Patient-Specific Detection of Perfusion Abnormalities Combining Within- and Between-Subject Variances in ASL

Camille Maumet, Pierre Maurel, Béatrice Carsin, Jean-Christophe Ferré, Christian Barillot

HAL Id: hal-00781247

https://inria.hal.science/hal-00781247

Submitted on 25 Jan 2013

Patient-Specific Detection of Perfusion Abnormalities Combining Within- and Between-Subject Variances in ASL

Camille Maumet, Pierre Maurel, Béatrice Carsin, Jean-Christophe Ferré, Christian Barillot

Patient-Specific Detection of Perfusion Abnormalities Combining Within- and Between-Subject Variances in ASL

Camille Maumet[*†‡§], Pierre Maurel[*†‡§], Béatrice Carsin[¶],

Jean-Christophe Ferré[*†‡§¶], Christian Barillot[*†‡§]

Project-Team Visages

* University of Rennes 1, Faculty of medecine, F-35043 Rennes, France
† INSERM, U746, F-35042 Rennes, France
‡ CNRS, IRISA, UMR 6074, F-35042 Rennes, France
§ Inria, VISAGES project-team, F-35042 Rennes, France
¶ CHU Rennes, Department of Neuroradiology, F-35033 Rennes, France

Abstract: Patient-specific patterns of abnormal perfusion are useful indicators in the diagnosis and monitoring of patients suffering from a microvascular dysfunction. Arterial Spin Labeling (ASL), as an entirely non-invasive magnetic resonance imaging technique, is a promising tool to image perfusion and the General Linear Model (GLM) could be used in order to quantitatively detect abnormal patterns of perfusion at the individual level in ASL.

In this paper, patient-specific perfusion abnormalities were therefore identified by comparing a single patient to a group of healthy controls using a mixed-effect hierarchical GLM. Two approaches are currently in use to solve hierarchical GLMs: (1) the homoscedastic approach assumes homogeneous variances across subjects and (2) the heteroscedastic approach is theoretically more efficient in the presence of heterogeneous variances but algorithmically more demanding. In practice, in functional magnetic resonance imaging studies, the superiority of the heteroscedastic approach is still under debate. Due to the low signal-to-noise ratio of ASL sequences, within-subject variances have a significant impact on the estimated perfusion maps and the heteroscedastic model might be better suited in this context.

In this paper we studied how the homoscedastic and heteroscedastic approaches behave in terms of specificity and sensitivity in the detection of patient-specific ASL perfusion abnormalities. Validation was undertaken on a dataset of 25 patients diagnosed with brain tumors and 36 healthy volunteers. We showed evidence of heterogeneous within-subject variances in ASL and pointed out an increased false positive rate of the homoscedastic model. In the detection of patient-specific brain perfusion abnormalities with ASL, modeling heterogeneous variances increases the sensitivity at the same specificity level.

Key-words: Arterial Spin Labeling, Hypo-perfusion, Hyper-perfusion, General Linear Model, Within-subject variance, Heteroscedasticity

Détection d'anomalies de perfusion en prenant en compte les variances intra- et inter- sujets en imagerie par marquage de spins artériels

Résumé :   Dans ce travail, les anomalies de perfusion chez un patient sont identifiées, en imagerie par marquage de spins artériels (Arterial Spin Labeling), grâce à une comparaison entre la carte de perfusion du patient et un modèle de perfusion normale obtenu à partir d'un groupe de contrôles sains. Pour effectuer cette comparaison, deux approches sont possibles : (1) l'approche homoscédastique qui suppose une homogénéité des variances chez les différents individus et (2) l'approche hétéroscédastique qui est théoriquement plus efficace en présence de variances hétérogènes mais plus coûteuses algorithmiquement. En pratique, en Imagerie par Résonance Magnétique fonctionnelle, la supériorité de l'approche hétéroscédastique est encore sujet à débat. En ce qui concerne les séquences par marquage de spins artériels (ASL), les variances intra-sujet ont un impact significatif sur l'estimation de la perfusion et on peut se demander si le modèle hétéroscédastique n'est pas plus adapté dans ce contexte.

Nous étudions ici comment les approches homoscédastique et hétéroscédastique se comportent en termes de sensibilité et spécificité pour la détection d'anomalies de perfusion en ASL. La validation est effectuée sur un ensemble de 25 patients présentant des tumeurs cérébrales et 36 volontaires sains. Nous montrons l'hétérogénéité des variances intra-sujets en ASL, ainsi que l'augmentation du taux de faux positifs dû au modèle homoscédastique. En ce qui concerne la détection des anomalies de perfusion cérébrale en ASL, modéliser l'hétérogénéité des variances augmente la sensibilité pour le même niveau de spécificité.

Mots-clés : Marquage de spins artériels, Hypo-perfusion, Hyper-perfusion, variance intra-sujet, Hétéroscédasticité

1   Introduction

Brain perfusion is the biological process that ensures the delivery of oxygen and nutrients to the cerebral tissues by means of micro-circulation. As an indicator of the well-being of the tissues, perfusion is a useful measurement for diagnosis in clinical practice. Identifying regions of abnormal perfusion, either hypo-perfusions or hyper-perfusions, can help understanding the mechanism of a disease and taking care of the patient. For instance, for patients diagnosed with tumors, the clinician is interested in hyper-perfusions that would reveal the grade of the tumor [44], or help differentiating between post-radiation necrosis and tumor recurrence [38]. In strokes, the extent of hypo-perfusion and mismatch with diffusion imaging provide an insight on the possible recovery of the tissue [19].

Arterial Spin Labeling (ASL), a Magnetic Resonance Imaging (MRI) technique introduced in the early 1990's, allows to measure the level of perfusion through a quantitative index: the cerebral blood flow (CBF). Contrary to standard perfusion imaging, including Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) in nuclear medicine, or Dynamic Susceptibility weighted Contrast (DSC) in MRI, ASL is completely non-invasive and does not require the injection of an exogenous contrast agent. In ASL, blood water, used as an endogenous tracer, is labeled with a radio-frequency pulse. After a delay called *inversion time*, a labeled image of the brain is acquired. A control image is also acquired without prior labeling. The difference between the labeled and the control image leads to a perfusion-weighted map. To increase the Signal-to-Noise Ratio (SNR) of the measure, the acquisition is repeated several times. The absence of radiation and the use of an endogenous tracer, which removes the need of injecting a contrast agent, are clear advantages of ASL over standard perfusion imaging techniques. The ASL sequence however suffers from a low SNR, which is still a serious obstacle for its use in clinical practice.

Since the introduction of ASL, and despite its low SNR, a large number of studies have demonstrated its usefulness in identifying patterns of abnormal perfusion at the group level (e.g. [34]). To date, identification of individual patterns of hypo- and hyper-perfusions in patients with ASL mainly relies on visual analysis [50, 12]. Very few ASL studies focused on voxelwise quantitative perfusion abnormality detections at the individual level. In [33], a template-based analysis was presented in order to detect individual activation patterns in functional ASL data. Interestingly, they also applied their method to a pathological case in order to detect a hypo-perfusion co-located with a dysplasia in an epileptic patient. While their approach opens the field to the detection of patient-specific perfusion abnormalities with ASL, a single pathological case was presented and no quantitative validation performed. Their detections relied on z-scores, but in a more general setting, the most widespread approach to compare voxel-wise maps in neuroimaging is the massively univariate General Linear Model (GLM). To detect differential patterns between two groups with repeated measurements a subtype of GLM is employed: a mixed-effect hierarchical two-sample t-test with two levels: subject and group. In this context, two variance components are of interest: the within-subject variance (or the measurement error, estimated from the repeated ASL acquisitions of a single subject) and the between-subject variance.

In the functional MRI community, where similar statistical models are applied, two approaches are currently in use to solve hierarchical GLMs. On the one hand, the homoscedastic approach, also termed "summary statistics" [31], or referred as "ordinary least square estimation" [26, 27] or "conventional group analysis" [11], assumes homogeneous within-subject variances or negligible within-subject variances by comparison to between-subject variance. On the other hand, the heteroscedastic approach, also referred as "full mixed-effect" [16, 39, 35], "mixed-effect model" [11] or "weighted least square estimation" [26, 27] models heterogeneous within-subject variances. There is indeed a large panel of homonyms to refer to these two approaches, in cur-

rent practice they are also sometimes referred as "mixed-effect" and "random-effect" approaches [20]. However, in a statistical sense both the homoscedastic and the heteroscedastic approaches are mixed-effect models. That is why, in the context of this paper, we chose to term the two approaches "homoscedastic model" and "heteroscedastic model" to identify them according to their main difference: the homoscedasticity (constant within-subject variance across subjects) assumption.

The homoscedastic model is theoretically less efficient in the presence of heterogeneous within-subject variances. However, in practice, the true variance components are unknown and the superiority of the heteroscedastic model is therefore questioned. Whether modeling heterogeneous variances should be preferred over the homoscedastic model is still under discussion in the fMRI community. Several authors outlined the benefits of heteroscedastic models including [48, 6, 47, 26, 25, 39, 11]. However, [27] showed that the homoscedastic approach is still valid with near optimal sensitivity in the context of one-sample t-tests. Nonetheless, the same authors acknowledged that, in two-sample t-tests, appropriate modeling of heterogeneous within-subject variances might be crucial [27, 35]. Recently the conclusions regarding one-sample t-tests were revisited leading to opposite conclusions [11]. In the different software packages currently available to deal with fMRI data both approaches are represented: SPM[1] favors the homoscedastic approach [18] while FSL[2] [47], AFNI[3] [13] and fmristat[4] [48] use the heteroscedastic model.

In ASL, the importance of the within-subject variance has been outlined by [40]. While we previously investigated the ability to detect hypo- and hyper-perfused regions using a template of normal perfusion in two conference papers [22, 23], here, we focus on the comparison of the homoscedastic and heteroscedastic approaches. We further improve on these papers by providing quantitative measure of heterogeneity. Also, a larger database is studied and a novel quantitative validation is performed.

In this paper, we focus on quantitative detections of pathological brain perfusion abnormalities at the individual level. To this aim, we employ and compare two GLM-based models: the homoscedastic and heteroscedastic approaches. We test whether the assumptions underlying the homoscedastic approach are verified in pulsed ASL datasets. We furthermore study how the homoscedastic and heteroscedastic approaches behave in terms of specificity and sensitivity in the detection of patient-specific perfusion abnormalities.

A quantitative validation is performed on a dataset of 25 patients diagnosed with brain tumors. Though there is no well-defined ground truth, this pathology was selected since patterns of hypo- and hyper-perfusions have been widely studied in this clinical context. The model of normal perfusion is computed out of the data of 36 healthy volunteers.

Section 2 begins with a presentation of the homoscedastic and heteroscedastic models employed in the detection of patient-specific brain perfusion abnormalities with ASL. Then, the experiments designed to test the assumptions of the homoscedastic model and measure sensitivity and specificity are presented. Section 3 presents the datasets under study: the acquisition and pre-processing are described. The results are presented in section 4. Section 5 gives a discussion and concludes.

2 Methods

In 2.1, the homoscedastic and heteroscedastic models and their implementation in the context of patient-specific detection of perfusion abnormalities in ASL are presented. Then, in 2.2, the

---

[1] http://www.fil.ion.ucl.ac.uk/spm
[2] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki
[3] http://afni.nimh.nih.gov/afni
[4] http://www.math.mcgill.ca/keith/fmristat

experiments undertaken in order to test the assumptions of the homoscedastic model and to compare the homoscedastic and heteroscedastic models are described.

2.1   Detection of patient-specific perfusion abnormalities using a mixed-effect hierarchical two-sample t-test

This section presents the common massively univariate GLM usually employed for voxel-based analysis in the neuroimaging community. The main assumption behind the GLM is the linearity of the effects. Moreover, Gaussian noise is usually assumed. In the standard approach, the GLM is defined for each voxel, which is why this approach is often termed "massively univariate". For ease of notation the voxel index is omitted in the remainder of the paper.

In the context of this paper, we aim at outlining areas of abnormal perfusion in the perfusion map of a subject of interest by comparison to a group of healthy controls. We hence focus on a subtype of GLM: a *one-versus-many* mixed-effect two-sample t-test. As several measurements are available for each subject, a hierarchical model is defined with 2 levels: subject and group.

Since the subject under study is usually an individual suffering of a pathological condition, the term *patient* will be used in the following. This approach is nevertheless suited for the comparison of any single subject (control, patient) to a group of individuals.

A description of the subject level is proposed in 2.1.1. In 2.1.2 the group level is presented along with a description of the two main related approaches: the homoscedastic and heteroscedastic models. Lastly, hypothesis testing performed in order to outline the perfusion abnormalities is discussed in 2.1.3.

2.1.1   Subject level (First level)

Given a voxel, for each subject s we have:

$$Y_s = X_s \, \beta_s + \epsilon_s, \tag{1}$$

where Y_s is a vector containing the observations at the given voxel, β_s is the set of subject parameters to be estimated, X_s is the subject-level design matrix and ϵ_s contains the residual errors. In the following, we assume that the data can be described with a single parameter per subject, β_s, however more parameters (e.g. nuisance covariates as in [42]) could be considered. In fMRI, X_s is closely related to the time-course of the paradigm under study. In ASL, at the subject level, the data under study is a 4D volume of CBF maps containing V volumes (1 per repeated acquisition). The observations Y_s are therefore repeated measurements of a same underlying value and the X_s matrix is a vector of ones. Then, the model reduces to:

$$Y_s = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_s + \epsilon_s.$$

Assuming Gaussian noise of the errors, ϵ_s follows a normal distribution: $\epsilon_s \sim \mathcal{N}(0, \sigma_s^2)$. While, in fMRI, temporal autocorrelation must be accounted for [2], in ASL, thanks to the subtraction process between control and labeled scans, we can reasonably assume white noise [1, 26]. The subject parameter β_s is thus estimated by Ordinary Least Squares (OLS):

$$\hat{\beta}_s = \frac{1}{V} \sum_{i=1}^{V} y_{s,i}, \tag{2}$$

where $y_{s,i}$ is the i$^{\text{th}}$ element of vector Y_s. Similarly, the sampling variance of $\hat{\beta}_s$ is estimated by:

$$\widehat{\text{Var}}(\hat{\beta}_s) = \frac{\hat{\sigma}_s^2}{V} \quad \text{where} \quad \hat{\sigma}_s^2 = \frac{1}{V-1} \sum_{i=1}^{V} \left(y_{s,i} - \hat{\beta}_s \right)^2. \tag{3}$$

2.1.2 Group level (Second level)

In a general setting, in a group of n subjects, the subject parameters $(\beta_s)_{1 \leq s \leq n}$ can be combined using the following model:

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = X_G\,\beta_G + \gamma_G, \tag{4}$$

where X_G is the group-level design matrix, β_G the group parameters and γ_G the residual error term (outlining the intra-group variability). In a one-sample t-test, X_G would be a vector of ones and β_G contain a single estimate. In our context, the subjects are part of two groups ($n-1$ controls in the first group and 1 patient in the second) and β_G is a vector with 2 elements. Let $\beta_{controls}$ and $\beta_{patient}$ be the control group and patient parameters. Without loss of generality, we assume that subjects 1 to $n-1$ are part of the control group and subject n is the patient of interest. Then the second-level model reduces to:

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{n-1} \\ \beta_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{controls} \\ \beta_{patient} \end{bmatrix} + \gamma_G. \tag{5}$$

Assuming Gaussian errors, γ_G follows a normal distribution $\gamma_G \sim \mathcal{N}(0, \sigma_G^2)$, where σ_G^2 is the common between-subject (also termed "within-group" or even "group") variance of both groups. Indeed, since in a *one-versus-many* study, a single subject is available in the patient group, estimating a different variance for each group (control, patient) is impractical.

However, the true subject parameters are in fact unknown and, in practice, their estimates from the first level are used in the second-level leading to a slightly different model:

$$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{n-1} \\ \hat{\beta}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{controls} \\ \beta_{patient} \end{bmatrix} + \gamma_{G_C}. \tag{6}$$

The new error term γ_{G_C} is impacted by two combined sources of variations: the error measurement on the subject parameters (also termed within-subject variance) and the between-subject variance. Each element of γ_{G_C} therefore follows a normal distribution: $\gamma_{G_C}^s \sim \mathcal{N}\left(0, \sigma_G^2 + \frac{\sigma_s^2}{V}\right)$.

Linear combinations of the group parameters can be calculated using a particular contrast. Here we are interested in the patient versus control group contrast:

$$b = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_{controls} \\ \beta_{patient} \end{bmatrix} = \beta_{controls} - \beta_{patient}. \tag{7}$$

Two approaches have been proposed in the neuroimaging literature to solve the system (6) and find an estimate of the patient versus control group contrast, $\hat{b}$, and its associated sampling variance, $\widehat{\text{Var}}(\hat{b})$. The homoscedastic approach assumes homoscedasticity while the heteroscedastic

approach explicitly models heterogeneous within-subject variances. Both models are described hereafter.

Homoscedastic model   The homoscedastic model is based on the assumption that the within-subject variance is either negligible by comparison to the between-subject variance (i.e. $\sigma_s^2 \ll \sigma_G^2$, $\forall\, 1 \leq s \leq n$) or, roughly constant across subjects (i.e. $\sigma_s^2 \approx \sigma_{SUB}^2$, $\forall\, 1 \leq s \leq n$). Then each element of γ_{G_C} follows a normal distribution:

$$\gamma_{G_C}^s \sim \mathcal{N}(0, \sigma_{G_C}^2) \tag{8}$$

where $\sigma_{G_C}^2$ is the combined within- and between-subject variance. Depending on the assumption, we have $\sigma_{G_C}^2 = \sigma_G^2 + \dfrac{\sigma_{SUB}^2}{V}$ or $\sigma_{G_C}^2 \approx \sigma_G^2$ and the combined within- and between-subject variance is therefore constant across subjects. The sphericity assumption (no heteroscedasticity and no autocorrelation) hence holds and the system (6) is then solved by OLS to get:

$$\hat{\beta}_{controls}^{HOMO} = \frac{1}{n-1} \sum_{s=1}^{n-1} \hat{\beta}_s$$
$$\hat{\beta}_{patient}^{HOMO} = \hat{\beta}_n, \tag{9}$$

The associated sampling variances are:

$$\widehat{\mathrm{Var}}(\hat{\beta}_{controls}^{HOMO}) = \frac{\hat{\sigma}_{G_C}^2}{n-1}$$
$$\widehat{\mathrm{Var}}(\hat{\beta}_{patient}^{HOMO}) = \hat{\sigma}_{G_C}^2 \tag{10}$$
$$\text{where} \quad \hat{\sigma}_{G_C}^2 = \frac{1}{(n-1)-1} \sum_{s=1}^{n-1} \left(\hat{\beta}_s - \hat{\beta}_{controls}^{HOMO}\right)^2.$$

As a consequence, the patient versus control group contrast b is estimated by:

$$\hat{b}^{HOMO} = \hat{\beta}_{controls}^{HOMO} - \hat{\beta}_{patient}^{HOMO} = \frac{1}{n-1} \sum_{s=1}^{n-1} \hat{\beta}_s - \hat{\beta}_n, \tag{11}$$

and the sampling variance of this estimator is:

$$\widehat{\mathrm{Var}}(\hat{b}^{HOMO}) = \hat{\sigma}_{G_C}^2 \left(\frac{1}{n-1} + 1\right). \tag{12}$$

Heteroscedastic model   In the heteroscedastic model, heterogeneous variances are accounted for. As described earlier, in the general case, each element of γ_{G_C} follows a normal distribution:

$$\gamma_{G_C}^s \sim \mathcal{N}(0, \sigma_G^2 + \frac{\sigma_s^2}{V}). \tag{13}$$

Due to the non-sphericity of the measurement errors (as a consequence of heteroscedasticity), the system (6) is solved by using a weighted least square to get:

$$\hat{\beta}_{controls}^{HETERO} = \frac{1}{\sum_{j=1}^{n-1} w_j} \sum_{s=1}^{n-1} w_s\, \hat{\beta}_s$$
$$\text{where } w_s = \frac{1}{\hat{\sigma}_G^2 + \hat{\sigma}_s^2} \tag{14}$$
$$\hat{\beta}_{patient}^{HETERO} = \hat{\beta}_n.$$

The associated sampling variances are:

$$\widehat{\mathrm{Var}}(\hat{\beta}_{controls}^{HETERO}) = \frac{1}{\sum_{s=1}^{n-1} w_s}$$
$$\widehat{\mathrm{Var}}(\hat{\beta}_{patient}^{HETERO}) = \hat{\sigma}_G^2 + \hat{\sigma}_n^2. \tag{15}$$

The within-subject variance estimates, $\hat{\sigma}_s^2$, are computed at the subject level as described in eq. (3). In this paper, we compute $\hat{\sigma}_G^2$ using a recent computationally efficient approach available in the MEMA function of the AFNI software package and described in [11].

The patient versus control group contrast b is estimated by:

$$
\begin{aligned}
\hat{b}^{HETERO} &= \hat{\beta}_{controls}^{HETERO} - \hat{\beta}_{patient}^{HETERO} \\
&= \frac{1}{\sum_{j=1}^{n-1} w_j} \sum_{s=1}^{n-1} w_s \hat{\beta}_s - \hat{\beta}_n
\end{aligned}
\tag{16}
$$

and the sampling variance of this estimator is:

$$\widehat{Var}\big(\hat{b}^{HETERO}\big) = \frac{1}{\sum_{s=1}^{n-1} w_s} + \hat{\sigma}_G^2 + \hat{\sigma}_n^2. \tag{17}$$

If homoscedasticity is respected then it can easily be proven that this model reduces to the homoscedastic model. However, in the presence of heteroscedasticity, the heteroscedastic model has two main advantages by comparison to the homoscedastic model:

1. In the control group, observations with high within-subject variances are downweighted in order to provide a more efficient estimate of the control group parameter, $\hat{\beta}_{controls}^{HETERO}$.

2. While with the homoscedastic model the sampling variance of the patient group estimate, $\widehat{\mathrm{Var}}(\hat{\beta}_{patient}^{HOMO})$, depends solely on variance estimations performed in the control group, the heteroscedastic sampling variance, $\widehat{\mathrm{Var}}(\hat{\beta}_{patient}^{HETERO})$, takes advantage of both the control group (estimation of σ_G^2) and the patient (estimation of σ_n^2) data.

The impact on the efficiency of the control group estimate (point 1) might be subtle and leads to no substantial improvement in one-sample t-tests if the homoscedasticity assumption is not overly altered [27]. However, the impact on the variance estimate of the patient parameter (point 2) can be large if the within-subject variance of the patient of interest is very different from the control subjects. In ASL studies, the large influence of within-subject variance has been described in [40]. Furthermore, patients are known to be less cooperative than control subjects which could potentially lead to within-subject variance inflation due to increased movement. We therefore expect that adjusting for the proper patient within-subject variance, using the heteroscedastic instead of the homoscedastic model, will lead to more accurate results in the detection of patient-specific perfusion abnormalities.

2.1.3   Hypothesis testing

Under the null hypothesis:

$$H_0 : \beta_{controls} = \beta_{patient}, \tag{18}$$

the estimated patient versus control group contrast, $\hat{b}$, divided by its estimated sampling standard deviation $(\widehat{\mathrm{Var}}(\hat{b}))^{\frac{1}{2}}$ follows a t-distribution with $n-1$ degrees of freedom.

A probability under the null hypothesis can therefore be calculated for each voxel with:

$$P(X < x) \text{ where } X = \frac{\hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}} \sim \mathcal{T}_{n-1}, \tag{19}$$

and x is the value taken by X at the voxel of interest. Equation (19) gives the probability to have a hyper-perfusion, similarly, hypo-perfusions can be detected by substituting $P(X < x)$ by $P(X > x)$. The thresholding of this probability map gives the detections.

2.2   Evaluation framework

In this subsection, we first present how the ground truth hypo- and hyper-perfusions were determined based on anatomical and perfusion information (2.2.1). Then, the experiments undertaken to verify the assumptions of the homoscedastic model (2.2.2) and to compare the homoscedastic and heteroscedastic models (2.2.3) are presented.

2.2.1   Ground truth

Quantitative assessment of the detections and comparison between detection methods are challenging tasks. This is mainly because, like in many other medical imaging problems, the ground truth is not clearly stated. We chose to evaluate this framework on patients diagnosed with tumor pathology because perfusion abnormalities have been widely studied in this clinical context [9, 38, 43, 44, 10]. Also, as perfusion is a useful clinical information, DSC is part of the clinical routine for these patients. Based on clinical knowledge, we used a semi-automatic procedure that took advantage of the complementary anatomical (T1w-Gd, T2w FLAIR) and perfusion (DSC) information to get an estimation of the ground truth.

True positives   According to clinical knowledge, T1w-Gd hypersignals are usually indicative of the presence of hyper-perfusions [44]. However, in some tumor types they are related to a different phenomenon. For instance, in lymphomas the hypersignal identified on the T1w-Gd map does not correspond to a hyper-perfusion [44]. Moreover, neoangiogenesis, which is characterized by hyper-perfusions in ASL, can spread out of the T1w-Gd hypersignal in particular in high grade tumors such as glioblastoma [17]. Ground truth hyperperfused regions based on the T1w-Gd hypersignal are therefore imprecise.

To overcome these limitations, we applied a two-step procedure taking advantage of both the anatomical and perfusional information (delivered by DSC CBF). In order to get an estimation of the ground truth, we implemented a method inspired by the hotspot technique [28] commonly used in clinical practice. To this aim, the tumor was first segmented using a semi-automated method based on the T2w and T1w-Gd images and visually inspected by an expert neuro-radiologist. Then, we compared the tissue segmented as part of the tumor to its controlateral counterpart in the DSC CBF map. Voxels overtaking the lower and upper deciles were identified as potential hypo- and hyper-perfusions. Each potential perfusion abnormality was then visually inspected by an expert neuro-radiologist and manually corrected if needed. Special care was taken in order to avoid inclusion of hyper-perfusions related to the presence of arteries.

Due to its low SNR, ASL is not well suited to measure low levels of perfusion [45]. That is why, we focused on hyper-perfusions for sensitivity estimation. Hypo-perfusions were nevertheless retained for specificity calculations.

Out of the 17 patients included in this study who underwent a DSC sequence, 9 presented hyper-perfusions, 16 hypo-perfusions and 8 both.

False positives   According to clinical knowledge, in the absence of metastasis, the perfusion abnormalities should be confined to the affected tissue (tumor and oedema) identifiable on T1w-Gd and T2w. The proportion of non-affected tissue detected as a perfusion abnormality was used as a measure of the false positive rate.

In the control group, in which no detections were expected, an additional estimate of the specificity was calculated by leave-one-out cross-validation.

2.2.2   Validity of the assumptions of the homoscedastic model

The homoscedastic model makes the assumption that within-subject variance is either negligible by comparison to between-subject variance, or roughly constant across subjects [27]. In order to test each of these assumptions, we performed two experiments.

Negligible within-subject variance   First, following [11], we measured the proportion of total variability that occurred within subjects with the following index, defined at each voxel for each subject s:

$$\lambda_s = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_G^2 + \hat{\sigma}_s^2} \tag{20}$$

Values close to 1 mean that the within-subject variance (σ_s^2) is preponderant compared to between-subject variance (σ_G^2) and values close to 0 mean that between-subject variance holds the majority of the total variance. This measure is provided as an output of the MEMA function in AFNI [11].

Constant within-subject variance across subjects   In a second experiment, we focused on the within-subject variance to verify whether it could be assumed roughly constant across subjects. To this aim, we calculated an average within-subject variance for each subject. This index was computed as suggested by [27] by averaging the within-subject variance for voxels within the interquartile range of the nonzero between-subject variance. Given this index of within-subject variance, we searched for outliers in the control group with Rosner's test [36], assuming that the distribution of the variance estimate was approximately normal. Then, we checked whether the control and patient groups had significantly different medians with a non-parametric Kruskal-Wallis test.

2.2.3   Comparison of the homoscedastic and heteroscedastic models

Detections at a fixed threshold   In order to assess the difference that heteroscedasticity modeling would induce, we compared the sensitivity and specificity of the homoscedastic and heteroscedastic models. We compared both approaches in a usual setting, where the detections were identified at a threshold $p < 0.05$ with False Discovery Rate (FDR) correction for multiple comparisons [7]. As the data is usually pre-smoothed with a Gaussian kernel at the end of the pre-processing, we studied 6 kernel sizes defined by their full-width at half maximum (FWHM): 0 mm^3 (i.e. no smoothing), 4 mm^3, 6 mm^3, 8 mm^3, 10 mm^3, 12 mm^3.

ROC analysis   In order to further assess the difference that heteroscedasticity modeling would induce, we compared the sensitivity and specificity of the homoscedastic and heteroscedastic models with Receiver-Operating-Characteristics (ROC) curves. In order to draw the ROC curves, we used 122 p-values (uncorrected), equally spaced in the logarithmic space. ROC curves provide a way to measure the performance of a classifier without focusing on a single threshold. We calculated the area under the curve as an indicator of the classification accuracy. Since the size

of the expected detections is much smaller than the number of voxels that must not be detected, the area of interest in the ROC curve is the one of high specificity. That is why, as previously proposed in the literature, we focused on the area under the curve corresponding to false positive rates ranging from 0% to 10% [37]. As in the previous experiment, different smoothing kernels were studied.

3   Materials

This section starts with a presentation of the datasets under study (3.1). The pre-processing steps applied to the data before the statistical analysis are then described (3.2).

3.1   Data acquisition

25 patients diagnosed with brain tumors and 36 healthy volunteers were involved in this study. One control subject and four patients were excluded because of strong borderzone signs [49]. The final dataset therefore included 21 patients (13 males, 8 females, age: 55.2 ± 14.1 years) and 35 healthy volunteers (16 males, 19 females, age: 27.7 ± 6.4 years).

Data acquisition was performed on a 3T Siemens Verio MR scanner with a 32-channel headcoil. Patients were scanned in the context of clinical practice. The imaging protocol included a 3D T1-weighted anatomical sequence (T1w) (TR: 1900 ms, TE: 2.27 ms, FOV: 256 mm $\times$ 256 mm $\times$ 176 mm, flip angle: 9°, resolution: 1 mm $\times$ 1 mm $\times$ 1 mm), a PICORE Q2TIPS sequence [46] with crusher gradients (TR: 3000 ms, TE: 18 ms, FOV: 192 mm $\times$ 192 mm, flip angle: 90°, resolution: 3 mm $\times$ 3 mm, slice thickness: 7 mm, inter-slice gap: 0.7 mm, TI: 1700 ms, TI_{wd}: 700 ms, 60 repetitions, mSENSE parallel imaging with accelerating factor of 2). In addition to these sequences, the patients also underwent a 3D T1w post gadolinium (T1w-Gd) sequence (TR: 1900 ms, TE: 2.27 ms, flip angle: 9°, FOV: 250 mm $\times$ 250 mm $\times$176 mm, resolution: 1 mm $\times$ 1 mm $\times$ 1 mm) and a 2D T2w FLAIR sequence (TR: 9000 ms, TE: 90 ms, FOV: 220 mm $\times$ 199.4 mm, flip angle: 150°, resolution: 0.69 mm $\times$ 0.69 mm, slice thickness: 4 mm). Out of the 22 patients, 17 subjects also underwent a DSC sequence (GRE EPI, TR: 1500 ms, TE: 30 ms, FOV: 230 mm$\times$ 230 mm, flip angle: 90°, in plane resolution: 1.8 mm $\times$ 1.8 mm, slice thickness: 4 mm, inter-slice gap: 1.2 mm).

3.2   Pre-processing

In this subsection, we present how raw ASL images were processed in order to compute CBF maps (3.2.1) and then normalized in intensity (3.2.2). Lastly, DSC pre-processing is briefly described (3.2.3)

3.2.1   CBF estimation with ASL

Image pre-processing was performed using SPM8 (Statistical Parametric Mapping 8, Wellcome Department of Imaging Neuroscience, University College, London) in Matlab R2012a (Mathworks, Natick, MA). The anatomical image of each subject was segmented using the unified segmentation [4]. A subject-specific anatomical brain mask was created, excluding voxels with less than 50% of brain tissue in subsequent statistical analyses. A six-parameter rigid-body registration of the ASL volumes was carried out in order to reduce undesired effects due to subject motion. Rigid coregistration onto the whole-brain anatomical map was then performed based on mutual information. The average of unlabeled volumes was used to estimate the geometrical transformation to apply to each volume. Pair-wise subtraction of the control and labeled

scans was then computed. A standard kinetic model [8] was applied in order to get ASL CBF, according to the following equation:

$$f = 6000 \times \frac{\lambda \, \Delta M}{2 \, M_0 \, \alpha \, TI_{wd} \, \exp^{-(TI + idx_{sl} * TI_{sl})/T1_b}} \tag{21}$$

where f is the 4D CBF map in mL.100g^{-1}.min^{-1}, M_0 the acquired M$_0$ map (first volume of the ASL serie), $\lambda = 0.9$ mL.g^{-1} the blood/tissue water partition coefficient, $\alpha = 0.95$ the labeling efficiency, ΔM the 4D perfusion-weighted map, $TI = 1.7$ s the inversion time [15], idx_{sl} the slice index (0 for the first slice), $TI_{sl} = 0.045$ s the duration of acquisition of one slice, $TI_{wd} = 0.7$ s the temporal width of the bolus, $TI_b = 1.5$ s the T1 of blood [41]. The 6000 factor allows the conversion from mL.g^{-1}.s^{-1} to mL.100g^{-1}.min^{-1} which is the standard unit for CBF. We bring the attention of the reader to the fact that, contrary to what is usually done in ASL pre-processing, ΔM represents the set of perfusion-weighted maps (one volume per repetition) instead of a single perfusion-weighted map obtained by averaging across the repetitions. This is necessary in order to allow for the measurement of the within-subject variance.

Spatial normalization parameters estimated during the segmentation step were then applied to the T1 and ASL CBF maps in order to normalize the subjects into the ICBM-452 T1 template space [24]. This registration algorithm was selected since it gives good results even in the presence of large anatomical lesions [14].

3.2.2 Intensity normalization

Intensity normalization is a common pre-processing step in PET or SPECT analysis [3] where the measured values are not quantitative. ASL produces quantitative CBF maps, but the large inter-subject variability in global CBF [30] induces a strong correlation across voxels. In voxelwise ASL detection studies, where the focus is on local variations across the brain, intensity normalization is therefore advised to increase the sensitivity [5]. Given $CBF_{s,v}$, the original CBF value of subject s at voxel v, $Perf_{s,v}$, the normalized intensity value and θ_s the cross-voxel normalization parameter, we have:

$$CBF_{s,v} = \theta_s \times Perf_{s,v}. \tag{22}$$

The most widespread approach to estimate θ_s in ASL data processing is to compute the mean CBF signal found in a given region of interest (ROI). This ROI is either limited to an anatomical region, which is known to be avoided by the pathology under study, or covers the whole normal grey matter [5]. A threshold of 70% of grey matter is often chosen [32]. The latter approach was preferred in this study since, in a general setting, no brain region is free of anomalies across all pathologies. Unphysiological negative perfusion estimates were excluded from the normalization mask.

Furthermore, as areas of abnormal perfusion should not be included in the normalization mask, we used an iterative scheme. At the first step, the normalization parameter was computed based on a mask covering the entire grey matter of the patient map under study. A first pass outlined the corresponding perfusion abnormalities as described in section 2.1.3. On the next step, the detected voxels were excluded from the intensity normalization estimation. These two steps were iterated until convergence.

3.2.3 CBF estimation with DSC

The DSC images were processed using MR manufacturer software by manually choosing an arterial input function to calculate: the CBF, cerebral blood volume, and mean transit time

maps. The method is based on a deconvolution algorithm as described in [29]. Similarly to ASL, DSC CBF maps were coregistered on anatomical maps and spatially normalized.
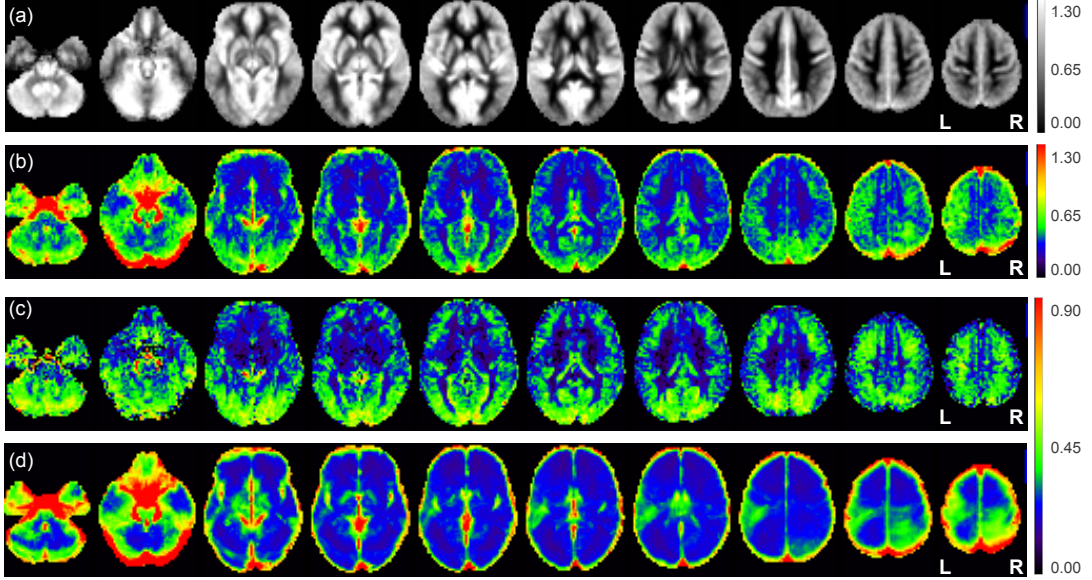

Figure 1: Parameter estimates of the homoscedastic and heteroscedastic model in the control group, computed from 35 healthy subjects. a) Mean perfusion estimate $\hat{\beta}_{controls}^{HETERO}$. b) Combined within- and between-subject standard deviation estimate from the homoscedastic model $\hat{\sigma}_{G_C}$. c) Between-subject standard deviation estimate from the heteroscedastic model $\hat{\sigma}_G$. d) Root square of the average within-subject variance in the control group. Perfusion is expressed in normalized units (ratio to mean grey matter perfusion). Axial slices are displayed in neurological convention.

4   Results

In this section we start with a graphical presentation of the parameter estimates of the homoscedastic and heteroscedastic models in the control group (4.1). Then, we present the results of the experiments regarding the validity of the assumptions of the homoscedastic model (4.2). In a last subsection, we display the results of the quantitative comparison between the homoscedastic and heteroscedastic models (4.3).

4.1   ASL template: a model of normal perfusion

In fig. 1, the parameter estimates of the homoscedastic and heteroscedastic models are displayed. Each parameter is defined as a voxelwise map. First, an estimate of the control group parameter is provided ($\hat{\beta}_{controls}^{HETERO}$ is displayed, the homoscedastic estimate $\hat{\beta}_{controls}^{HOMO}$ is visually nearly identical). As expected, the CBF is higher in the cortex and the basal ganglia than in white matter. Then, three standard deviation estimates are displayed:

- the combined within- and between-subject standard deviation estimate $\hat{\sigma}_{G_C}$ from the homoscedastic model (1 b);
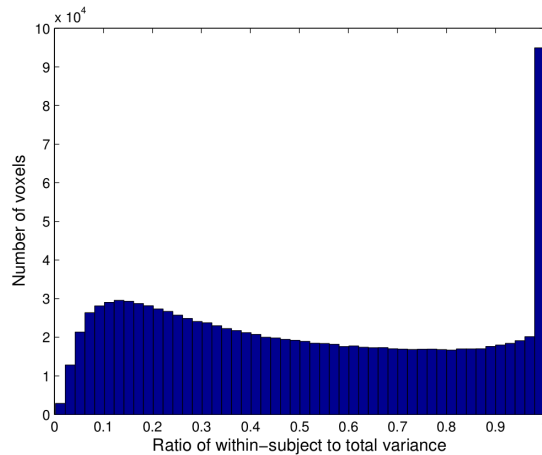
Figure 2: Histogram of the ratio of within-subject variance onto total variance in the control group. Both variance estimates have a significant impact depending on the voxels.

- the between-subject standard deviation estimate $\hat{\sigma}_G$ from the heteroscedastic model (1 c);

- the root square of the average within-subject variance estimates in the control group $(\frac{1}{n-1} \sum_{s=1}^{n-1} \hat{\sigma}_s^2)^{\frac{1}{2}}$. This map is not part of the estimated standard deviations but is provided as a visual example of expected within-subject standard deviation in the control group (1 d).

The high variance values observed in the vascular structures, such as the transverse sinus, are in concordance with the findings of [40]. This pattern is clearly visible in the combined within- and between-subject variance estimate of the homoscedastic model and captured by the within-subject variance estimate in the heteroscedastic model. The main variations observed in the between-subject variance, as estimated in the heteroscedastic model, are related to inter-subject misregistrations in the cortex. An increased variance is also visible in the occipital lobe, probably related to increased arterial transit times in these regions [21].

4.2   Testing the assumptions of the homoscedastic model

The homoscedastic model stands on one of the following assumptions: either the within-subject variance is negligible by comparison to between-subject variance, or the within-subject variance is roughly constant across subjects.

4.2.1   Relative weights of between-subject and within-subject variances in the control group

In order to verify if the within-subject variance can be assumed negligible by comparison to between-subject variance, fig. 2 presents the histogram of the ratio of within-subject to total variance in the control group (λ_s from eq. (20)). Overall, a total of 1 094 790 voxels are considered. Values close to one indicate a preponderance of the within-subject variance, whereas values close to zero outline a dominating between-subject variance. Clearly, both components of variance have an important impact as the λ_s index spans the complete range of values between 0 and 1. The large peak indicating voxels with a zero between-subject variance was also observed in
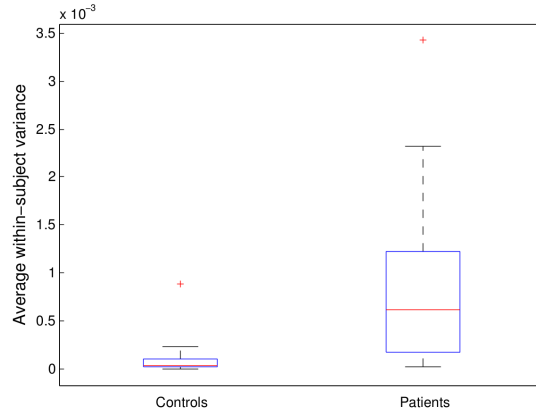
Figure 3: Average estimated within-subject variances in control and patient groups. For each box, the red line corresponds to the median and the top and bottom lines of the blue square are the upper and lower quartiles of the distribution.

[11] and might be the consequence of calculation inaccuracies. While the true between-subject variance might not be exactly zero, these voxels nevertheless present a very small between-subject variance, negligible by comparison to the within-subject variance. The quartiles of the distribution are 0.23 and 0.78, so that the outermost 25% of the voxels (on each side) have either a dominant within-subject variance or a dominant between-subject variance. In the remaining 50% voxels, both variance components have a significant impact. In conclusion, it cannot be assumed that within-subject variance is negligible in comparison to between-subject variance.

4.2.2 Cross-subject comparison of within-subject variances

In this experiment, we tested whether within-subject variance could be assumed roughly constant across subjects. As proposed by [27], we calculated an average within-subject variance across voxels. This led to an index of within-subject variance per subject. Fig. 3 displays the within-subject variance indexes in the control and patient groups.

In the control group, it is clear that one of the subjects presents an unexpected high variance by comparison to the other controls. Retrospectively, we found out that this subject moved substantially more than other controls and was identified as uncooperative by the MR physicist during the acquisition. As expected from fig. 3, we found one outlier in the control group with Rosner's test.

The within-subject variance was significantly higher in the patient group than in the control group ($p < 0.05$ with Kruskal-Wallis test). This might be explained by the fact that patients tend to have more difficulties to lie still during the acquisition due to their pathological condition.

The index proposed by [27] is an average across voxels and therefore focused on global variations of the within-subject variance. However, strong variations, sometimes caused by artefacts, that appear locally can also be a concern. As an example, fig. 4a displays the estimated within-subject standard deviation for a control subject presenting locally atypical patterns. While the high variance induced by the presence of large vessels is a pattern shared across subjects (as previously described in fig. 1, last panel), the high variance observed bilaterally in the frontal lobe is specific of this control subject. We hypothesize that these strong variations are the consequence of motion during the acquisition that was not correctly compensated during the pre-processing. Even if this subject was not previously outlined as an outlier, these atypical patterns of variance
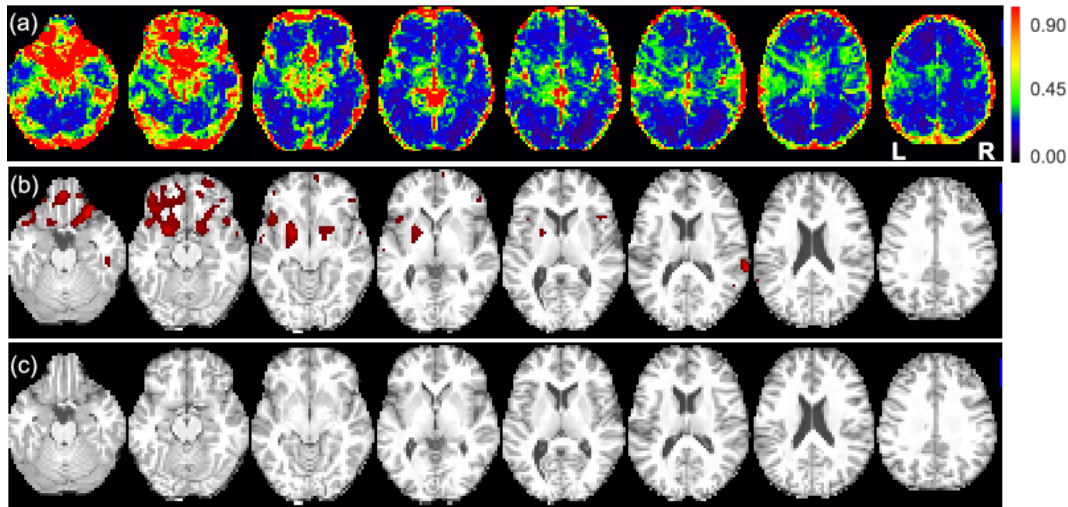
Figure 4: a) Estimated within-subject standard deviation in a control subject presenting locally atypical values. b) T1w map with false positive detections (in red) by leave-one-out cross-validation (smoothing FWHM = 8 mm^3) with the homoscedastic model (b) and the heteroscedastic model (c). No false positive detections are observed with the heteroscedastic model.

Table 1: Sensitivity and specificity in the control (first row) and patient (second and third rows) groups with the homoscedastic and heteroscedastic models for different smoothing kernels (FWHM = [0, 4, 6, 8, 10, 12] mm^3), at $p < 0.05$ FDR corrected.

	Homoscedastic model						Heteroscedastic model					
	0	4	6	8	10	12	0	4	6	8	10	12
Specificity (controls)	1.00	0.99	0.99	**0.99**	0.99	0.99	1.00	1.00	**1.00**	1.00	1.00	1.00
Specificity (patients)	0.92	0.89	0.87	**0.86**	0.85	0.84	1.00	0.99	**0.99**	0.99	0.99	0.98
Sensitivity (hyper)	0.49	0.59	0.63	**0.66**	0.67	0.67	0.51	0.59	**0.61**	0.58	0.57	0.55

might have a detrimental impact on the analysis if not properly taken into account.

In conclusion, variations across subjects of the within-subject variance appear to be important in ASL. This might be a consequence of the low SNR of this technique, since small artefacts in the original control and labeled scans can lead to substantial variations in the perfusion-weighted map after subtraction. In the next sections we investigate whether modeling heteroscedasticity can improve the detections of perfusion abnormalities.

4.3   Comparison of homoscedastic and heteroscedastic models

In this subsection, the homoscedastic and heteroscedastic models are compared in term of sensitivity and specificity to detect patient-specific perfusion abnormalities with ASL (as described in 2.1.3).

4.3.1   Fixed threshold

Quantitative analysis   Table 1 presents the sensitivity to detect hyper-perfusions and the specificity both in the patient group and by leave-one-out cross-validation in the control group
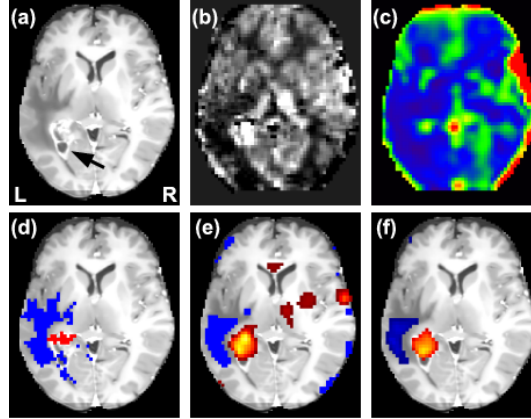
Figure 5: Detections of perfusion abnormalities with the homoscedastic and heteroscedastic models in a patient suffering from a high grade glioma. a) T1w-Gd map, the tumor site is pointed by a black arrow. b) Patient ASL CBF estimate $\hat{\beta}_n$. c) Within-subject standard deviation of ASL CBF $\hat{\sigma}_n^2$. d) T1w-Gd map with ground truth overlaid. T1w-Gd map with hypo- (blue color-map) and hyper-perfusions (hot color-map) overlaid for the homoscedastic (e) and heteroscedastic (f) models. Modeling heterogeneous variances (heteroscedastic model) reduces the false positive detections while preserving the true detections. Axial slices are displayed in neurological convention.

with the homoscedastic and heteroscedastic models for different smoothing kernels (FWHM = [0, 4, 6, 8, 10, 12] mm^3), at $p < 0.05$ FDR corrected.

Overall, the heteroscedastic model leads to a decrease in false positive rate. This is particularly noticeable in the patient group where the specificity is improved for each smoothing studied. As expected, in the control group, this effect is also observed but to a lesser extent. This is probably due to the fact that the hypothesis of homoscedasticity is better suited for the control subjects. The increase in false positive rate, with the homoscedastic model, is accompanied by a relative increase in sensitivity, which does not seem significant given the specificity loss.

Qualitative analysis At the individual level, as illustrated in fig. 4, local atypical patterns of variance can further lead to a substantial increase in false positives with the homoscedastic model. The unexpected high standard deviation in the frontal lobe for the control subject presented in fig. 4 leads to false positive hyper-perfusions with the homoscedastic model whereas the heteroscedastic model does not get any false positive (smoothing FWHM = 8 mm^3). Fig. 5 further illustrates the benefits of the heteroscedastic model compared to the homoscedastic model in a patient subject. The motion artefacts and hyper-signals induced by arteries indeed correspond to regions of high within-subject standard deviation. Modeling heterogeneous variances reduces the artefactual detections in these regions while preserving the quality of the true detections.

To investigate whether, at a same specificity rate, the sensitivity differs between the two approaches, we employed a ROC analysis as described in the next section.

4.3.2   ROC analysis

Quantitative comparison Fig. 6 presents the ROC curves in the patient group for the homoscedastic and heteroscedastic models. The average over the studied smoothing kernels is plotted
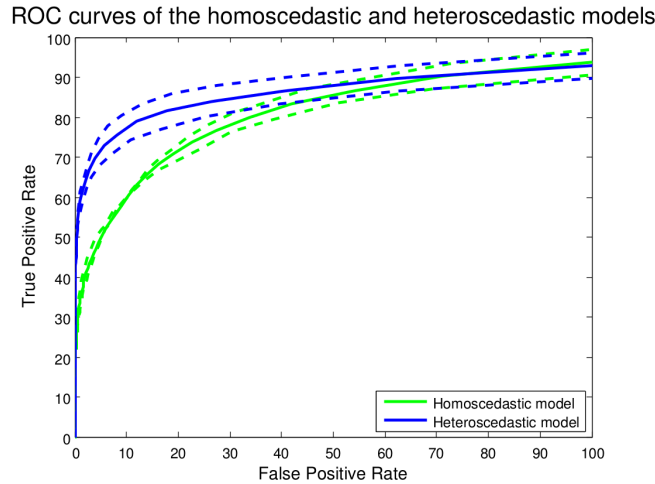
Figure 6: ROC curves for perfusion abnormality detections with the homoscedastic and heteroscedastic models. The average ROC curves across the studied smoothings are plotted in plain line. Dotted lines are plotted one standard deviation away from the average.

Table 2: Area under the ROC curve for false positive rates ranging between 0% and 10% with the homoscedastic and heteroscedastic models. The heteroscedastic model outperforms the homoscedastic model.

| | Homoscedastic model | | | | | | Heteroscedastic model | | | | | |
	0	4	6	8	10	12	0	4	6	8	10	12
Area under the ROC curve	0.46	**0.49**	**0.49**	**0.49**	0.48	0.48	0.63	0.70	**0.72**	**0.72**	0.69	0.65

along with the spread provided by the standard deviation over the 6 measurements. The ROC curve of the heteroscedastic model is substantially closer to the ideal classifier than the homoscedastic curve.

Table 2 presents the area under the ROC curves for false positives rates ranging from 0 to 10%, for the homoscedastic and heteroscedastic models with different smoothing kernels. The heteroscedastic model outperforms the homoscedastic model with an increased area under the curve. The best values are 0.72 and 0.49 with the heteroscedastic and homoscedastic model respectively. For both approaches, a maximum area under the curve is reached for a smoothing kernel of 8 mm^3.

Qualitative comparison In order to illustrate the advantage of the heteroscedastic over the homoscedastic model, we chose 3 representative subjects and compared the methods at fixed false positive rate and true positive rate. To this aim, we selected the uncorrected p-values that would lead to a pre-specified false positive rate (respectively true positive rate) from the ROC analysis. We worked with data smoothed with a Gaussian kernel of 8 mm^3 that led to the best area under the curve with both models. It is worth noting that in the previous section the sensitivity and specificity were estimated at the group level and that the 3 subjects presented in this part were chosen so that they would best illustrate the group findings. Fig. 7 presents the detections obtained with both methods on the 3 selected patients. The first subject, which presents a small hyper-perfusion, is studied at a true positive rate of 50%. In the two remaining

patients, the methods are compared at a false positive rate of 0.1%.

Patient 15 suffers from a gliosarcoma in the left hemisphere close to the parahippocampal region. The lesion displays a small hyper-perfusion in its dorsal part, small hypo-perfusions are seen in the surrounding edema. At a false positive rate of 50%, the homoscedastic model displays a larger number of false positives than the heteroscedastic model. Patient 16 was diagnosed with a high grade tumor in the left temporal lobe. The lesion is characterized by a large hyper-perfusion and a surrounding hypo-perfusion. With a false positive rate of 0.1%, the hyper-perfusion is correctly located with both methods. The extent of the hyper-perfusion is however better covered by the heteroscedastic model. Patient 6 suffers from a meningioma partly hyperperfused. Similarly to patient 16, at a false positive rate of 0.1%, both methods detect the hyper-perfusion but the heteroscedastic model is clearly more sensitive.

These 3 cases illustrate the loss of sensitivity of the homoscedastic model by comparison to the heteroscedastic model at the same specificity level.

5   Discussion and conclusion

We have compared two approaches to quantitatively outline patient-specific pathological patterns of abnormal perfusion in ASL data based on the massively univariate GLM: the homoscedastic and heteroscedastic models.

We demonstrated that the assumptions underlying the homoscedastic model are not verified in ASL studies. More precisely, the within-subject variance cannot be considered as negligible by comparison to between-subject variance, nor constant across subjects. In fMRI data, small deviations from homoscedasticity have shown to not overly alter the results in one-sample mixed-effects GLM analysis [27]. Here, we showed that modeling heterogeneous within-subject variances is essential in order to reach a satisfactory level of specificity in a mixed-effect two-sample t-test comparing a patient to a group of controls in ASL. These results are in line with a recent study by [11] where heteroscedastic mixed-effects GLM were shown to provide more accurate results in fMRI.

In the context of this paper, we defined a single parameter in the subject-level design matrix and focused on perfusion-weighted images obtained after pair-wise subtraction of the control and labeled scans. Other authors have suggested that taking the complete ASL time-course into account (before subtraction) would lead to more efficient estimates of perfusion in fMRI [26]. Also, additional regressors can be introduced as nuisance covariates for denoising purposes in the subject-level design matrix as suggested in [42]. Modifying the subject-level design matrix would change the estimated subject parameters and their sampling variances. However, an heteroscedastic model would still be appropriate in this setting so that the conclusions of this paper remain valid even with a different subject-level model.

We demonstrated that within-subject variance captures important information regarding the subject-specific spatial distribution of noise in ASL data. We also outlined that patient-specific brain perfusion abnormalities can be correctly detected using ASL if the heterogeneous within-subject variances are properly modeled. We therefore advise the use of heteroscedastic models in ASL studies.

References

[1] G K Aguirre, J a Detre, E Zarahn, and D C Alsop. Experimental design and the relative sensitivity of BOLD and perfusion fMRI. *NeuroImage*, 15(3):488–500, 2002.
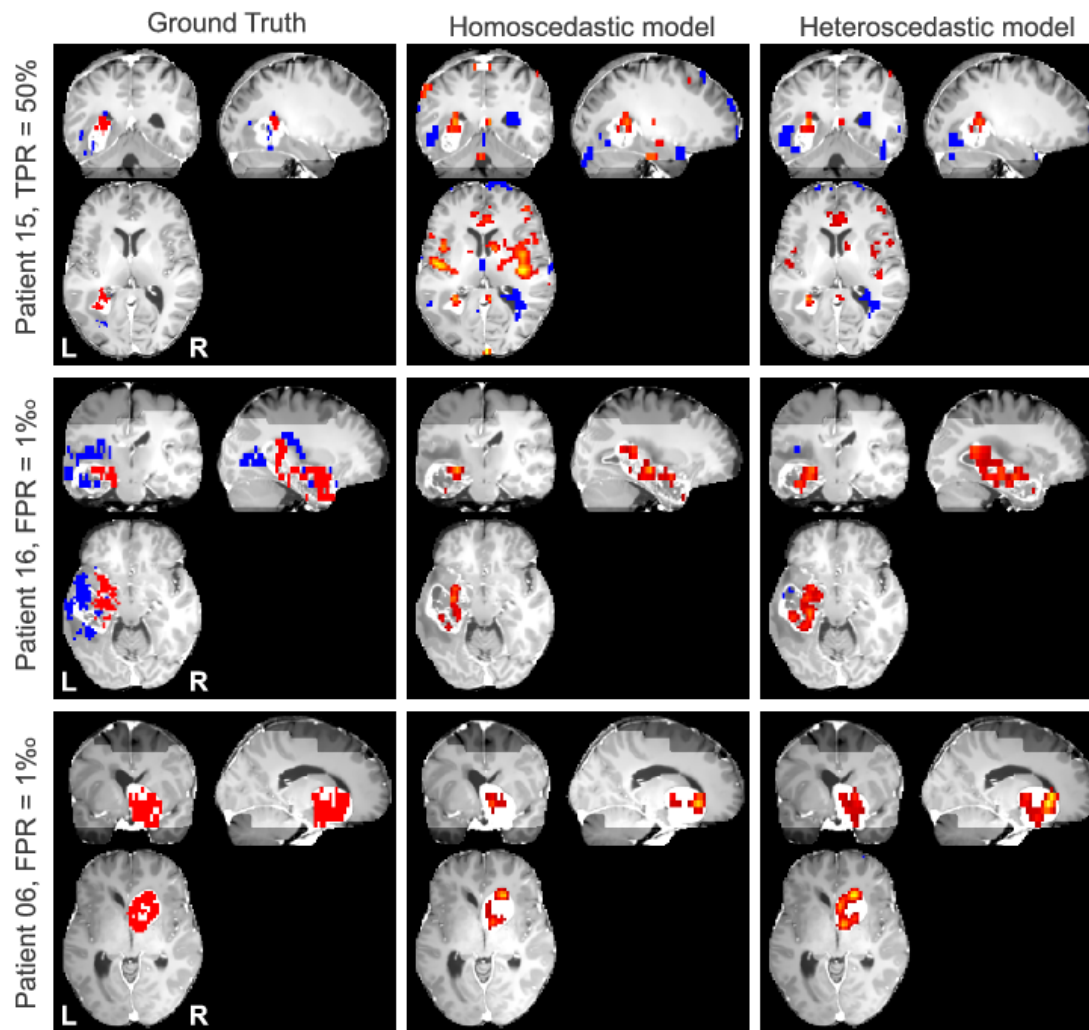
Figure 7: Perfusion abnormalities detections in 3 patients with homoscedastic and heteroscedastic models. From left to right: ground truth hyper-perfusions (red) and hypo-perfusions (blue) overlaid on the T1w-Gd map; detections with the homoscedastic model with a smoothing kernel of FWHM = 8 mm^3, hyper-perfusions (hot colormap) and hypo-perfusions (blue); detections with the heteroscedastic model with a smoothing kernel of FWHM = 8 mm^3, hyper-perfusions (hot colormap) and hypo-perfusions (blue).

[2] Geoffrey K Aguirre, E Zarahn, and M D'Esposito. Empirical analyses of BOLD fMRI statistics. II. Spatially smoothed data collected under null-hypothesis and experimental conditions. *NeuroImage*, 5(3):199–212, April 1997.

[3] S Arndt, T Cizadlo, D O'Leary, S Gold, and N C Andreasen. Normalizing counts and cerebral blood flow intensity in functional imaging studies of the human brain. *NeuroImage*, 3(3 Pt 1):175–84, 1996.

[4] John Ashburner and Karl J Friston. Unified segmentation. *NeuroImage*, 26(3):839–851, 2005.

[5] Sina Aslan and Hanzhang Lu. On the sensitivity of ASL MRI in detecting regional differences in cerebral blood flow. *Magnetic Resonance Imaging*, 28(7):928–935, 2010.

[6] Christian F. Beckmann, Mark Jenkinson, and Stephen M Smith. General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, 20(2):1052–1063, 2003.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[8] Richard B. Buxton, Lawrence R. Frank, Eric C. Wong, B Siewert, S Warach, and R R Edelman. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magnetic Resonance in Medicine*, 40(3):383–396, 1998.

[9] J.A. Chalela, David C Alsop, J.B. Gonzalez-Atavales, J.A. Maldjian, S.E. Kasner, and John A Detre. Magnetic resonance perfusion imaging in acute ischemic stroke using continuous arterial spin labeling. *Stroke*, 31(3):680–687, 2000.

[10] S Chawla, Sumei Wang, Robert Christian Wolf, J H Woo, Jiongjiong Wang, D M O'Rourke, K D Judy, M S Grady, E R Melhem, and H Poptani. Arterial spin-labeling and MR spectroscopy in the differentiation of gliomas. *American Journal of Neuroradiology*, 28(9):1683–1689, 2007.

[11] Gang Chen, Ziad S Saad, Audrey R Nath, Michael S Beauchamp, and Robert W Cox. FMRI group analysis combining effect estimates and their variances. *NeuroImage*, 60(1):747–765, 2012.

[12] Tai-Yuan Chen, Lee Chiu, Tai-Ching Wu, Te-Chang Wu, Chien-Jen Lin, Shih-Chuan Wu, and Yu-Kun Tsui. Arterial spin-labeling in routine clinical practice: a preliminary experience of 200 cases and correlation with MRI and clinical findings. *Clinical imaging*, 36(4):345–352, 2012.

[13] R W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and biomedical research, an international journal*, 29(3):162–73, June 1996.

[14] Jenny Crinion, John Ashburner, Alex Leff, Matthew Brett, Cathy Price, and Karl Friston. Spatial normalization of lesioned brains: performance evaluation and impact on fMRI analyses. *NeuroImage*, 37(3):866–875, 2007.

[15] Jean-Christophe Ferré, Jan Petr, Elise Bannier, Christian Barillot, and Jean-Yves Gauvrit. Improving quality of arterial spin labeling MR imaging at 3 tesla with a 32-channel coil and parallel imaging. *Journal of Magnetic Resonance Imaging*, 1239:1233–1239, 2012.

[16] Karl J. Friston, K E Stephan, T E Lund, A Morcom, and S Kiebel. Mixed-effects and fMRI studies. *NeuroImage*, 24(1):244–52, 2005.

[17] B Hakyemez, C Erdogan, I Ercan, N Ergin, S Uysal, and S Atahan. High-grade and low-grade gliomas: differentiation by using perfusion MR imaging. *Clinical radiology*, 60(4):493–502, 2005.

[18] AP Holmes and KJ Friston. Generalisability, Random Effects & Population Inference. *Proceedings of Fourth International Conference on Functional Mapping of the Human Brain, NeuroImage 7, S754*, 1998.

[19] S Huck, H U Kerl, M Al-Zghloul, C Groden, and I Nölte. Arterial spin labeling at 3.0 Tesla in subacute ischemia: comparison to dynamic susceptibility perfusion. *Clinical neuroradiology*, 22(1):29–37, 2012.

[20] Martin a Lindquist, Julie Spicer, Iris Asllani, and Tor D Wager. Estimating and testing variance components in a multi-level GLM. *NeuroImage*, 59(1):490–501, January 2012.

[21] Bradley J MacIntosh, Nicola Filippini, Michael a Chappell, Mark W. Woolrich, Clare E Mackay, and Peter Jezzard. Assessment of arterial arrival times derived from multiple inversion time pulsed arterial spin labeling MRI. *Magnetic Resonance in Medicine*, 63(3):641–647, 2010.

[22] Camille Maumet, Pierre Maurel, Jean-Christophe Ferré, and Christian Barillot. A contrario detection of focal brain perfusion abnormalities based on an ASL template. In *Proceedings of the IEEE 9th International Symposium on Biomedical Imaging (ISBI 2012)*, pages 1176–1179, 2012.

[23] Camille Maumet, Pierre Maurel, Jean-Christophe Ferré, and Christian Barillot. A comprehensive framework for the detection of individual brain perfusion abnormalities using arterial spin labeling. In *Proceedings of the 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2012)*, pages 542–549, 2012.

[24] J Mazziotta, A Toga, and A Evans. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

[25] Sébastien Mériaux, Alexis Roche, Ghislaine Dehaene-Lambertz, Bertrand Thirion, and Jean-Baptiste Poline. Combined permutation test and mixed-effect model for group average analysis in fMRI. *Human brain mapping*, 27(5):402–10, May 2006.

[26] Jeanette a Mumford, Luis Hernandez-Garcia, Gregory R Lee, and Thomas E Nichols. Estimation efficiency and statistical power in arterial spin labeling fMRI. *NeuroImage*, 33(1):103–14, October 2006.

[27] Jeanette A Mumford and Thomas Nichols. Simple group fMRI modeling and inference. *NeuroImage*, 47(4):1469–1475, 2009.

[28] T Noguchi, T Yoshiura, A Hiwatashi, O Togao, K Yamashita, E Nagao, T Shono, M Mizoguchi, S Nagata, T Sasaki, S O Suzuki, T Iwaki, K Kobayashi, F Mihara, and H Honda. Perfusion imaging of brain tumors using arterial spin-labeling: correlation with histopathologic vascular density. *American Journal of Neuroradiology*, 29(4):688–93, 2008.

[29] L Ostergaard, R M Weisskoff, D A Chesler, C Gyldensted, and B R Rosen. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. Part I: Mathematical approach and statistical analysis. *Magnetic Resonance in Medicine*, 36(5):715–725, 1996.

[30] Laura M Parkes and John A Detre. ASL: Blood Perfusion Measurements Using Arterial Spin Labelling. In Paul Tofts, editor, *Quantitative MRI of the Brain: Measuring Changes Caused by Disease.*, chapter 13. John Wiley & Sons, Ltd, 2003.

[31] William D Penny and A J Holmes. Random-Effects Analysis. In Elsevier, editor, *Human Brain Function*, pages 843–850. San Diego, 2004.

[32] Jan Petr, Jean-Christophe Ferre, Elise Bannier, Helene Raoult, Jean-Yves Gauvrit, and Christian Barillot. Construction and evaluation of a quantitative arterial spin labeling brain perfusion template at 3T. In *Proceedings of the 8th IEEE International Symposium on Biomedical Imaging (ISBI 2011)*, pages 1035–1038, 2011.

[33] Jan Petr, Jean-Christophe Ferré, HÃ©lÃ¨ne Raoult, Elise Bannier, Jean-Yves Gauvrit, and Christianan Barillot. Template-based approach for detecting motor task activation-related hyperperfusion in pulsed asl data. *Human Brain Mapping*, (in press), 2012.

[34] Amy Pinkham, James Loughead, Kosha Ruparel, Wen-Chau Wu, Eve Overton, Raquel Gur, and Ruben Gur. Resting quantitative cerebral blood flow in schizophrenia measured by pulsed arterial spin labeling perfusion MRI. *Psychiatry research*, 194(1):64–72, 2011.

[35] Russell A. Poldrack, Jeanette Mumford, and Thomas Nichols. *Handbook of functional MRI data analysis.* Cambridge University Press, Cambridge, 2011.

[36] B Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

[37] P Skudlarski, R T Constable, and J C Gore. ROC analysis of statistical methods used in functional MRI: individual subjects. *NeuroImage*, 9(3):311–329, 1999.

[38] T Sugahara, Y Korogi, S Tomiguchi, Y Shigematsu, I Ikushima, T Kira, L Liang, Y Ushio, and M Takahashi. Posttherapeutic intraaxial brain tumor: the value of perfusion-sensitive contrast-enhanced MR imaging for differentiating tumor recurrence from nonneoplastic contrast-enhancing tissue. *American Journal of Neuroradiology*, 21(5):901–909, 2000.

[39] Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, 35(1):105–20, March 2007.

[40] Roberto Viviani, Eun-Jin Sim, Hanna Lo, Sven Richter, Sebastian Haffer, Nadine Osterfeld, Jan Thöne, and Petra Beschoner. Components of variance in brain perfusion and the design of studies of individual differences: the baseline study. *NeuroImage*, 46(1):12–22, 2009.

[41] Yang Wang, Andrew J Saykin, Josef Pfeuffer, Chen Lin, Kristine M Mosier, Li Shen, Sungeun Kim, and Gary D Hutchins. Regional reproducibility of pulsed arterial spin labeling perfusion imaging at 3T. *NeuroImage*, 54(2):1188–1195, 2011.

[42] Ze Wang. Improving cerebral blood flow quantification for arterial spin labeled perfusion MRI by removing residual motion artifacts and global signal fluctuations. *Magnetic resonance imaging*, July 2012.

[43] Carsten Warmuth, Matthias Günther, and Claus Zimmer. Quantification of Blood Flow in Brain Tumors: Comparison of Arterial Spin Labeling and Dynamic Susceptibility-weighted Contrast-enhanced MR Imaging. *Radiology*, 228(4):523–532, 2003.

[44] M A Weber, S Zoubaa, M Schlieter, E Jüttler, H B Huttner, K Geletneky, C Ittrich, M P Lichy, A Kroll, J Debus, F L Giesel, M Hartmann, and M Essig. Diagnostic performance of spectroscopic and perfusion MRI for distinction of brain tumors. *Neurology*, 66(12):1899–1906, 2006.

[45] Max Wintermark, Musa Sesay, Emmanuel Barbier, Katalin Borbély, William P Dillon, James D Eastwood, Thomas C Glenn, Cécile B Grandin, Salvador Pedraza, Jean-François Soustiel, Tadashi Nariai, Greg Zaharchuk, Jean-Marie Caillé, Vincent Dousset, and Howard Yonas. Comparative overview of brain perfusion imaging techniques. *Stroke*, 36(9):e83–e99, 2005.

[46] Eric C. Wong, Richard B. Buxton, and Lawrence R. Frank. Quantitative imaging of perfusion using a single subtraction (QUIPSS and QUIPSS II). *Magnetic Resonance in Medicine*, 39(5):702–708, 1998.

[47] Mark W. Woolrich, Timothy E J Behrens, Christian F. Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–47, April 2004.

[48] Keith J Worsley, C H Liao, J Aston, V Petre, G H Duncan, F Morales, and a C Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, January 2002.

[49] Greg Zaharchuk, Roland Bammer, Matus Straka, Ajit Shankaranarayan, David C Alsop, Nancy J Fischbein, Scott W Atlas, and Michael E Moseley. Arterial spin-label imaging in patients with normal bolus perfusion-weighted MR imaging findings: pilot identification of the borderzone sign. *Radiology*, 252(3):797–807, 2009.

[50] Greg Zaharchuk, Ibraheem S El Mogy, Nancy J Fischbein, and Gregory W Albers. Comparison of arterial spin labeling and bolus perfusion-weighted imaging for detecting mismatch in acute stroke. *Stroke*, 43(7):1843–1848, 2012.