-plot comparing healthy control to the left temporal lobe epilepsy patients. The corpus callosum (CC) and internal capsule (IC) are chosen as white matter reference regions [5, 11]. The CC notably has a significant between group differences (anova p-value of 0.04). The cerebrospinal fluid (CSF) and internal capsule (IC) on the other hand are uncorrelated between the different cohorts (anova p-values of 0.98 and 0.12).

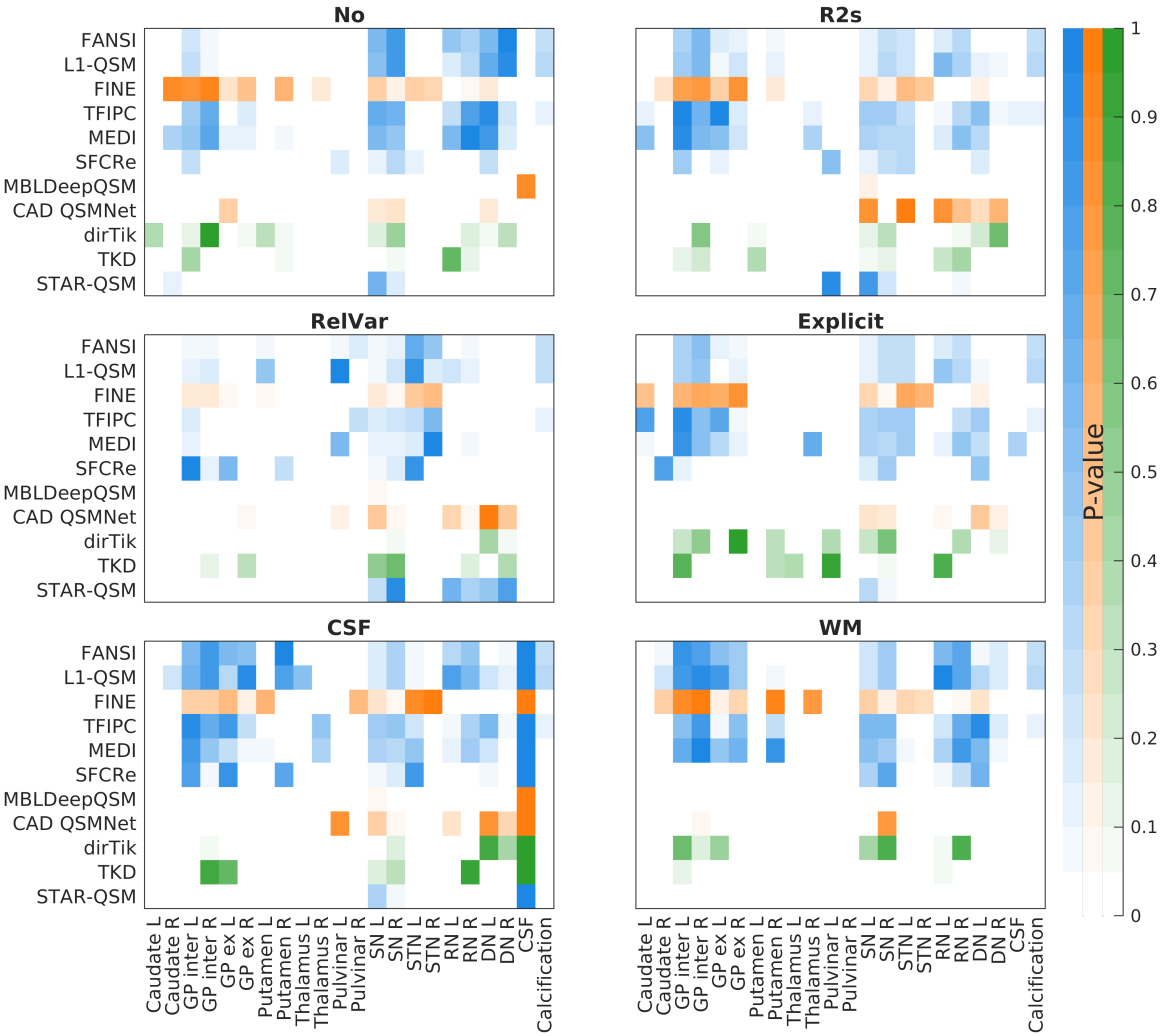

Figure 3: Challenge dataset t-test p-values based on various reference regions. Any p-value above 0.05 is colored (relative to the reconstruction type) and is not dissimilar to the ground truth (that is, all colored squares denote "correct" reconstructions).

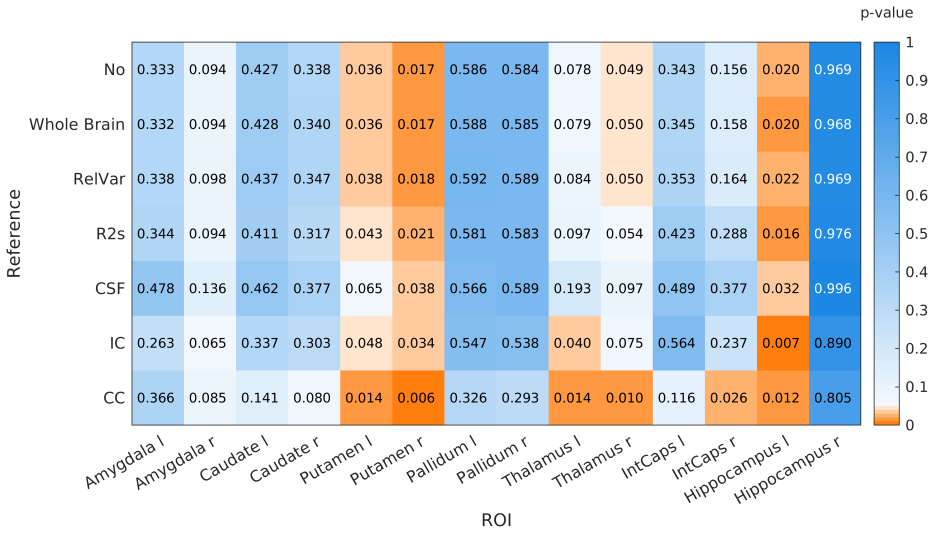

Figure 4: Groupwise anova p-values for the different references. These indicate whether there is a significant between group difference, but does not indicate which groups are different, values below 0.05 are indicated in blue and are expected to be statistically significant.

5 Discussion

From the challenge dataset referencing can "improve" the accuracy of some of the reconstructions. The overall totals come out highest for the R_2^* reference region at 94 correct regions. Interestingly the "air" ROI, which is not usually one of interest shows how referencing can bias the baseline, as this ROI is correct when no referencing is performed but is incorrect in all other cases. It could be argued that this ROI is unimportant, thereby reducing the total number of correct ROIs of importance without referencing to 83. We do not think this is the right way to think about this however, as referencing is a global correction, therefore bias in any region is an important consideration, and one cannot argue that bias that "corrects" a region over another is an improvement overall.

Regarding the epilepsy dataset, it is important to note that the Anova (groupwise) results influence which ROIs need to be analysed post-hoc (multiple-comparisons). From these results (Figure 5) we can see that whole brain and relative variance referencing don't change any significances, whereas R_2^* referencing makes the right thalamus barely insignificantly different. Larger differences are seen in the local anatomy-based reference regions, where the CSF reduces significances overall, the corpus callosum increases them (due to its correlation between groups, as noted in the results), and the internal capsule changes right for left thalamus.

In the multiple comparisons results the differences are less pronounced, and all global methods show the same (significance) results. Here, the biggest difference can be seen when referencing the corpus callosum, between right and left temporal lobe epilepsy, clearly indicating that this reference region has a significant difference between these two groups, thereby biasing the results to exacerbate the differences between them. Again it can

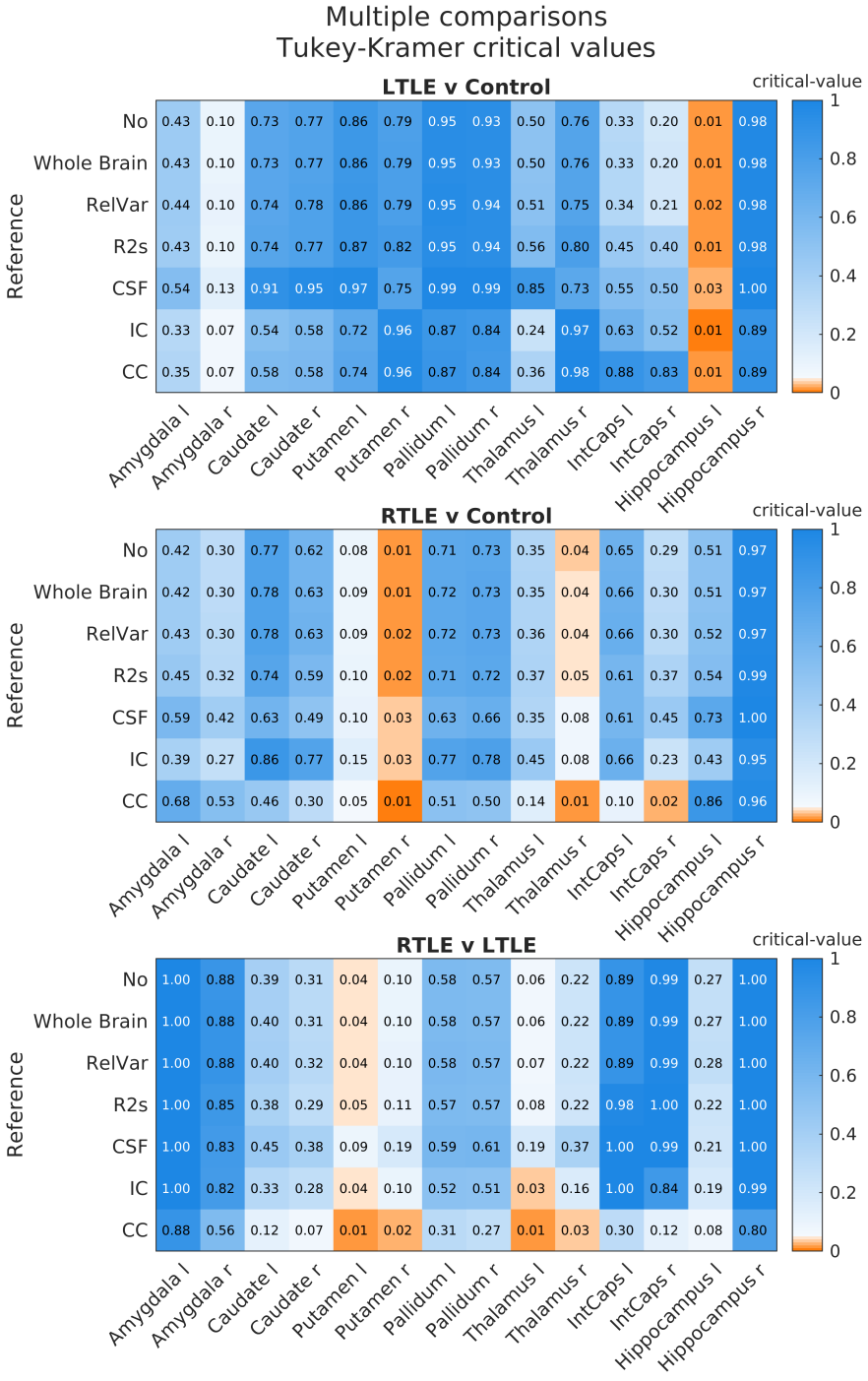

Figure 5: Tukey-Kramer multiple comparisons analysis critical values for the different references. These indicate between which of the groups there is a significant difference, values below 0.05 are indicated in orange and are statistically significant.

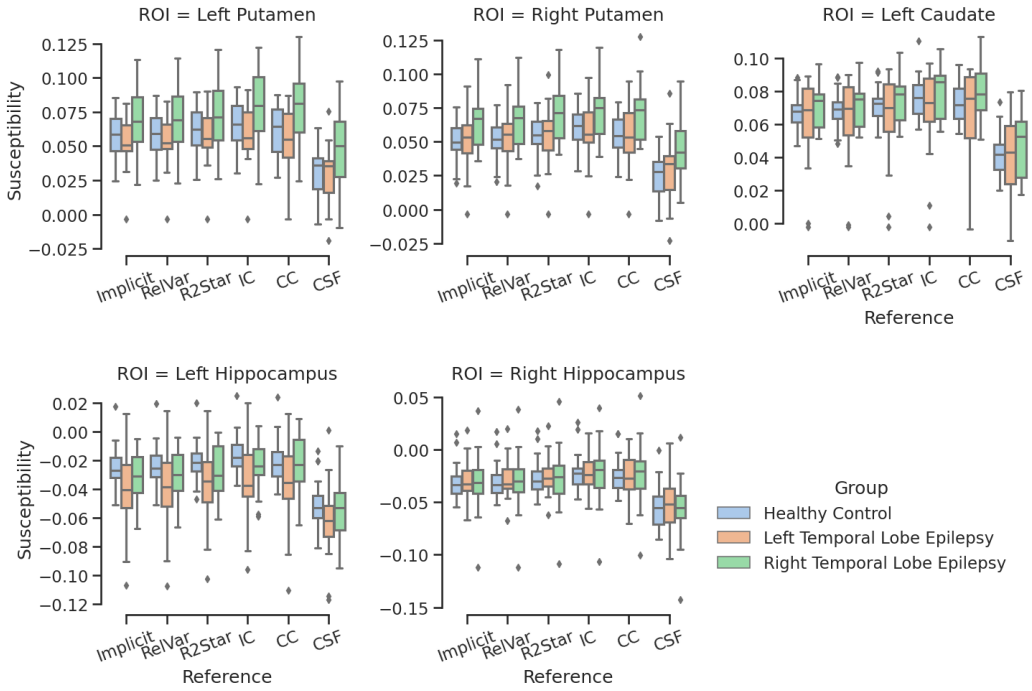

Figure 6: Boxplot of the statistically significant (groupwise) ROIs from the original study. The boxes indicate changes in variance as well as means of the groups depending on the reference method used.

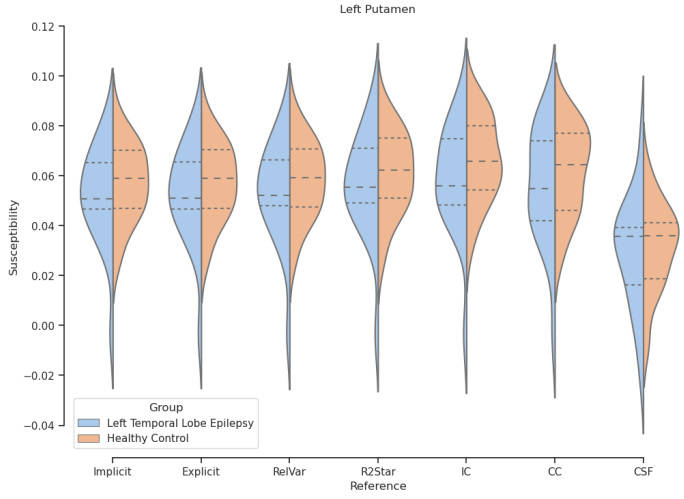

Figure 7: Violin plots of the left putamen in the epilepsy dataset. Comparing left temporal lobe epilepsy to healthy controls. The lines inside the violin plot denotes the quartiles (1st, 2nd (mean) and 3rd quartiles).

be observed that CSF referencing increases the critical values, decreasing the significant differences that are observed. The internal capsule, while not significantly different between groups does highlight a significant difference in the left thalamus when used as a reference. This might indicate that a combination of internal capsule and left thalamus may share disease characteristics between the groups, but it is difficult to investigate such a relationship without biasing our results through unintentional p-hacking.

To illustrate the effect of referencing on a specific ROI we show the Left putamen in Figure 7, where the largest difference can be observed for CSF referencing. Here the means (long dashed lines) of HC and LTLE cohorts are almost the same, when in the other datasets they are noticeably different. Besides this, both IC and CC referencing changes the shape of the distribution (when looking at the LTLE cohort almost in opposite ways), whereas there is less difference in the distributions between the global reference regions.

That brings us to an important consideration, does choosing a reference region based on a-posterior results (i.e., referencing with several regions and then choosing the reference region which shows the correlations / significant results we expect to see) equate to p-hacking? We believe this to be the case, especially when considering anatomical regions. Therefore, it is important to choose a reference region before performing a study, and only to change this choice when it turns out the reference region is significantly correlated between groups (as can be seen in the CC here), which should be checked before performing the referencing and subsequent age correction. Additionally, it can be worthwhile to check the age-related correlation of the reference region across subjects as it is important not to bias the results towards age related effects. That is, there should be no age-related changes in the reference region of choice.

6 Conclusions

This paper furthers our understanding of referencing, both from a theoretical-statistical point of view, making clear what the expected consequence is with respect to the distribution of the reference region, as well as from a practical point of view, giving examples of how referencing can influence statistical results in a clinical investigation. We show that referencing to the whole brain or low variance regions (as well as to a lesser degree low R_2^* regions) does not influence the statistical analysis significantly, in the sense that the results are the same regardless of the reference region chosen (specifically the post-hoc multiple-comparisons tests). Therefore, we would suggest referencing with one of these regions, as this can reduce the overall bias of the susceptibility estimation by reducing the inter-scan variability (while preserving the inter-subject variability as much as possible). More specific regions of interest (such as the CSF or other specific anatomical regions) have the potential to introduce additional inter-subject variability and are more likely correlated with disease, age, or other subject specific traits, which would be "averaged out" when choosing larger or less specific regions for referencing.

7 Acknowledgements

The authors would like to thank Dr. Jon Clayden for his valuable insights and corrections of our statistical analysis, and Jinjia Chen for her suggestion to investigate the relaxation rate

of tissue for low susceptibility regions. PF and KS are supported by the European Research Council Consolidator Grant DiSCo MRI SFN 770939. OCK is supported by EPSRC Doctoral Training Partnership (EP/R513143/1) and EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) (EP/S021930/1).

References

[1] Q. C. . O. Committee, B. Bilgic, C. Langkammer, J. P. Marques, J. Meineke, C. Milovic, and F. Schweser, "Qsm reconstruction challenge 2.0: Design and report of results," *Magnetic Resonance in Medicine*, vol. 86, no. 3, pp. 1241–1255, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.28754

[2] C. Milovic, C. Tejos, and P. Irarrazaval, "Structural similarity index metric setup for qsm applications (xsim)," in *5th International Workshop on MRI Phase Contrast & Quantitative Susceptibility Mapping, Seoul, Korea*, 2019.

[3] P. Fuchs, C. Milovic, and K. Shmueli, "Region-of-Interest Based Statistical Analysis of the 2019 QSM Challenge," in *Proceedings 31. Annual Meeting International Society for Magnetic Resonance in Medicine*, vol. 31. Toronto, Canada: http://archive.ismrm.org/2023/1044.html, 2023, p. 1044.

[4] J. Acosta-Cabronero, M. J. Betts, A. Cardenas-Blanco, S. Yang, and P. J. Nestor, "In vivo mri mapping of brain iron deposition across the adult lifespan," *Journal of Neuroscience*, vol. 36, no. 2, pp. 364–374, 2016. [Online]. Available: https://www.jneurosci.org/content/36/2/364

[5] S. Straub, T. M. Schneider, J. Emmerich, M. T. Freitag, C. H. Ziener, H.-P. Schlemmer, M. E. Ladd, and F. B. Laun, "Suitable reference tissues for quantitative susceptibility mapping of the brain," *Magnetic Resonance in Medicine*, vol. 78, no. 1, pp. 204–214, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26369

[6] P. Ravanfar, S. M. Loi, W. T. Syeda, T. E. Van Rheenen, A. I. Bush, P. Desmond, V. L. Cropley, D. J. R. Lane, C. M. Opazo, B. A. Moffat, D. Velakoulis, and C. Pantelis, "Systematic review: Quantitative susceptibility mapping (qsm) of brain iron profile in neurodegenerative diseases," *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.618435

[7] Z. Liu, P. Spincemaille, Y. Yao, Y. Zhang, and Y. Wang, "Medi+ 0: Morphology enabled dipole inversion with automatic uniform cerebrospinal fluid zero reference for quantitative susceptibility mapping," *Magnetic resonance in medicine*, vol. 79, no. 5, pp. 2795–2803, 2018.

[8] F. Schweser, A. L. R. D. Martins, J. Hagemeier, F. Lin, J. Hanspach, B. Weinstock-Guttman, S. Hametner, N. Bergsland, M. G. Dwyer, and R. Zivadinov, "Mapping of thalamic magnetic susceptibility in multiple sclerosis indicates decreasing iron with disease duration: a proposed mechanistic relationship between inflammation and oligodendrocyte vitality," *Neuroimage*, vol. 167, pp. 438–452, 2018.

[9] O. C. Kiersnowski, G. P. Winston, L. Caciagli, E. Biondetti, M. Elbadri, S. Buck, J. S. Duncan, J. S. Thornton, K. Shmueli, and S. B. Vos, "Quantitative susceptibility mapping identifies hippocampal and other subcortical grey matter tissue composition changes in temporal lobe epilepsy," *Human Brain Mapping*, 2023.

[10] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949. [Online]. Available: http://www.jstor.org/stable/3001913

[11] P. Ravanfar, S. M. Loi, W. T. Syeda, T. E. Van Rheenen, A. I. Bush, P. Desmond, V. L. Cropley, D. J. R. Lane, C. M. Opazo, B. A. Moffat, D. Velakoulis, and C. Pantelis, "Systematic review: Quantitative susceptibility mapping (qsm) of brain iron profile in neurodegenerative diseases," *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.618435

[12] I. A. L. Lim, A. V. Faria, X. Li, J. T. Hsu, R. D. Airan, S. Mori, and P. C. van Zijl, "Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: Application to determine iron content in deep gray matter structures," *NeuroImage*, vol. 82, pp. 449–469, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811913006411

[13] M. J. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin, "Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion," *IEEE transactions on medical imaging*, vol. 34, no. 9, pp. 1976–1988, 2015.

[14] F. Prados, M. J. Cardoso, N. Burgos, C. Wheeler-Kingshott, S. Ourselin, C. Angela, M. Gandini, and S. Ourselin, "Niftyweb: web based platform for image processing on the cloud," in *24th scientific meeting and exhibition of the international society for magnetic resonance in medicine (ISMRM)*, 2016, pp. 7–13.

[15] CMIC, "Niftyweb," http://niftyweb.cs.ucl.ac.uk/program.php?p=GIF, 8 2023.

[16] G. P. Winston, M. J. Cardoso, E. J. Williams, J. L. Burdett, P. A. Bartlett, M. Espak, C. Behr, J. S. Duncan, and S. Ourselin, "Automated hippocampal segmentation in patients with epilepsy: available free online," *Epilepsia*, vol. 54, no. 12, pp. 2166–2173, 2013.

[17] S. J. Kiebel, J. Ashburner, J.-B. Poline, and K. J. Friston, "Mri and pet coregistration—a cross validation of statistical parametric mapping and automated image registration," *Neuroimage*, vol. 5, no. 4, pp. 271–279, 1997.