STEAM - Statistical Template Estimation for Abnormality Mapping: a Personalized DTI Analysis Technique with Applications to the Screening of Preterm Infants

Brian G. Booth[a], Steven P. Miller[b,e,f], Colin J. Brown[a], Kenneth J. Poskitt[c,e], Vann Chau[b,c,f], Ruth E. Grunau[d,e], Anne R. Synnes[d,e], Ghassan Hamarneh[a]

[a] *Simon Fraser University, Burnaby, Canada*
[b] *The Hospital for Sick Children, Toronto, Canada*
[c] *British Columbia's Children's and Women's Health Hospitals, Vancouver, Canada*
[d] *The Child & Family Research Institute, Vancouver, Canada*
[e] *University of British Columbia, Vancouver, Canada*
[f] *University of Toronto, Toronto, Canada*

Abstract

We introduce the STEAM DTI analysis engine: a whole brain voxel-based analysis technique for the examination of diffusion tensor images (DTIs). Our STEAM analysis technique consists of two parts. First, we introduce a collection of statistical templates that represent the distribution of DTIs for a normative population. These templates include various diffusion measures from the full tensor, to fractional anisotropy, to 12 other tensor features. Second, we propose a voxel-based analysis (VBA) pipeline that is reliable enough to identify areas in individual DTI scans that differ significantly from the normative group represented in the STEAM statistical templates. We identify and justify choices in the VBA pipeline relating to multiple comparison correction, image smoothing, and dealing with non-normally distributed data. Finally, we provide a proof of concept for the utility of STEAM on a cohort of 134 very preterm infants. We generated templates from scans of 55 very preterm infants whose T1 MRI scans show no abnormalities and who have normal neurodevelopmental outcome. The remaining 79 infants were then compared to the templates using our VBA technique. We show: (a) that our statistical templates display the white matter development expected over the modeled time period, and (b) that our VBA results detect abnormalities in the diffusion measurements that relate significantly with both the presence of white matter lesions and with neurodevelopmental outcomes at 18 months. Most notably, we show that STEAM produces personalized results while also being able to highlight abnormalities across the whole brain and at the scale of individual voxels. While we show the value of STEAM on DTI scans from a preterm infant cohort, STEAM can be equally applied to other cohorts as well. To facilitate this whole-brain personalized DTI analysis, we made STEAM publicly available at `http://www.sfu.ca/~bgb2/steam`.

Keywords:
Diffusion Tensor Imaging, Preterm Infants, Brain Development, Statistical modeling, Voxel-based Analysis, Outcome Prediction

1. Introduction

Worldwide, more than one in ten infants are born prematurely (earlier than 37 weeks gestational age) and are at high risk of adverse neurodevelopmental outcome [1]. This abnormal neurodevelopment is believed to be due to white matter dysmaturation or injuries acquired over the period of the infant's neonatal intensive care [2, 3]. As a result, there has been a strong effort in identifying these white matter abnormalities early. The earlier these abnormalities are detected, the sooner clinicians can intervene and either improve a preterm infant's neurodevelopmental health, or set up appropriate rehabilitative care to aid in that child's growth and maturation.

Diffusion tensor imaging (DTI) provides us with the ability to probe both white matter organization and integrity, potentially making it a valuable tool to identify such white matter abnormalities. That potential has been examined in recent studies with various links being identified between diffusion measures (like fractional anisotropy [FA] and mean diffusivity [MD]) and neurodevelopmental outcome [4, 5, 6, 7, 8, 9, 10, 11]. These group-based studies have provided us with a further understanding of how DTI-based abnormalities could indicate future neurodevelopmental delay, but they stop short of applying those conclusions to individual cases. As a result, there's still the open question of, "Given the DTI scan of a single preterm infant, what cues can we extract from that *one* scan to gauge whether that infant will have an adverse neurode-

*Corresponding author at: School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. Tel: 1-778-782-5509, *Email address:* bgb2@sfu.ca (Brian G. Booth)

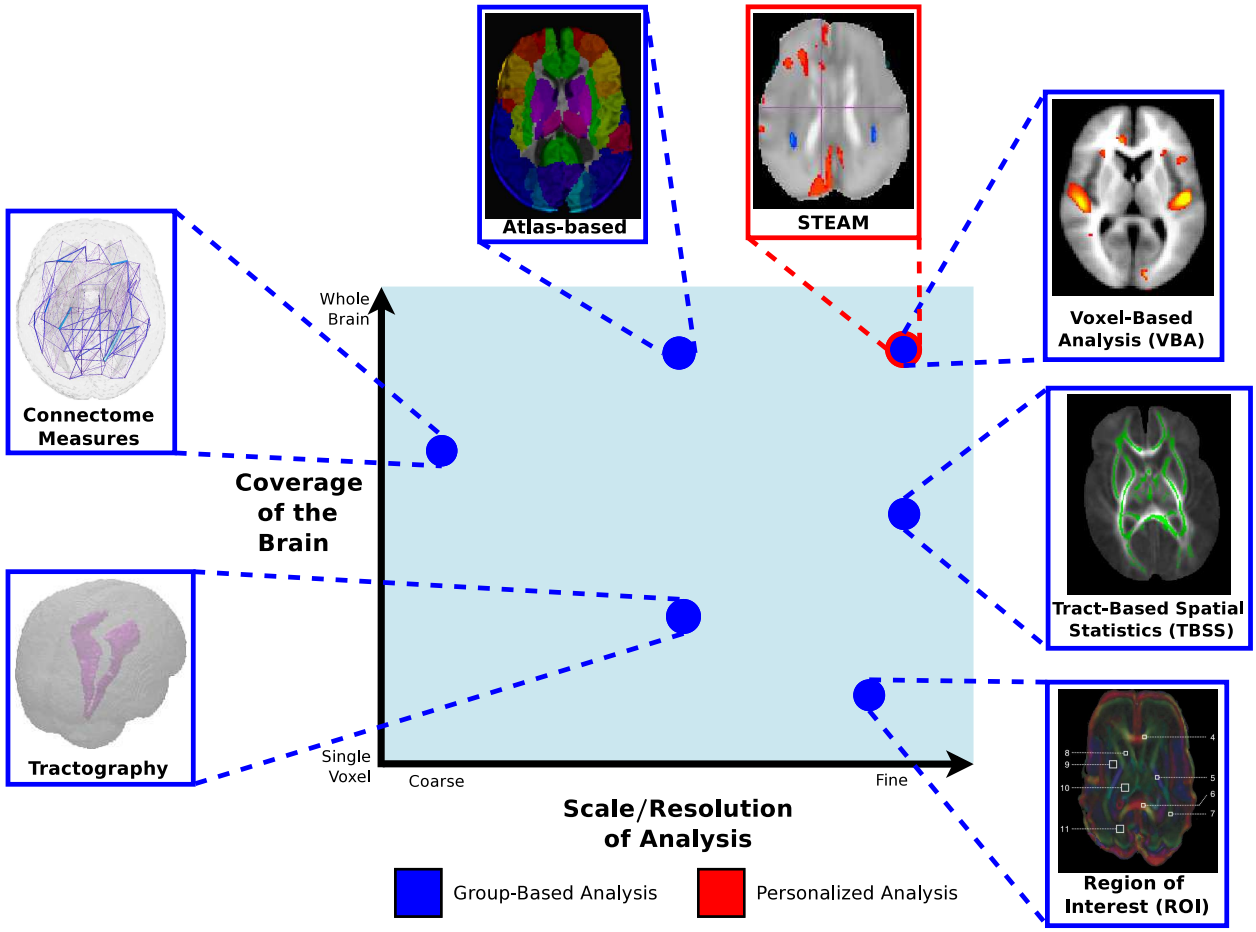


Figure 1: Popular preterm infant DTI analysis techniques presented according to the amount of the brain included in the analysis (coverage), the resolution at which the statistical analysis is undertaken (scale), and whether the analysis technique is group-based or personalized. Note that our proposed techinque, STEAM, is the only technique that provides a personalized, fine-scale analysis of the whole brain.

velopmental outcome?"

The question of subject-specific outcome projection has been examined in structural MRI with observer rating systems being proposed based on the number and size of white matter lesions [12] or on the presence of intraventricular hemorrhages (IVH) [13]. However, these techniques are specific to structural MRI and do not harness the potential DTI has of providing additional diagnostic information. Alternatively, an argument could be made that recent DTI group studies, like those cited earlier, identify where to look and what to look for. Unfortunately if an experimental group contains a wide variety of abnormalities (compared to a control group), that experimental group variance makes it a challenge to identify statistically significant differences in group studies. This limitation in group studies is a concern with preterm infants as recent studies have noted that adverse neurodevelopmental outcomes in preterm infants can manifest themselves in multiple ways [2]. This intra-group variability in preterm infant group-level analysis might be masking certain outcome-predictive image cues. To capture this intra-group variability and determine its relevance, we require a technique that produces a personalized result. By identifying abnormal diffusion measurements at the level of an individual DTI scan, we would have the potential to identify and examine the impact of variability within experimental or control groups.

Our goal is to generate a subject-specific DTI analysis technique which can flag brain regions with abnormal DTI measurements that are indicative of future neurodevelopmental delay. Such an analysis technique can take on many forms, as is evidenced by the presence of multiple comparable group-based analysis techniques. Figure 1 summarizes group-based analysis techniques, placing them according to their *scale* (i.e., the spatial resolution at which the statistical analysis is performed) and their *coverage* (i.e., the amount of the DTI scan that is analyzed). In terms of scale, existing techniques range from computing statistics at the level of each scan (e.g., connectome mapping [14]), to the level of each segmented region (e.g., tractography-based [8] or atlas-based [11]), to the level of individual voxels (e.g., Tract-based Spatial Statistics [7]). In terms of coverage, we see techniques that cover small regions of interest (e.g., ROI-based analysis [9, 10]), to the white matter skeleton (e.g., Tract-based Spatial Statistics [7]), to the whole brain (e.g., Voxel-based analysis [4]).

Different choices of scale and coverage result in algorithms with different strengths. While a fine scale analysis technique has the ability to accurately localize specific abnormalities, a coarse scale analysis technique can detect small but widespread changes in the brain [15]. Similarly, a full-brain analysis is valuable in an exploratory setting where the location of structural abnormalities are not known a priori. On the other hand, a localized analysis allows for the examination of specific brain structures without introducing potential confounding factors from the rest of the brain.

In our context of subject-specific DTI screening, we would prefer an exploratory analysis technique that can localize potential abnormalities. Covering the whole brain would be valuable as recent ROI-based studies in subcortical white matter [10] and cortical gray matter [16] suggest that not only are patterns of maturation observable within those regions on a DTI scan, but also that deviation from the normal maturation pattern could be indicative of adverse neurodevelopmental outcome. Adding those regions to the more frequently studied deep white matter would provide us with greater potential to identify clinically valuable abnormalities. Equally, analyzing a DTI scan at the scale of its individual voxels would give us a greater ability to identify and localize regions with abnormal diffusion measurements. This advantage features prominently in the voxel-based analysis works of Aeby et al. [4] and Gimenez et al. [17] where diffusion measurements in small, but clinically important, brain regions were identified as being related to neurodevelopmental outcome.

It is evident from Figure 1 that voxel-based analysis (VBA) [18] provides the greatest coverage and finest scale of analysis compared to other techniques. Using two groups of scans, VBA spatially aligns all scans from both groups, then computes statistical tests at each voxel, resulting in a fine-scale, group-based, statistical analysis of diffusion measurements over the whole brain. While VBA is attractive due to its ability to maximize scale and coverage, it is currently limited to group-based studies and is susceptible to various design decisions in the analysis pipeline [19, 20]. In particular, the choice of image registration algorithm, the size of the image smoothing kernel, the choice of multiple comparison correction scheme, and the decision of how to handle non-normally distributed data can all impact the conclusions that one draws from a VBA analysis [19]. As a result of these VBA susceptibilities, one has to take care in setting up and documenting a VBA pipeline.

With these points in mind, we desire the ability to perform a subject-specific analysis of a DTI scan in a similar fashion to VBA while obtaining a reliable result that can be related to neurodevelopmental outcome. It is towards this goal that we introduce STEAM: Statistical Template Estimation for Abnormality Mapping. The STEAM technique consists of two parts. First, we generate a collection of 3D statistical template images that capture, at the scale of each individual voxel, the distribution of diffusion measurements for a group a preterm infants with normal developmental outcome. This template collection acts as our normative statistical model and can be computed offline (i.e., during downtime) from a control group's DTI scans. Second, we use these statistical templates to perform VBA by spatially aligning a new DTI scan to the corresponding template and applying single-sample statistical tests. In this fashion, we are able - for the first time - to compare a single preterm infant's DTI scan to a normative preterm population and generate results at the level of individual voxels. As part of this VBA-style analysis, we provide a thorough review of what choices we make in our analysis pipeline to ensure a reliable outcome for this form of whole brain DTI analysis. We also provide the code for STEAM, as well as our statistical templates, online to those who wish to use this analysis technique[1]. Finally, we show that our STEAM analysis engine - the incorporation of our preterm statistical DTI templates into a VBA pipeline - can provide further insights into how abnormal diffusion measurements can impact the neurodevelopment of individual subjects, a benefit that existing analysis technique cannot provide.

The remainder of the paper is organized as follows. Section 2 provides an overview of our STEAM analysis engine, including the creation of statistical templates (in Section 2.1), the voxel-based analysis (in Section 2.2), and the generation of a full collection of DTI templates (in Section 2.3). Section 3 provides an overview of the cohort we scanned to validate this technique, the imaging parameters we used, and the decisions made on which DTI scans to include in the statistical templates. Section 4, we examine and validate STEAM by comparing our statistical templates and VBA results to existing literature. Finally, in Section 5, we conclude with a discussion on the suitability of VBA for preterm DTI analysis and the potential for this technique in future studies.

2. Methods: The STEAM Analysis Engine

At a high level, STEAM works by producing a statistical model for the DTI scans of healthy preterm infant brains, then compares a new DTI scan to that model. We refer to the model creation step as *statistical template estimation* (i.e., the STE in STEAM) while we call the model comparison step *abnormality mapping* (i.e., the AM in STEAM). We present both steps in the following sub-sections before rounding out STEAM by addressing different diffusion measurements (e.g., FA, MD, etc...) and different image scales.

2.1. Statistical Template Estimation

To facilitate subject-specific VBA for preterm infants, we first establish a statistical reference model for the normative preterm DTI brain scan. We provide this model

[1] `http://www.sfu.ca/~bgb2/steam`

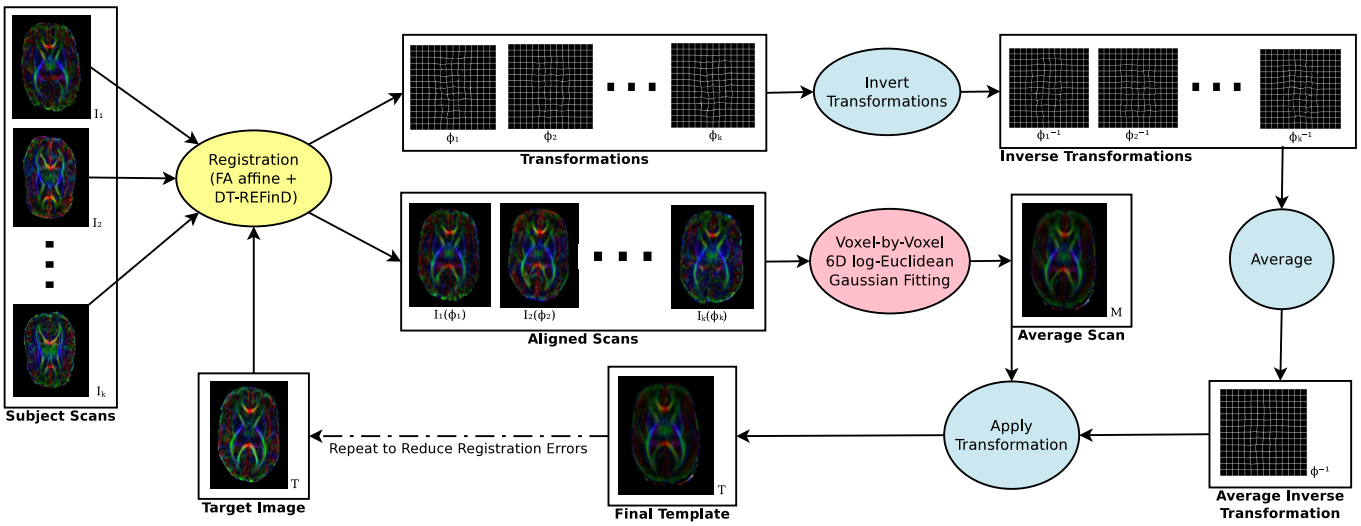


Figure 2: Flow diagram of the template creation procedure of Guimond et al. [21]. The subject's scans are aligned to a given target image and averaged. The corresponding image transformations are inverted, averaged, then applied to the average image to adjust the template to an unbiased shape and size. Multiple iterations are then done to reduce registration error. Each step in the pipeline is colour-coded by section, with the image alignment (yellow) discussed in Section 2.1.1, the model fitting (red) discussed in Section 2.1.2, and the bias correction steps (blue) discussed in Section 2.1.3.

through the construction of population-specific statistical templates. These templates are computed offline from a collection of DTI scans from a normative control group, resulting in a succinct statistical model of a normal preterm infant population.

The templates are generated using a DTI-extended version of the scalar image atlas-building technique of Guimond et al. [21]. An overview of the technique is presented in Figure 2. The technique of Guimond et al. involves three basic steps: (a) aligning all given scans to a chosen target image T; (b) computing, at each voxel $\mathbf{x}$, the mean $M(\mathbf{x})$ and (co-)variance $S(\mathbf{x})$ of the image data, and (c) transforming the resulting mean and (co-)variance images to an "average" frame of reference that is not biased by the choice of the initial target image T.

2.1.1. Image alignment

The template creation procedure of Guimond et al. begins with a sample set of DTI scans, $W = \{I_1, \cdots, I_k\}$, from a control group and the objective of spatially aligning all scans to the same frame of reference. In order to begin this alignment process, a scan from the group is selected as our initial target image T and all other DT images $I_i (i \in [1, k])$ are registered to the target using the state of the art in DT image registration techniques. While the choice of target introduces a bias on brain shape and size, we discuss how to correct for that bias in Section 2.1.3.

The registration is performed by independently aligning each DTI scan I_i to the chosen target image T in two steps. First, we perform an affine registration using FSL's Linear Image Registration Tool (FLIRT) [22, 23] to remove any pose or scale differences between the given image I_i and the target T. The affine transformation obtained from this registration step was one that maximized the normalized mutual information between the FA of the given image I_i and that of the target T. Thus, this registration step aligns each image to the target image as best as possible without non-linearly warping the images themselves.

The affine registration is then followed by a deformable registration using DT-REFinD [24]: a full tensor version of the deformable demons algorithm [25]. The resulting deformation from DT-REFinD minimizes the log-Euclidean [26] sum-squared difference between the given tensor image I_i and the target T. We relaxed the smoothness parameters for the DT-REFinD registration (using $\sigma_{def} = 1$, $\sigma_{update} = 0.0$, max. step length $= 2.0$ voxels [24]) as they were found, by qualitative inspection, to give good results over a variety of ventricle sizes. This registration step introduces non-linear warping to the sample DTI scans $I_i (j \in [1, k])$ in order to better align them to the target image T. Note that DT-REFinD operates using the full diffusion tensor and not a measure, like FA, that is derived from the tensors. The use of the full diffusion tensor in image registration has been shown to significantly improve registration accuracy [27] over the use of individual tensor features.

2.1.2. Voxel-wise Model Fitting

Once all DTI scans have been anatomically aligned, we proceed with fitting a multivariate Gaussian distribution to the tensor data at each voxel. The choice of Gaussian distribution here is equivalent to the commonly-used t-test in standard VBA frameworks and comes with the same assumptions of normalcy on the image data.

It has been well established that tensors do not form a vector space [28] and so algebraic operations on ten-

sors are not guaranteed to give a tensor as a result. Due to this constraint, we model the tensor data in the log-Euclidean space, thereby ensuring that we respect the manifold of positive semi-definite 2^{nd} order tensors [26]. The log-Euclidean mean tensor image M is computed as

$$M(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} \operatorname{logm}\left(\phi_i \circ I_i(\mathbf{x})\right). \tag{1}$$

where ϕ_i is the concatenation of the affine and non-rigid deformations for the scan I_i, and matrix logarithm function $\operatorname{logm}(\cdot)$ is used to map the tensors to log-Euclidean space. The log-Euclidean tensor covariance image $S(\mathbf{x})$ is computed as well, though with the limited number of images commonly used in preterm DTI analysis studies, we rely on the shrinkage approach of Schäfer and Strimmer to get a more reliable estimate of the voxel-by-voxel covariance matrices than the maximum likelihood estimator can provide [29]. Note that the mean and covariance images are kept in the log-Euclidean space throughout this work to ensure that the tensors are manipulated properly.

As each multivariate Gaussian distribution is fit at each voxel, it becomes important to record whether the log-Euclidean tensors at a voxel are indeed well represented by a multivariate Gaussian distribution. Therefore, we employ a Henze-Zirkler multivariate normalcy test at each voxel to determine the probability that the given log-Euclidean tensors were obtained from a multivariate Gaussian distribution [30]. These probabilities are captured in an image, P, for each template and can then be taken into consideration when performing VBA. We will discuss how P can be used while performing VBA in Section 2.2.1.

2.1.3. Removing Target Image Bias

Once the Gaussian distributions are fit at each voxel, the resulting mean and covariance images are still in the reference space of the chosen target image T. Therefore, the choice of target image impacts the size and shape of the brain in the resulting mean and covariance images. In order to remove this bias to the target, it is necessary to deform both the mean image M and covariance image S so that the brain shape and size in both images more closely resembles the shape and size of the average brain of the population.

As given by Guimond et al., the target image bias correction can be computed from the computed deformations ϕ_j $(j \in [1, k])$:

$$\bar{\phi}^{-1}(x) = \frac{1}{k}\sum_{i=1}^{k} \phi_i^{-1}(x) \tag{2}$$

where ϕ_i^{-1} is the inverse of the deformation that warps image I_i to the target. By inverting those deformations, we obtain warps that deform the target towards each individual sample image I_i. Averaging these inverted deformations generates the correction warp $\bar{\phi}^{-1}$ that deforms the template towards an "average" brain shape and size.

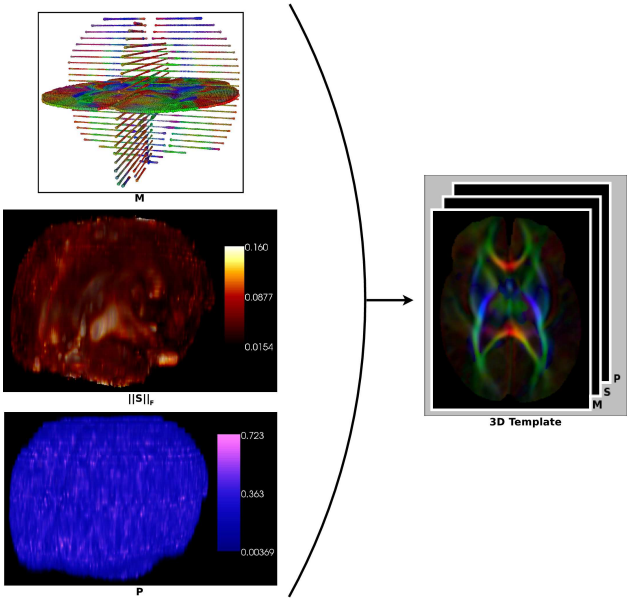


Figure 3: STEAM centers around a 3D statistical template modeling the DTI scan of a healthy preterm infant brain. This template consists of the three 3D images seen on the left: a mean image M, a covariance image S, and a normalcy p-value image P. To conserve space, we will refer to a 3D statistical template using the stacked visualization on the right.

The correction warp is applied to the mean image M to obtain a new target image T

$$T = \bar{\phi}^{-1} \circ M \tag{3}$$

and the atlas-creation process then repeats itself, using this new target, in order to remove any errors that may be caused by the registration algorithm limited ability to match to a target image far from the average brain shape and size. The template-creation procedure is repeated three times as this number of iterations has been empirically shown to be enough for the algorithm to converge [21, 27]. By doing this correction, the initial choice of target image T does not bias the final template [21]. Note that the computation of the covariance image S is only required for the final iteration when the bias correction steps have been applied and so it does not need to be corrected as the mean image is in (3).

Together, the mean image M, covariance image S, and p-value image P, form a single statistical template for the given normative population. An example of a statistical template is visualized in Figure 3. Note that the template provides a distribution of the diffusion tensors at each voxel as well as a measure of how trustworthy those distributions are, resulting in a full statistical description of what a normal preterm infant brain looks like on a DTI scan.

2.2. Abnormality Mapping

Our statistical template provides us with a model of how the preterm infant brain should look on an early DTI scan.

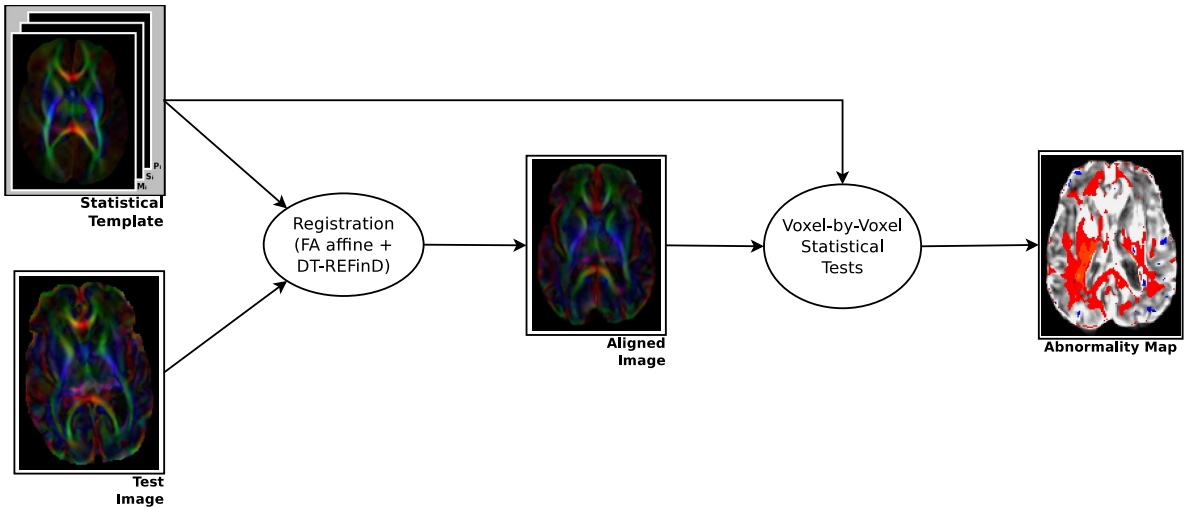


Figure 4: The proposed analysis pipeline for VBA. A new DTI scan is aligned to the STEAM statistical template, then values at each voxel are compared to the Gaussian distributions in the template using a χ^2-test. The voxels whose tensors are significantly different than their corresponding Gaussian distribution, after multiple comparison correction, are identified and visualized.

Such a model becomes valuable in its ability to assess the normality of a newly-obtained preterm DTI scan. This new scan can be aligned to that template and voxel-by-voxel statistical tests can be done to identify regions of the brain where the measured diffusion is significantly different from what we see in "normal" preterm infant population. This notion of aligning scans and detecting outliers at a voxel-by-voxel level is what underpins the concept of VBA.

The VBA framework consists of two main tasks: image alignment and voxel-by-voxel comparisons (Figure 4). Typically, VBA is done by aligning images from both the experimental and control groups, then performing a paired t-test (or T^2-test in the case of multivariate data) at each voxel to identify group differences [18]. In our case, we already have our statistical template as a normative model. As a result, we propose aligning individual scans to our template and performing single-sample statistical tests to identify voxels with significantly abnormal diffusion. By performing the VBA in this way, we are uniquely able to produce a subject-specific abnormality map that highlights where the diffusion abnormalities are. We are also able to shift the majority of the time-consuming registration steps to the template creation process, a process which can be run offline.

Given a new DTI scan I_{test}, our VBA process begins by aligning it to the mean image M in our statistical template. This image alignment step is performed in the same manner as in Section 2.1.1. First, we perform an affine registration with FA images to align image I_{test} to the mean image M of the template using FSL FLIRT [22, 23]. This registration step is followed by a full tensor deformable registration using DT-REFinD [24]. As in the atlas-creation step, these registration algorithms were selected and tuned to make use of all the information in the diffusion tensor during the image alignment process.

Once aligned to the template's image space, we can perform voxel-by-voxel statistical tests to identify outliers. For full tensor images we compute the Mahalanobis distance at each voxel $\mathbf{x}$ in log-Euclidean space:

$$\begin{aligned} d(\mathbf{x}) =& \operatorname{vec}\left(\operatorname{logm}\left(I_{test}(\mathbf{x})\right) - M(\mathbf{x})\right)^T S^{-1}(\mathbf{x}) \\ & \operatorname{vec}\left(\operatorname{logm}\left(I_{test}(\mathbf{x})\right) - M(\mathbf{x})\right) \end{aligned} \tag{4}$$

where $\operatorname{vec}(\cdot)$ is the matrix vectorization function. Note that M is already in the log-Euclidean space (as mentioned in Section 2.1.2) and does not require matrix logarithm transformation. Conceptually, the Mahalanobis distance can be seen as a multi-dimensional version of a z-score: a distance between a sample and a distribution's mean, divided by the variance of the distribution (in this case captured by the inverse of S).

To identify outliers, we first assume the tensors in the new scan I_{test} are from the same distributions represented by the template. Under this assumption, the Mahalanobis distances from (4) follow a Chi-squared distribution χ^2_p with $p = 6$ degrees of freedom (as there are six unique elements in each tensor) [31]. If these Mahalanobis distances fall outside the $(1 - \alpha)$-quantile of the distribution χ^2_6, then those distances are outliers of the distribution. For these outliers, we can reject (with confidence $1-\alpha$) the assumption that their log-Euclidean tensor values come from the distributions described in the preterm infant statistical template [31].

While this voxel-wise test above provides us with a way of identifying statistical outliers, it comes with two major caveats. First, it is possible that tensors at certain voxels in our statistical template are unlikely to follow a Gaussian distribution. For those voxels, the results of our statistical test may be unreliable, which raises questions about how those results should be reported. Second, we are performing multiple statistical tests and we have to

make our significance threshold α more stringent to account for these multiple comparisons. VBA results, like the ones STEAM produces, are known to be susceptible to these two design decisions [19], so we examine these two points in greater detail and justify the decisions we make with respect to both of them.

2.2.1. Addressing Non-Gaussian Data

When generating a statistical DTI template, we assume that at each voxel, the distribution of diffusion measurements can be modeled as Gaussian. In previous DTI VBA studies, this has been shown not to be the case [19]. Consequently, when creating the statistical templates, we test for normalcy at each voxel and provided the probability of the data being normally distributed through the image P (see Section 2.1.2). We can then see if the diffusion data at voxel $\mathbf{x}$ is normally distributed by checking whether $P(\mathbf{x}) > \alpha$, where α is a multiple comparison corrected significance threshold.

We should be wary of statistics computed at voxels where $P(\mathbf{x}) < \alpha$ as the normal distributions fit at these voxel locations do not accurately represent the underlying distribution of possible diffusion measurements. Depending on how conservative we want to be, we could discard the statistics computed at these voxels, weight the contribution of the statistics at these voxels by how likely their data is to be normally distributed, or simply use the statistics despite the lack of normalcy in the data at those locations.

2.2.2. Multiple Comparison Correction

When performing multiple statistical tests, as we are doing at each voxel in an image, there is always the potential that a certain fraction of voxel values will be marked as outliers by chance. For example, if we select $\alpha = 0.05$ for our Chi-squared Mahalanobis test above, then on average, 5% of the voxels in I_{test} will be marked as outliers when they shouldn't be. To address this phenomenon and obtain a more reliable threshold for detecting outliers, we can use various techniques to correct for multiple comparisons [32].

Generally, there is no consensus on how to perform multiple comparison correction [33]. Still, four techniques are commonly seen in the VBA community: (a) Bonferroni correction [34], (b) Gaussian Random Field Theory [35], (c) False Discovery Rate [36], and (d) Permutation testing [37]. The most conservative correction approach is Bonferroni correction which assumes that the Mahalanobis distances computed in (4) are independent from one voxel to another. When performing VBA, this assumption is overly strict as the values at image voxels are usually similar to their neighbors [33]. As a result, Bonferroni correction is rarely used as it can miss certain outliers by ignoring this neighborhood correlation.

Gaussian Random Field (GRF) Theory shares some similarities with Bonferroni correction as it scales the significance threshold α by the number of independent statistical comparisons. However, GRF theory assumes that the neighboring image values follow a Gaussian-like profile (i.e., smooth), thereby allowing us to more accurately estimate the number of independent statistical comparisons [35]. Unfortunately, due to sharp transitions between narrow fiber tracts, the typical DTI scan is usually not smooth enough for GRF theory to provide a notably less conservative correction scheme than Bonferroni [33]. Smoothing the DT images with a Gaussian kernel would alleviate this problem, but the amount of smoothing required to get a less conservative significance threshold is relatively high (greater than 6 mm full width at half-max Gaussian smoothing given our cohort size) [33]. Such extensive smoothing would also blur out narrow fiber tracts, making it more difficult to localize abnormalities with our statistical template[2]. For these reasons, GRF theory is ill-suited to correct for multiple comparisons in DTI studies.

Permutation testing is popular for multiple comparison correction when performing group studies as it makes few assumptions about the relationships between input statistics [37]. Instead, permutation tests computes statistical group differences for various random permutations of labelings for the group members. The significance threshold is then chosen from the distribution of these group differences so that only a fraction (α) of the random group label permutations were above this threshold. Permutation testing has become popular as it is based on distance metrics and not any specific distribution. This condition has allowed for some preprocessing of statistical maps as is done in Threshold-Free Cluster Enhancement [38]. Unfortunately, applying permutation testing in STEAM would require maintaining two groups: one group containing only I_{test}, and the other containing all the DTI scans used to create the statistical template. Maintaining these two groups may be cumbersome if the number of scans used to create the template is large. Also, retaining the DTI scans used for template creation leads to redundancy between those scans and the template itself. For these reasons, permutation testing is not ideal for multiple comparison correction in STEAM.

As a result of the limitations of these other methods, we use False Discovery Rate (FDR) to perform multiple comparison correction. Like permutation testing, FDR does not make any assumptions on the relationships between input statistics or on the smoothness of the images, yet it is less conservative than Bonferroni correction or GRF theory [36]. FDR also has the benefit of being applicable to the single-sample statistical tests that are performed in STEAM.

By taking into consideration multiple comparison correction, as well as non-Gaussian distributed data, the STEAM analysis engine address two major caveats surrounding voxel-based analysis, resulting in a technique

[2]Note that this GRF-based smoothing for multiple-comparison correction is different from the scale-space smoothing discussed later in section 2.3.2.

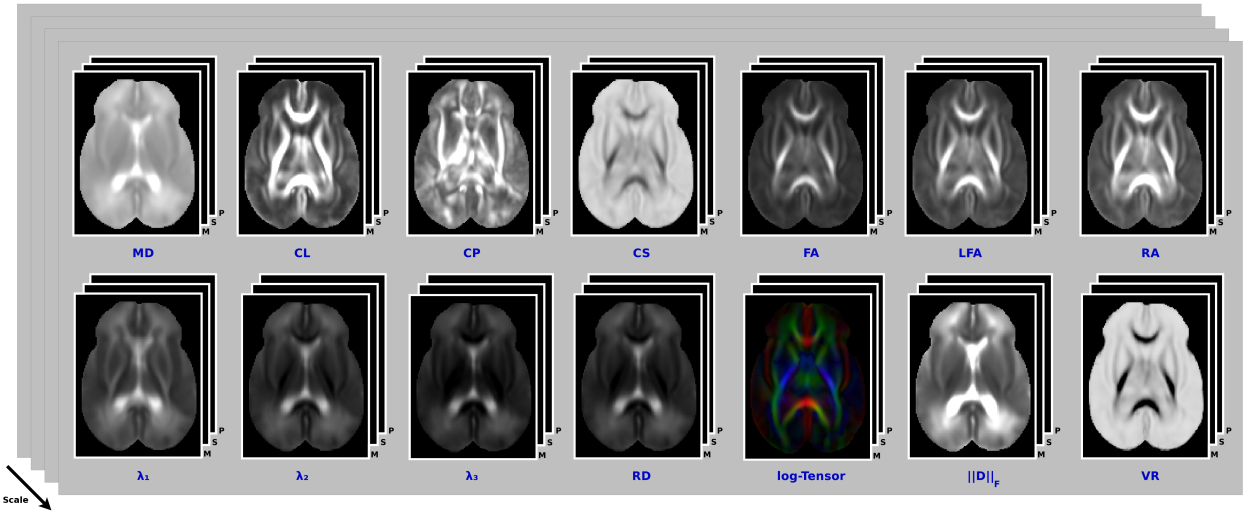


Figure 5: A visualization of the various diffusion tensor measures for which we generated preterm infant statistical templates. Among these measures are mean diffusivity (MD), the tensor shape measures (c_l, c_p, c_s) defined in [39], fractional anisotropy (FA), log-Euclidean FA (LFA), relative anisotropy (RA), each individual eigenvalue $(\lambda_1, \lambda_2, \lambda_3)$ of the diffusion tensor, radial diffusivity (RD), tensor norm ($\|\mathbf{D}\|_F$), and volume ratio (VR). Note that templates for all of these measures are computed at multiple spatial scales.

that can be used to identify abnormalities at the voxel level in a single DTI scan.

2.3. Completing the Template Collection

While we have presented STEAM with respect to a full tensor statistical template, it has been common to examine preterm infant DTI scans using simpler features of interest from the diffusion tensors, in particular FA and MD [4, 17]. It is also common in VBA to smooth the images being analyzed in order to examine their content at different spatial scales [20]. STEAM also provides this functionality through an expansion of its statistical template collection, a collection that we describe below.

2.3.1. Templates for Diffusion Features

While we described STEAM's template-creation technique with respect to the full diffusion tensor image, the same technique is used to generate statistical templates for various tensor features including fractional anisotropy, mean diffusivity, and all others shown in Figure 5. As in the full tensor case, the DTI scans are aligned using a combination of FSL FLIRT and DT-REFinD to obtain the same deformations ϕ_i $(i \in [1, k])$ as in the full tensor analysis. Once the scans are aligned, the scalar diffusion feature (e.g., FA, MD) are computed from the aligned tensor images $\phi_i \circ I_i (i \in [1, k])$. The model fitting step then simplifies to computing scalar mean and variances at each voxel, while using the Lilliefors test for normalcy at each voxel [40]. The template bias correction is then computed from the deformations ϕ_i $(i \in [1, k])$ in the same way as in the full tensor case.

We refer to these individual scalar templates by their mean image M_f, variance image S_f, and normalcy p-value image P_f, where f is the tensor feature being modeled (e.g., $f = FA$, or $f = MD$, etc...). STEAM can compute statistical templates for 14 different diffusion features (including the full log-Euclidean tensor), obtaining the mean, (co-) variance, and p-value images shown in Figure 5.

When a new DTI scan needs to be analyzed on a specific tensor feature of interest, STEAM performs the analysis in a similar fashion to the full tensor case. The DTI scan is aligned to the mean image of the full tensor template using FSL FLIRT and DT-REFinD. Once aligned, we compute the scalar diffusion feature (e.g., FA, MD) from the aligned test image and compare it to the statistical template for that feature. Note that since the same image registration algorithms were used regardless of the choice of template, each template should be anatomically aligned. The statistical test for the scalar images simplifies to the z-test and the same multiple comparison correction is applied. In this way, STEAM produces VBA results for individual tensor features from a single DTI scan.

2.3.2. Templates for Different Image Scales

It has also been general practice in the VBA community to smooth images before performing VBA. The rationale behind performing this smoothing is to (a) reduce the number of false outliers identified due to misregistration, (b) to make the image data at each voxel more likely to be normally distributed, and (c) to introduce a spatial scale to the analysis [20]. Unfortunately, it has been well noted that different levels of smoothing can result in very different conclusions being drawn from a VBA analysis [19, 20]. It has been suggested that when VBA is performed on DTI scans, researchers should provide results for a range of smoothing scales [20].

In order to accommodate different spatial scales, we expand STEAM to produce templates at various smoothing scales. Specifically, Gaussian smoothing is applied to the aligned images $\phi_i \circ I_i (i \in [1, k])$ prior to the compu-

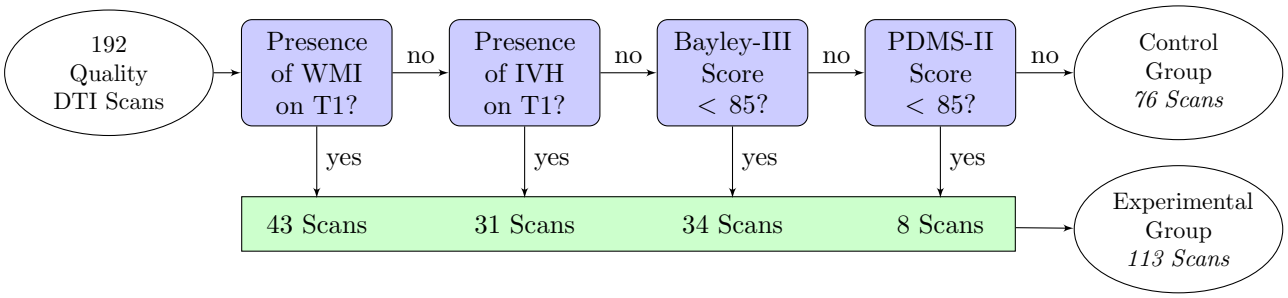


Figure 6: Exclusion criteria for the group of preterm infant DTI scans used to create the normative statistical templates. The number of scans excluded by each criteria are listed below the corresponding criteria (WMI = white matter injury, IVH = intraventricular hemmorhage, Bayley-III = Bayley Scales of Infant and Toddler Development, PDMS-II = Peabody Developmental Motor Scales). Excluded scans of sufficient quality were used to validate the VBA aspect of STEAM. A more detailed description of that VBA test set is given with the corresponding experiments in Section 4. Note that scans were included in the template only if the infant's measures of neurodevelopment are within 1 standard deviation of the normal mean (> 85). Further details on these exclusion criteria are given in Section 3.2.

tation of the mean and (co-) variance images described earlier. STEAM can create templates smoothed with a range of Gaussian filters whose full width at half maximum (FWHM) values have been seen in previous DTI VBA literature [20]. When a new DTI scan I_{test} is obtained for analysis, STEAM can smooth I_{test} with the equivalent Gaussian function, then perform VBA at that spatial scale in the same manner as described earlier. In this fashion, STEAM can perform VBA and report results across multiple spatial scales.

By introducing n spatial scales, and 14 different diffusion features, a full STEAM analysis engine includes $14n$ statistical templates, each of which include a mean, (co-) variance, and p-value images as shown in Figure 5. These additional templates give STEAM the added flexibility to isolate specific diffusion abnormalities that manifest at different spatial scales.

3. Materials: Cohort and Imaging

To validate STEAM, we make use of an existing cohort of 195 premature newborns born between 24 to 32 weeks gestational age (GA) at the Childrens & Womens Health Centre of British Columbia, 177 are described in Chau et al. [10] and an additional 18 infants recruited since that work was published. Cohort exclusion criteria included 1) congenital malformation or syndrome; 2) antenatal infection; or 3) large parenchymal hemorrhagic infarction (> 2 cm) detected using head ultrasound scanning. This prospective study was approved by the University of British Columbia Clinical Research Ethics Board. The newborns enrolled in this cohort were evaluated with MRI scans in the neonatal period (outlined below) and had neurodevelopmental assessments at a corrected age of 18 months with the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III) [41] and the Peabody Developmental Motor Scales, Second Edition (PDMS-II) [42]. The 3 composite scores (cognitive, language and motor scores) of the Bayley-III have a mean of 100 and standard deviation of 15. The PDMS-II provides a more sensitive assessment of motor function yielding gross, fine and total motor scores with a mean of 100 and standard deviation of 15.

Table 1: Deomgraphics of the experimental and control groups for the preterm infant cohort used in this paper to validate STEAM. Experimental and control groups are defined in Section 3.2. P-values are shown between the two groups. Note that there are no significant differences in birth age, scan age, sex, or brain volume between the experimental and control groups.

Cohort Demographic	Control Group	Experimental Group	P-Value
Number of Subjects	55	79	–
Number of Scans	76	113	–
Sex (M/F)	30 / 25	38 / 41	–
Avg. GA at Birth	28.3251	27.9127	0.2882
Avg. PMA at Scan	35.1952	35.2002	0.9390
Brain Volume (cm^3)	286.25	269.83	0.3129

3.1. Magnetic Resonance Imaging

Of the 195 very preterm neonates, 170 were scanned within the first weeks of life once they were clinically stable. One hundred and fifty-two (152) of these 170 infants were scanned again at term-equivalent age, with 0.85 to 15.28 (7.98±3.32) weeks between scans. The resulting 322 diffusion MRI scans cover the age range of 27.86 to 46.42 (36.38±4.89) weeks post-menstrual age (PMA).

Our MRI studies were carried out on a Siemens (Erlangen, Germany) 1.5T Avanto using VB 13A software and included the following sequences: 3D coronal volumetric T_1-weighted images (repetition time [TR], 36 ms; echo time [TE], 9.2 ms; field of view [FOV], 200 mm; slice thickness, 1 mm, no gap) and a multi-slice 2D axial EPI diffusion MR acquisition (TR 4900 ms; TE 104 ms; FOV 160 mm; slice thickness, 3 mm; no gap) with 3 averages of 12 non-colinear gradient directions, resulting in an in-plane resolution of 0.625 mm. The DTI acquisition was repeated twice, once with a diffusion weighting (b-value) of 600 s/mm^2 and once with a diffusion weighting of 700 s/mm^2. The two DTI acquisitions were then combined to create a single diffusion tensor image. The combined

Table 2: Demographics for the preterm infant cohort divided by age window for which we created a statistical template. Numbers are provided for the control group first, followed by the experimental group, with the p-value (1-way ANOVA) of the group differences shown in brackets. Note that there are significant age differences (highlighted in bold) between experimental and control groups for the second and fourth-youngest statistical template age windows.

Demographics	Template Post-Menstrual Age (PMA) Groups			
control/exp (p-val)	28-31 weeks	32-36 weeks	37-40 weeks	41-45 weeks
Number of Subjects	24 / 28	22 / 43	16 / 23	14 / 17
Males - Females	12-12 / 16-12	12-10 / 21-22	9-7 / 13-10	8-6 / 7-10
Number of Scans	24 / 28	22 / 45	16 / 23	14 / 17
Avg. GA at Birth	27.39 / 27.69 (0.55)	**29.31 / 27.83 (0.03)**	28.05 / 26.97 (0.13)	**28.40 / 29.76 (0.04)**
Avg. PMA at Scan	29.92 / 30.11 (0.32)	**32.84 / 33.86 (0.003)**	38.92 / 39.13 (0.43)	42.90 / 43.02 (0.68)
Brain Volume (cm^3)	169.54 / 177.57 (0.31)	220.92 / 240.05 (0.06)	**388.08 / 336.87 (0.03)**	445.23 / 430.92 (0.55)

diffusion weighted image set was preprocessed (i.e., eddy current corrected and skull stripped) using the FSL Diffusion Toolbox (FDT) pipeline[3] and tensors were then fit using RESTORE [43]: a weighted least-squares tensor fitting algorithm implemented in the Camino toolkit[4].

An experienced neuroradiologist (K.J.P.) reviewed the resulting MR images for presence of white matter injury (WMI), intraventricular hemorrhages (IVH), ventriculomegaly (VM), and poor image quality. The full neuroradiological review was performed on the T1 images using the following protocols. The presence of WMI was identified using a system found to be predictive of adverse neurodevelopmental outcome at 12 to 18 months of age [12]. We noted IVH using the grading of Papile et al. [13] and VM using the grading system of Cardoza et al. [44].

We employ a high standard for image quality by visibly checking for evidence of motion corruption and various image artifacts discussed in Tournier et al. [45] and Gallichan et al. [46]. To avoid corrupting our DTI analysis of the whole brain, we included a scan in our study only if the entire scan is free of all degradations. Of the 322 scans we collected, 192 of them met that stringent criteria and were included in this study. Of the 130 excluded scans, 42 we excluded due to excessive motion, 76 were removed due to vibrational artifacts similar to those described by Gallichan et al. [46], and the remaining 12 were removed due to the presence of other artifacts described by Tournier et al. [45].

3.2. Defining Experimental and Control Groups

To generate a set of statistical templates that capture the range of "normal" brain development, we first must define what criteria we use to decide whether an infant's DTI scan and neurodevelopmental outcome are normal, then determine which scans in our cohort fit that criteria. Those scans that fit these criteria will comprise our control group from which our statistical templates will be built.

The full control group selection criteria is shown in Figure 6. Infants were included in the control group if their scores on all six composite measures of neurodevelopment (Bayley-III and PDMS-II) where at least within 1 standard deviation of the normal mean (> 85). Of those infants that met this criteria, we further excluded any infants that showed they acquired brain injury, either white matter lesions or intraventricular hemorrhage, on MRI (as identified by the protocols described in Section 3.1). In our cohort, we obtained 76 scans from 55 infants that satisfied these criteria. The distribution by PMA of the DTI scans in the control group is given in Figure 7. Full demographics of these infants are given in Table 1.

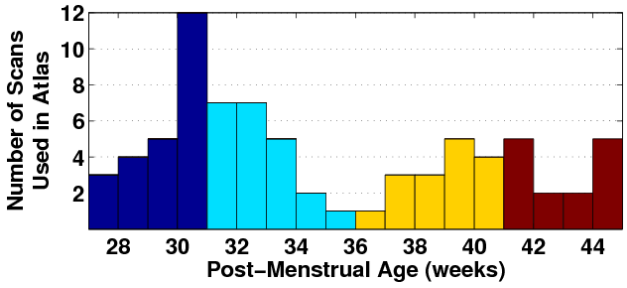


Figure 7: The distribution of DTI scans used to generate each statistical preterm infant template. Bars are color-coded based on the age windows used for each template (see Section 3.2).

Given a control group consisting of 55 infants and 76 scans, we have the luxury to sub-divide our control group according to PMA at time of scan. By performing this subdivision, we are able to reduce the amount of variance in our statistical templates that is caused by PMA, resulting in a greater ability to identify statistical abnormalities. We chose to sub-divide our control group into roughly 4-5 week time windows as highlighted by the different colors in Figure 7. This sub-division allows us to maintain a similar number of scans (i.e., similar statistical power) in each time window while also optimizing the trade-off between the number of scans per time window and the age-related image variance within each window. Full demographics for each sub-group are given in Table 2.

As can be seen in Table 2, the experimental and control sub-groups are generally similar to each other, with only the four bolded comparisons showing significant differences. As far as demographic differences across time windows, we saw the control group for the earliest age win-

[3]http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FDT
[4]http://cmic.cs.ucl.ac.uk/camino/

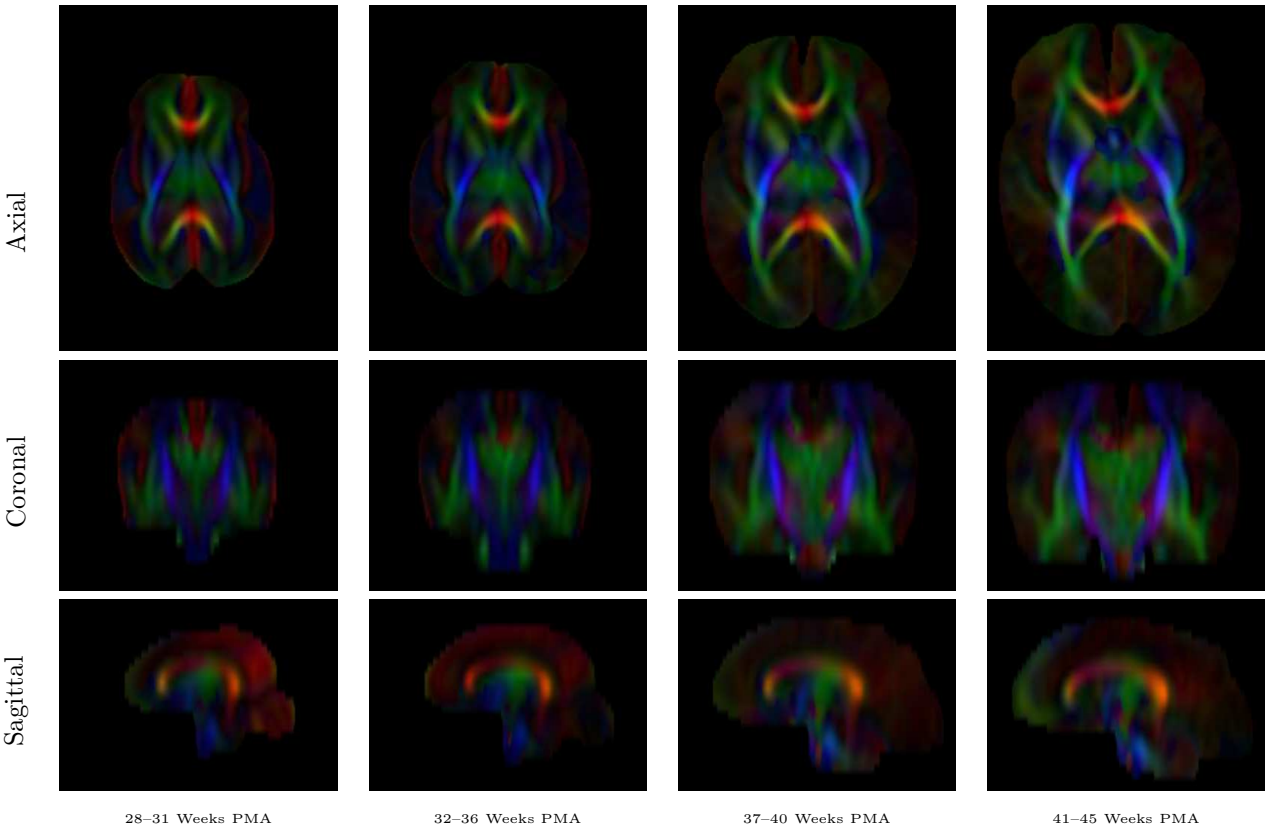


Figure 8: Axial, coronal, and saggital slices of the mean color FA maps for our four preterm infant DTI templates. Note that all figures are drawn with the same scaling so any change in size is due to brain development. Also note the template quality makes it easy to distinguish major fiber tracts.

dow had a significantly lower GA at birth than the second (32-36 weeks PMA, $p = 0.0016$) and fourth (41-45 weeks PMA, $p = 0.0459$) age windows, but no other significant differences were found between GA at birth for any pair of templates (smallest $p = 0.1112$). No other significant differences in demographics were found for the control groups across time windows.

4. Experimental Results

Our proposed STEAM analysis engine contains two steps: the creation of a normative statistical template collection, and a VBA pipeline to identify areas of abnormality in a single DTI scan. The following sections present results on both these steps, specifically how these results compare with known anatomical findings as well as how STEAM-generated results compare to existing single-scan evaluations of preterm infants (e.g., [12]). We examine the hypotheses that STEAM:

- Generates statistical templates that show the growth, development, and inter-subject variability we would expect to see in the normal preterm infant brain over the examined time period.
- Generates subject-specific abnormality maps that can be reliably interpreted and are consistent with a subject's structural MR evaluation (e.g., [12]).
- Identifies brain abnormalities that relate to neurodevelopmental outcome at a corrected age of 18 months.
- Identifies brain abnormalities that are separate to, yet complement those, that can be identified on structural MRI.

The results that follow support these hypotheses but should not be considered a full clinical evaluation of STEAM. Our goal is simply to show a broad proof of concept.

4.1. Validation of Normative Statistical Templates

Our STEAM statistical template creation procedure generated the four preterm DTI templates whose mean images, M, are displayed in Figure 8. Qualitatively, the templates displayed the expected anatomical organization of major fiber tracts with the Genu, Splenium, Optic Radiations, and Corticospinal Tracts clearly identifiable from each mean image. Further, we see greater lateral growth of these tracts as post-menstrual age increases, which agrees with earlier DTI findings [11, 47, 48]. This lateral growth is also consistent with histological findings that have identified a reduction of the subplate zone and an expansion of the white matter over this period [49].

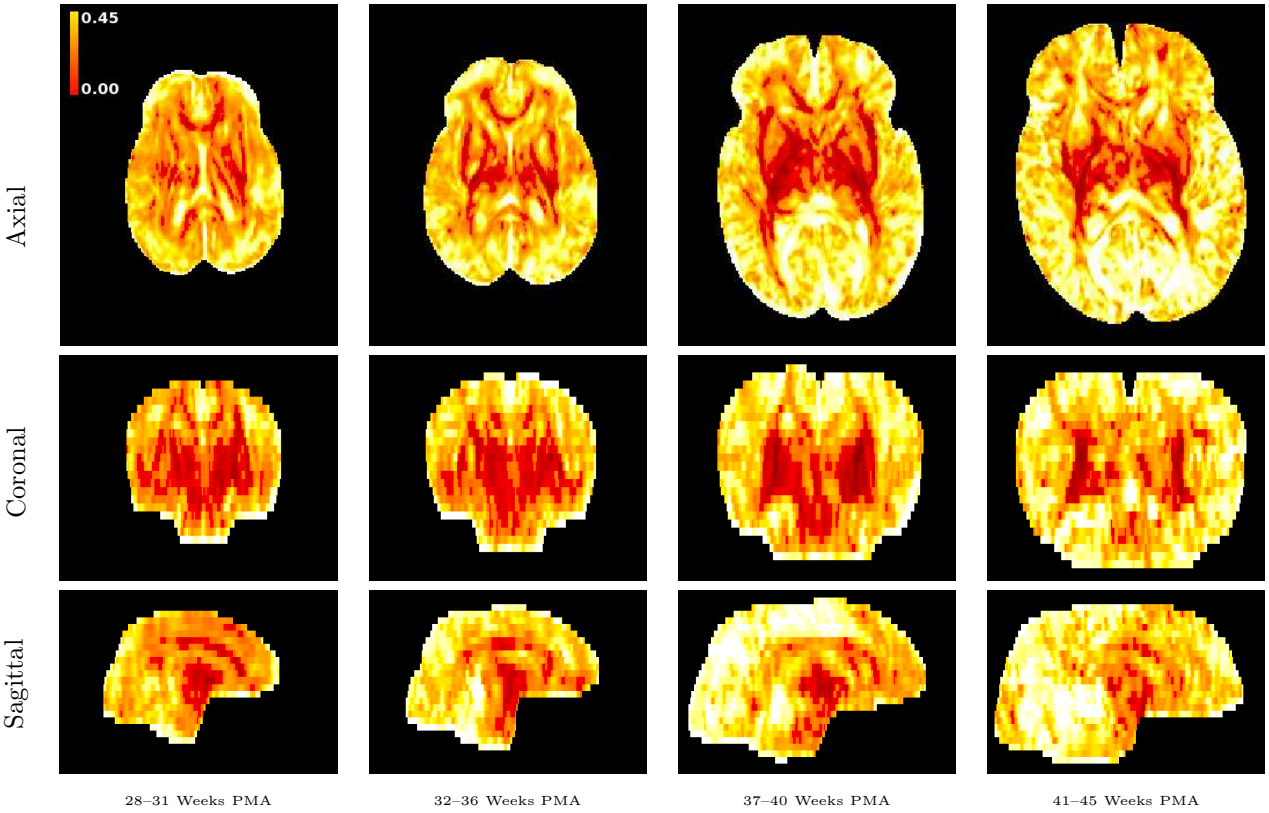


Figure 9: Axial, coronal, and saggital slices of the FA coefficient of variation for our four preterm infant DTI templates. Note that as the brain develops the inter-subject variability increases. Also, the variability is greater in the posterior part of the brain, suggesting greater development in that part of the brain over this 28-45 week PMA time period.

We also saw increased contrast between the major fiber tracts and the rest of the brain as gestational age increases. This result was expected as the maturation of the major fiber tracts during this period increases the FA within those tracts [3]. This increase in FA contrast is also aided by a decrease of FA in cortical and sub-cortical regions, a decrease that is consistent with a decrease in the radial organization of neurons [48]. This FA decrease has also been reported in an ROI-based study [16] and is likely due to dendritic arborization of neurons in the subplate zone [50].

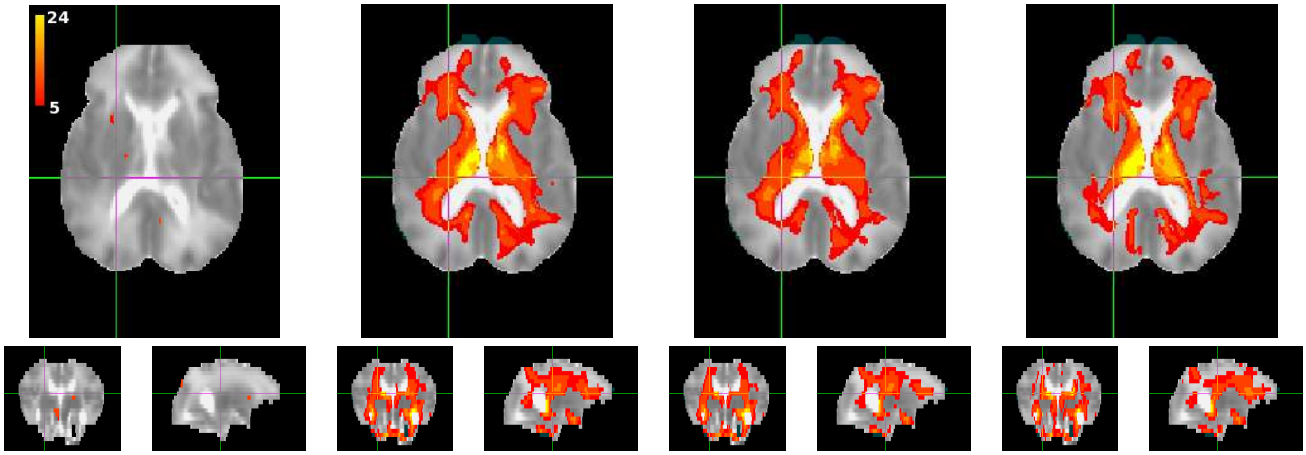
For each of our templates, we also examined the covariance at each voxel to determine where we see the greatest variability within the normal preterm infant brain. In particular, we examined the coefficient of variation images

$$c(\mathbf{x}) = \frac{\sigma(\mathbf{x})}{\mu(\mathbf{x})} \tag{5}$$

where the standard deviation, σ at each voxel $\mathbf{x}$ is displayed as a fraction of the mean value μ. Figure 9 shows a representative example: the coefficient of variation images for the FA templates. Qualitatively, we saw greater variation in FA in the posterior portion of the brain, which agrees with greater development of the occipital lobes during this time period [51, 52]. This greater development in the occipital lobes has also been shown in earlier DTI studies [14, 11, 48].

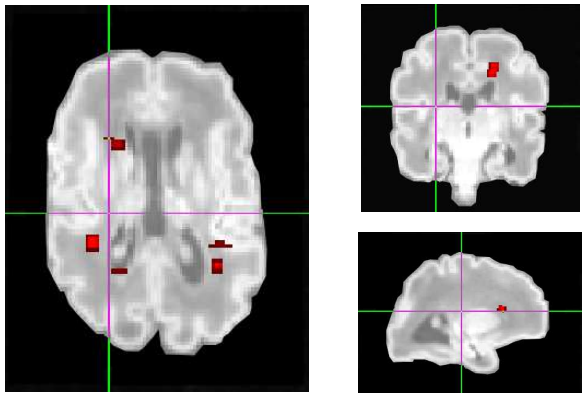
We further observed an increase in the coefficient of variation over time, suggesting that the brain structure of preterm infants becomes more diverse over time. This result agrees with the work of Brown et al. which showed increased variability in connectome measures over the same time period [14]. This result also agrees with the development of sulci over this age period [53] and that sulci have features that are unique to each individual [54]. We note that this trend appears to be independent of age at birth. Only the first template control group showed a significantly lower distribution of birth age compared to other templates (see Section 3.2), yet the coefficient of variation increases across the four templates.

To avoid excessive clutter in this paper, we have posted all our templates on the STEAM project website (`www.sfu.ca/~bgb2/steam`) where they can be viewed online (via a web-based 3D image viewer) and downloaded. The observations identified here were generally consistent across diffusion features, specifically the lateral growth of white matter tracts, the decreased radial organization for neurons in cortical regions, and greater variability in the posterior region of the brain. The consistency of these results with existing DTI and histological studies suggest that our STEAM templates capture the expected growth and variability previously identified in the normal preterm infant brain over the examined time period.

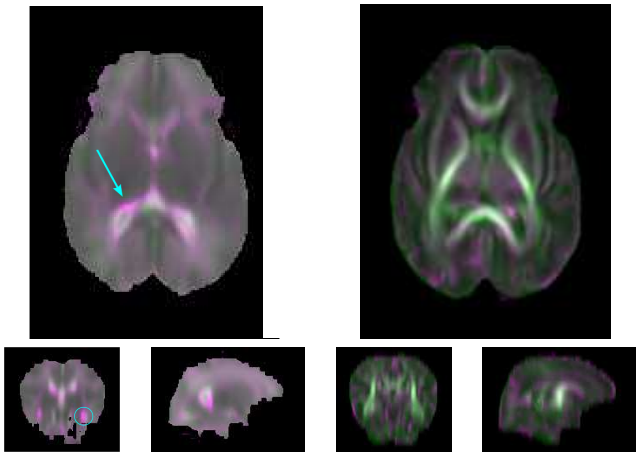


FA MD RD λ_1

(a) Widespread STEAM-detected abnormalities identified in deep gray and white matter



(b) Infant T1 rigidly-registered to template
(white matter lesions shown in red)



MD FA

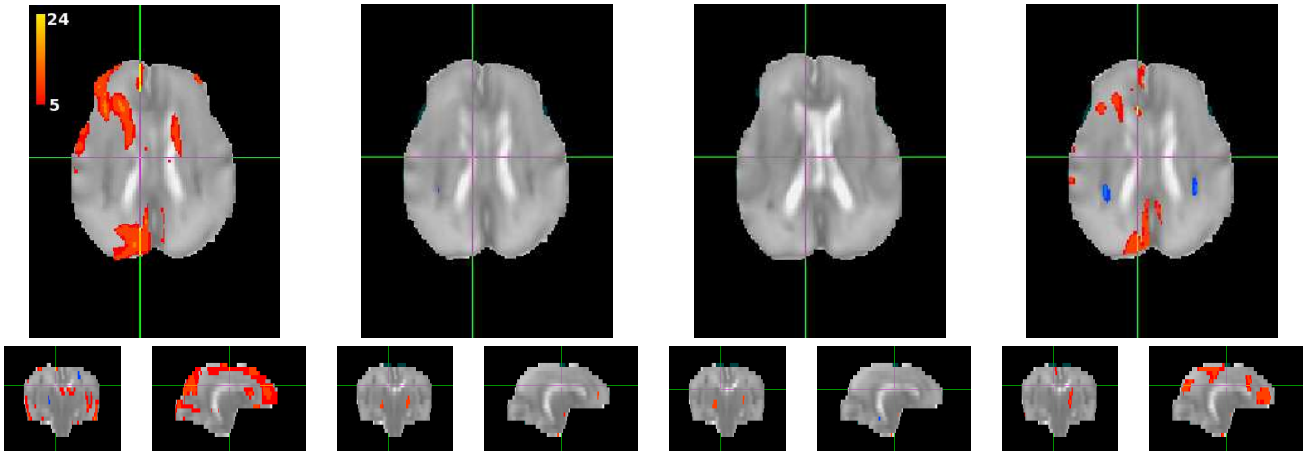
(c) Blended Images: Subject and Template Mean
(subject in purple, template in green)

MRI-based Scores (32 wks. PMA)			Bayley-III Test Scores (18 mo. corrected)	
WMI [12]	IVH [13]	VM [44]	Cognitive	Motor
2 - Moderate	0 - Absent	1 - Mild	55	46

Figure 10: A case study of how STEAM can be used to identify DTI abnormalities and how those detected abnormalities compare to structural MRI abnormalities and outcome. The results from STEAM's voxel-based analysis for FA, MD, RD, and λ_1 are shown in (a). These results show abnormally high MD, RD, and λ_1 over a large region encompassing deep gray and white matter. These results are consistant both with the presence of white matter lesions on the infant's T1 scan shown in (b) and the infant's significantly reduced neurodevelopmal test scores at 18 months corrected age (shown in the table above). The registration accuracy between the infant's DTI scan at the STEAM statistical template is shown in (c). We do see some misregistration around the posterior portion of the ventricles on the MD blended image, which is a result of ventriculomegaly. However, this misregistration is small in comparison to the STEAM-detected DTI abnormalities. The combination of all these results suggest that STEAM is identifying a true structural abnormality in this infant.

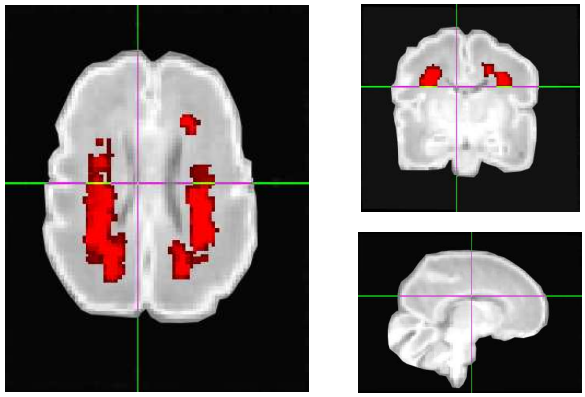
4.2. Subject-Specific Abnormality Maps: Proof of Concept

We further generated abnormality maps for the 113 DTI scans in our experimental group by comparing them to the STEAM templates of the appropriate postmenstrual age. These 113 scans were analyzed using our VBA pipelines for four common diffusion measures: fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), and axial diffusivity (λ_1). We focused on these four diffusion measures as they have been the focus of many previous preterm DTI studies [4, 10, 47, 55], allowing us to examine our results in the context of that earlier work. To account for multiple imaging scales, we generated abnormality maps for all scales from $0-8mm$ (at $1mm$ intervals) and marked a voxel as abnormal if its value was significantly different from the STEAM template on a majority of image scales.

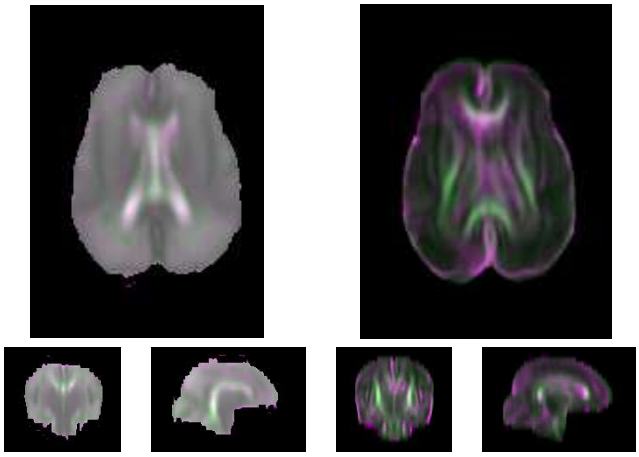


FA MD RD λ_1

(a) Widespread STEAM-detected abnormalities identified in sub-cortical gray and white matter



(b) Infant T1 rigidly-registered to template
(white matter lesions shown in red)



MD FA

(c) Blended Images: Subject and Template Mean
(subject in purple, template in green)

MRI-based Scores (29 wks. PMA)			Bayley-III Test Scores (18 mo. corrected)	
WMI [12]	IVH [13]	VM [44]	Cognitive	Motor
3 - Severe	2 - Moderate	0 - Absent	110	67

Figure 11: A second case study of how STEAM can be used to identify DTI abnormalities and how those detected abnormalities compare to structural MRI abnormalities and outcome. The results from STEAM's voxel-based analysis for FA, MD, RD, and λ_1 are shown in (a). These results show abnormally high FA and λ_1 in areas of cortical gray matter and superficial white matter on the left side of the brain. These results are consistant with the presence of nearby white matter lesions on the infant's T1 scan shown in (b) as well as the infant's significantly reduced motor test scores at 18 months corrected age (shown in the table above). The registration accuracy between the infant's DTI scan at the STEAM statistical template is shown in (c). While some of the abnormalities are around the cortex, there is no discernable registration error in these regions. The combination of all these results suggest that STEAM is identifying a true structural abnormality in this infant.

As a proof of concept, we present two notable cases to highlight how STEAM-generated abnormality maps can be interpreted and how they can be used to inform on the presence of anatomical abnormalities. The first case is shown in Figure 10. The STEAM-generated abnormality maps for the four diffusion measures are presented in Figure 10(a) and show the presence of widespread abnormalities in MD, RD, and axial diffusivity. The infant's corresponding T1 scan was rigidly aligned to the template space and is displayed in Figure 10(b). This infant's T1 scan showed the presence of multiple small hyperintense lesions, all of which were manually identified by a trained expert and highlighted in red. Finally, Figure 10(c) show the FA and MD images from the subject and the template mean in the complementary colors of purple and green respectively. The blending of these images results in shades

of gray where the images match and highlights differences between the aligned scans in one of the two image colors. These blended images allow us to examine for the presence of image registration errors that may impact the abnormality maps STEAM generates (e.g., neighboring green and purple structures would imply misalignment. A more thorough introduction to image blending can be found in [56]). The infant's T1 MRI scores, as well as their neurodevelopmental test scores at 18 months are shown in the table at the bottom of Figure 10.

As an initial step in interpreting these STEAM-generated abnormality maps, we looked for the possibility of image registration errors. By examining Figure 10(c), we do see purple regions around the posterior portion of the ventricles, suggesting that, after aligning to the template, the ventricles in the infant's DTI scan remained slightly larger than those in the template's mean image. This result is not surprising as this infant showed presence of mild ventriculomegaly (VM) that, apparently, our image registration techniques could not fully account for. Even so, the misregistration is limited to a much smaller region than the abnormalities present on the STEAM-generated abnormality maps. These detected abnormalities extend well beyond the periventricular region and into regions occupied by major white matter fiber tracts (e.g., splenium), and those major tracts are well-aligned as evidenced in the blended FA image in Figure 10(c). While misregistration would account for some of the detected abnormalities near the posterior portion of the ventricles, it alone cannot explain the widespread MD, RD, and axial diffusivity abnormalities identified by STEAM.

Instead, the infant's T1 scan provides some additional clues that corroborate the abnormalities identified by STEAM. Specifically, multiple hyperintense white matter lesions appear scattered in the same areas as the abnormalities identified by STEAM. It is believed that these lesions are an indication of a more widespread diffuse white matter injury in the neighboring tissue [2]. The abnormalities identified by STEAM match that description, suggesting that we are capturing a greater extent of the diffuse white matter injury than the lesions display on the T1 scan. This interpretation also agrees with the low neurodevelopmental test scores obtained at 18 months as one would expect that such widespread brain abnormalities would have a profound impact on later neurodevelopmental outcome.

The STEAM-generated results for a second infant are shown in Figure 11(a) along with their T1 scan in Figure 11(b), the blended FA and MD images in Figure 11(c), and the infant's T1 evaluation and neurodevelopmental scores in the corresponding table. The STEAM results suggest increased FA and axial diffusivity in the left occipital lobe, the left frontal-temporal lobe, and in various cortical regions. The blended MD image in Figure 11(c) shows no notable image registration error, while the blended FA image shows clear FA differences but none appear to be due to anatomical misalignment (which would appear as neighboring purple and green structures). The lack of registration error suggests that the abnormalities identified by STEAM are indicative of anatomical abnormalities.

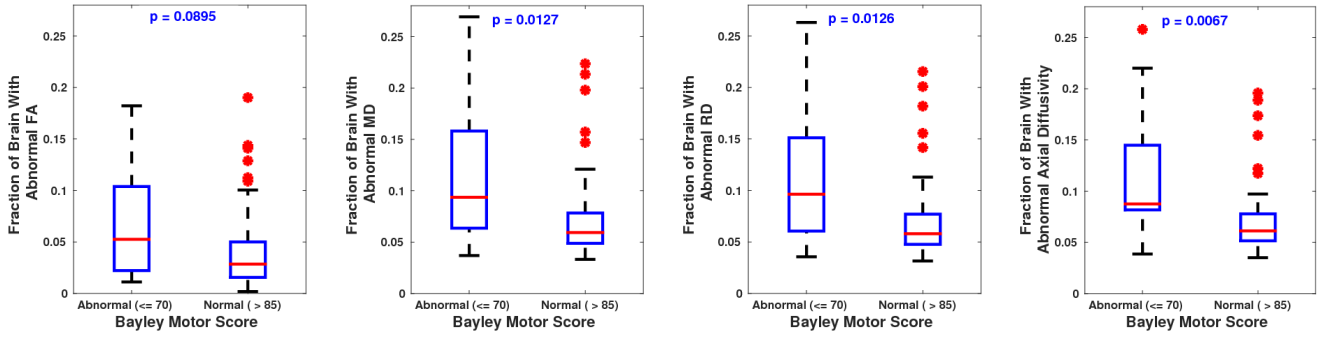
Comparing the STEAM abnormality maps to the infant's T1 scan, we see that the larger regions of STEAM-detected abnormalities are in the proximity of a large white matter lesion in the left hemisphere. This proximity suggests a relationship between the lesion and the nearby FA abnormalities that is consistent with previous findings [2]. Further, the increased FA in these parts of the subplate zone suggest a reduced maturation of those regions, a result commonly seen in the presence of injury [48]. One would hypothesize that the extent of these abnormalities would predict neurodevelopmental outcome as reflected in the lower than expected motor function at 18 months.

While these two cases show what abnormalities STEAM captures at the level of a single scan, further insights can be gathered by considering both scans together. First, we note that these two scans show very different patterns of abnormality despite the fact that both have T1 scans showing white matter lesions and both have altered functional outcomes at 18 months corrected age. If we performed a group-based study where the experimental group contained both of these infant's scans, the best that study would be able to do is identify brain regions where the overwhelming majority of experimental group scans were different than the corresponding control group. That group study would not capture potentially large intra-group differences as displayed in these two cases.

Further, it is interesting to note that the infant in Figure 10 showed a greater amount of STEAM-detected abnormalities, as well as lower neurodevelopmental scores at 18 months, than the infant in Figure 11. The amount of abnormality identified by STEAM in these two cases agrees with their later neurodevelopmental outcome. The same cannot be said for the scores obtained from the T1 scan. The T1 scan for the second infant (in Figure 11) showed a greater presence of white matter lesions than the first infant (in Figure 10), as well as presence of intraventricular hemorrhage, yet the second infant showed better on neurodevelopmental tests at 18 months corrected age. While these results are only for two cases out of many, it raises the question of whether the volume of STEAM-detected abnormalities may be, in part, indicative of future neurodevelopmental outcome. We examine the potential of that abnormality-outcome relationship in the following section.

4.3. Relating STEAM Abnormalities to Outcome

While STEAM can be used to generate personalized abnormality maps, the question remains as to whether the abnormalities identified by STEAM are indeed meaningful and clinically relevant. We saw in the previous section that the volume of STEAM-detected abnormalities was indicative of neurodevelopmental outcome for two selected infants. Here, we examine whether that trend holds for our cohort as a whole. As a proof of concept, we narrow



(a) FA (b) MD (c) RD (d) λ_1

Figure 12: Comparison of STEAM-detected abnormalities in fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), and axial diffusivity (λ_1) between infants with normal and abnormal motor development. P-values were computed using two-way ANCOVA with sex, birth age, and brain volume included as covariates. Note that for all diffusion features except FA, the extent of STEAM-detected abnormalities is significantly higher for the infants with abnormal motor outcome.

our goal to identifying whether the volume of STEAM-detected abnormalities can be used to differentiate between infants with normal and abnormal motor outcomes.

Table 3 lists our cohort's experimental group according to Bayley Motor Score (18 months corrected age) and PMA at scan. Within this experimental group, we identify three main sub-groups based on outcome: those with a normal motor outcome (Bayley score > 85, highlighted in green), those with a clinically abnormal outcome (Bayley score ≤ 70, highlighted in red), and borderline cases (highlighted in yellow). To compare the volume of STEAM-detected abnormalities between the normal and abnormal groups, we first quantify the *extent* of STEAM-detected abnormalities as,

$$Extent = \frac{Volume\ of\ Abnormal\ Region\ (in\ voxels)}{Brain\ Volume\ (in\ voxels)} \tag{6}$$

so that the volume of STEAM-detected abnormalities is normalized by the brain volume. We then hypothesize that the extent of STEAM abnormalities should be significantly higher in the abnormal group than in the normal group.

To test for this group difference, we perform a two-way ANCOVA (analysis of covariance) with sex, age at birth, and brain volume included as covariates. We note that each of these three covariates have been identified as having an impact on infant neurodevelopment [9, 57, 58, 59] and we wish to isolate the effect of these covariates from the group differences that STEAM may identify.

The two-way ANCOVA results are presented in the boxplots in Figure 12 for the four most commonly studied diffusion measures: FA, MD, RD, and λ_1. For three of the four diffusion measures, we saw a significantly higher extent of STEAM-detected abnormalities in the abnormal outcome group than in the normal outcome group (MD: $p = 0.0127$, RD: $p = 0.0126$, λ_1: $p = 0.0067$). In the case of FA, the p-value of 0.0895 was not significant, but was still small enough to suggest that a group difference could exist depending on the cohort and on the template settings. We discuss this point further in Section 5. Note

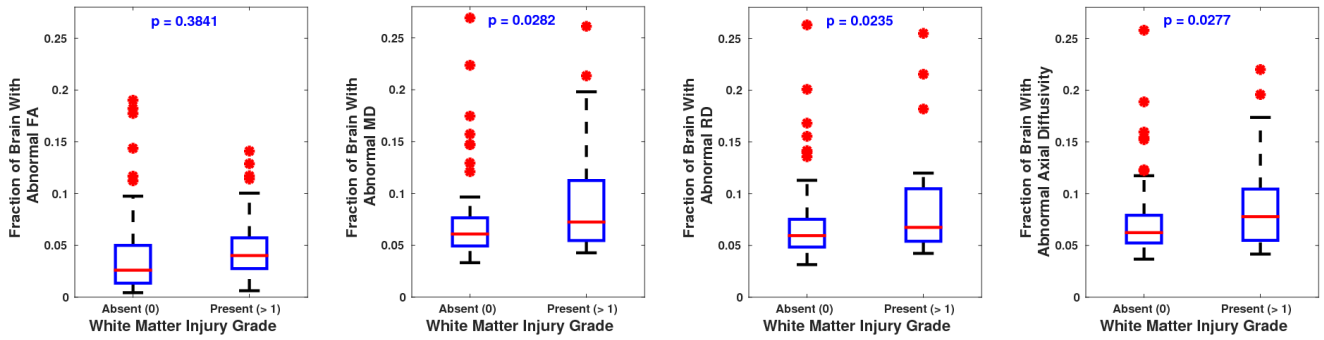
Table 3: The DTI scans used to test the relationship between our voxel-based analysis and motor outcome at 18 months corrected age. The number of scans are grouped according to post-menstrual age and Bayley motor score. Scans with clinically abnormal motor outcomes are highlighted in red while scans with normal motor outcomes are highlighted in green. Borderline, or "low normal", cases are highlighted in yellow.

Bayley Motor Score	Post-Menstrual Age (wks.) 27-31	32-36	37-40	41-45	Scan Total
> 100	5	8	7	3	23
$86 - 100$	14	23	11	9	57
$71 - 85$	7	7	3	5	22
≤ 70	2	7	2	0	11
Scan Total	28	45	23	17	113

that the experimental group does not include the scans used to create the statistical templates, so the group differences identified here are present despite of the fact that the scans in the normal motor outcome group showed abnormalities that eliminated them from being used in the templates themselves. Had our experimental group contained scans that met our control group criteria, we hypothesize that the group difference would be even larger.

With regard to the covariates of sex, age at birth, and brain volume, we found no significant relationships between them and the extent of STEAM detected abnormalities (smallest p-value = 0.0880 for brain volume and RD abnormalities, largest p-value = 0.8453 for birth age and λ_1 abnormalities). We believe that the lack of significant results for these covariates is a result of the fact that the templates do not differ along the dimensions of these covariates. When generating the statistical templates, all "healthy" scans are used equally and so the variability due to sex, age at birth, and brain volume is incorporated into the template itself, making it difficult to identify abnormalities related to those factors.

While the extent of STEAM abnormalities is able to differentiate, on average, between infants with normal and abnormal motor outcomes, we do note with red circles in



(a) FA (b) MD (c) RD (d) λ_1

Figure 13: Comparison of STEAM-detected abnormalities in fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), and axial diffusivity (λ_1) between infants with and without white matter lesions. P-values were computed using two-way ANCOVA with sex, birth age, and brain volume included as covariates. Note that for all diffusion features except FA, the extent of STEAM-detected abnormalities is significantly higher for the infants with white matter lesions, which is consistent with previous literature [2].

Figure 12 the presence of outlier results (i.e., false positives) for the normal outcome group. These outliers can be explained by the fact that the normal motor outcome group is comprised of scans that displayed at least one other measure of abnormality. Of the ten scans identified as outliers, three were from infants that had moderate to severe white matter injury as scored using [12], two were from infants that had low Bayley language scores (< 85), one showed presence of moderate IVH, and one was an outlier due to poor image registration (this outlier is shown in Figure 15 and discussed further in Section 5.1). The remaining three outlier scans were only outliers on the FA measure where the group difference was not significant.

Finally, we note that while the group differences were significant, the correlations between the extent of STEAM-detected abnormalities and Bayley motor scores were not (lowest p-value $= 0.087$ for axial diffusivity; p-value computed using a linear mixed effects model to account for the presence of multiple scans from the same infants). As a result, we cannot guarantee that the extent of STEAM-detected abnormalities are sufficient to predict motor outcome. It is likely that abnormality severity and location may also play roles for more refined predictions of outcome. Even so, these group differences are consistent with the results from the previous section and suggest that STEAM is capturing meaningful abnormalities across our cohort.

4.4. Comparing STEAM Abnormalities to T1 Abnormalities

While we have shown examples of how STEAM can generate personalized results and have shown that those results can be related to neurodevelopmental outcome, we have yet to look at how STEAM compares to existing ways of grading a preterm infant's individual MR scan. In particular, we are interested in knowing if the results obtained from STEAM are consistent with those obtained from another MRI grading system and whether STEAM provides any additional information over such a grading system. To examine this issue, we compare STEAM to a white matter injury (WMI) grading system that was found to be predictive of adverse neurodevelopmental outcome in preterm infants at 12 to 18 months of age [12]. This WMI grading system is based on the examination of the size and number of hyperintense white matter lesions seen on the infant's T1 MRI. These lesions have been identified as being indicative of axonal dematuration [2]. Given these lesion grades, we expect to see the extent of STEAM-detected abnormalities increase in the presence of lesions.

Table 4: The DTI scans used to test the relationship between our voxel-based analysis and the presence of white matter injury (WMI) scored according to [12]. The number of scans are grouped according to post-menstrual age and WMI score. Scans with significant white matter lesions are highlighted in red while lesion-free scans are highlighted in green. Scans with single, small lesions are highlighted in yellow.

WMI Grade [12]	Post-Menstrual Age (wks.) 27-31	32-36	37-40	41-45	Scan Total
0 - Absent	14	30	16	10	70
1 - Mild	3	6	5	3	17
2 - Moderate	6	8	2	1	17
3 - Severe	5	1	0	3	9
Scan Total	28	45	23	17	113

Table 4 lists our cohort's experimental group according to their white matter injury grade as defined in [12]. Lesion-free scans are identified in green while scans with a significant number of lesions are highlighted in red. More marginal cases containing single, small lesions (less than $2mm$ in diameter) are highlighted in yellow. To test whether the presence of lesions affects STEAM-detected abnormalities, we perform a two-way ANCOVA between the lesion-free group (in green) and the group of scans containing the presence of significant lesions (in red). We use ANCOVA here to test for a difference in STEAM-detected abnormalities between these groups while isolating the effects of sex, age at birth, and brain volume as covariates.

The two-way ANCOVA results are presented as boxplots in Figure 13 for the four studied diffusion measures: FA,

MD, RD, and λ_1. In all cases, we saw a larger amount of STEAM-detected abnormalities in the group with lesions than in the group without lesions. Excluding FA, the group differences were statistically significant (MD: $p = 0.0282$, RD: $p = 0.0235$, λ_1: $p = 0.0277$). In the case of FA, the larger STEAM-detected abnormalities in the group with lesions was not suggestive of a true group difference ($p = 0.3841$). Once again, note that these group differences are present despite the fact that our experimental group does not contain any of the scans used to create the statistical templates. As a result, the lesion-free scans used in this analysis still contain some abnormalities that eliminated them from being used to create the templates themselves.

While we are able to use STEAM abnormality extent to differentiate between lesion and non-lesion groups, we once again note the presence of outliers. These outlier scans in the non-lesion group can also be explained by some other measure of abnormality. Of the 12 scans identified as outliers, six were from infants with low Bayley motor scores (< 85), two were from infants with low Bayley Language scores (< 85), one showed presence of moderate IVH, and one scan was an outlier due to poor registration (this is the same outlier shown in Figure 15). The two remaining scans were only outliers on the FA measure where the group difference was not significant.

The presence of these outliers make it impossible for us to guarantee that the presence of white matter lesions always results in a greater extent of STEAM-detected abnormalities. In fact, Figures 10 and 11 show examples of scans where a greater lesion volume actually resulted in a lower extent of STEAM-detected abnormalities. That said, the group differences reported in Figure 13 do agree with the literature reviewed in [2] suggesting a link between hyperintense T1 lesions and more diffuse brain injuries.

5. Discussion

We have introduced herein the STEAM technique for the personalized analysis of DTI scans of the developing preterm infant brain. STEAM consists of two parts. First, we created a collection of statistical DTI templates for both the full diffusion tensor as well as a range of features derived from the diffusion tensors (e.g., FA, MD). Our template-creation pipeline is based on the technique of Guimond et al. and ensures an unbiased estimate of the average DTI scan of a population [21]. As part of that template estimation, we employed DT-REFinD, a full tensor DTI registration algorithm, to obtain the greatest accuracy we could in aligning anatomical structures across the DTI scans from our control group. The resulting templates contained the mean, variance, and normalcy p-value estimates at each voxel for a normative preterm infant population. This template estimation allows us to generate a statistical model *offline*, which reduces the amount of image registrations and statistical computations that need to be done to analyze a DTI scan on-the-fly.

The second component of STEAM is a full VBA processing pipeline that involves aligning an individual DTI scan to the template, then performing voxel-by-voxel statistical tests to identify abnormalities. Following the advice given in [19, 20], we examined various choices involved in setting up a VBA pipeline, in particular the multiple comparison correction scheme, what level of image smoothing to perform, and what to do with data that is not normally distributed. In all three cases, we followed the accepted convention in the VBA field and in the latter two cases, proposed the use of normalcy p-value images P_i and a collection of smoothed templates that capture a range spatial scales. The results from our STEAM analysis engine are summarized subject-specific abnormality maps, maps that no other existing technique provides.

We evaluated STEAM first qualitatively by showing that our generated templates display the type of brain development that is consistent with the reduction of the subplate zone [49, 11], the increased dendritic arborization in the cortex [16, 50], and the rapid development of the occipital lobes [48, 51, 52] that has been observed in previous DTI, MRI, and histological studies. We also showed qualitative examples of STEAM's voxel-based analysis on four common diffusion features (FA, MD, RD, λ_1) and identified how the resulting abnormality maps both corroborate and expand upon the results seen on T1 MRI scans [2].

We further evaluated STEAM quantitatively by performing VBA on the 113 DTI scans from our cohort that were not used in the creation of our templates. We showed that there exists a relationship between the extent of abnormalities detected by STEAM and neurodevelopmental outcome at 18 months corrected age. We also identified a relationship between the presence of white matter lesions and an increased volume of STEAM-detected abnormalities, which is consistent with existing literature [2]. These results serve as a proof of concept and show that STEAM is sufficiently reliable to be useful for preterm DTI analysis.

Finally, we have made our STEAM templates, as well as the source code for STEAM, publicly available to further facilitate research involving preterm DTI analysis. The code and templates are available at `http://www.sfu.ca/~bgb2/steam`. This STEAM website allows users to download the whole template collection, the source code, or even individual templates (Figure 14(a)). When a request is submitted (Figure 14(b)), a PHP script collects the requested files (in the case of the template images, they are provided in nifti format) into a single zip archive and emails a link to the archive to the requesting user. The STEAM website also allows for online 3D viewing of the STEAM templates used in this work (Figure 14(c)). Every image in our STEAM experiments can be viewed through this online image viewer and users can browse between different diffusion measures, age ranges, statistics, and image scales (Figure 14(d)). While we do provide access to our statistical templates, we recommend that users create their own templates for their studies as

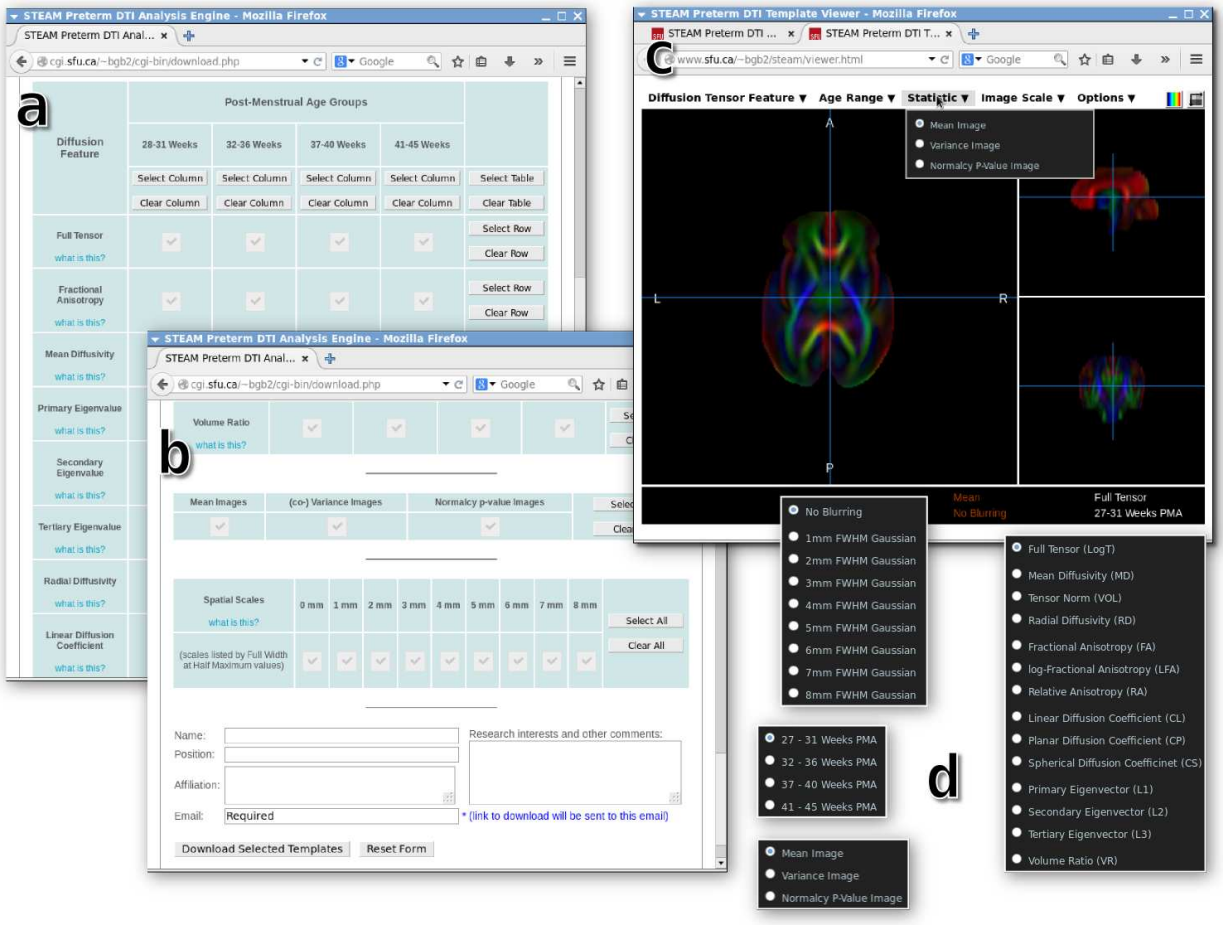


Figure 14: A collection of views from the STEAM website. From the website, one can download the STEAM source code, template collections (a) or even individual templates within each the collection (b). Also, the STEAM website boasts an online image viewer (c) that allows an interested user to examine each STEAM statistical template used in this work (d).

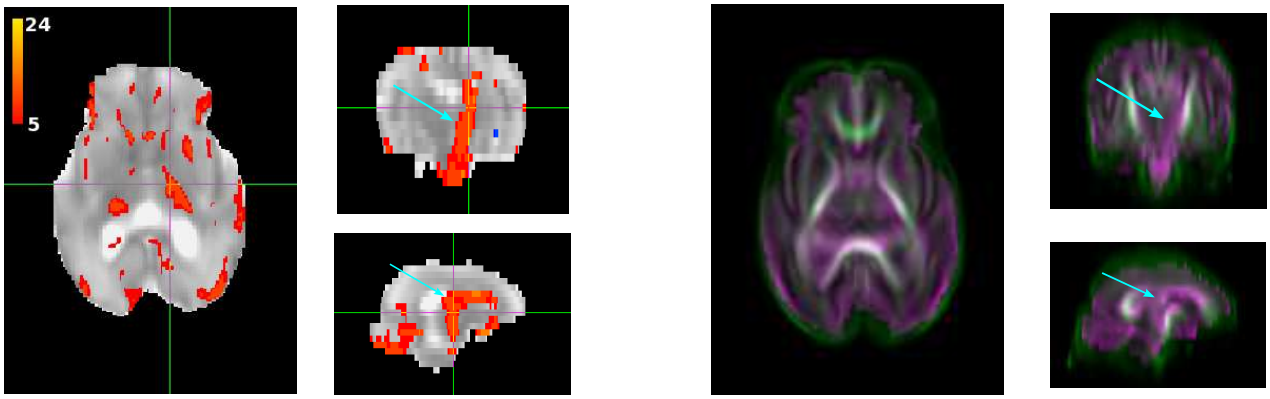
the choice of scanner and imaging protocol can impact diffusion measurements [60].

5.1. Limitations of STEAM

While STEAM has the strength of providing a fine-scale, personalized assessment of DTI abnormality over the whole brain, it is not without limitations. The abnormality detection that STEAM performs is based on VBA: a technique that has seen its share of criticism over the years [61, 20, 19]. The primary criticism has been the impact of registration quality on VBA results. If the image registration step does not succeed in aligning the anatomy in the DTI scans, the resulting statistical tests would not be comparing properties of the same anatomical region. The result of this registration error in the template creation step would be increased variance in the template, variance that could hide true abnormalities from being detected by STEAM. If there is registration error when comparing a new scan to the template, then misaligned structures would be identified as abnormalities. One such example was shown in Figure 10(c) where the posterior portion of an infant's ventricles did not align to the template's mean image.

We have attempted to mitigate the impact of image registration errors in multiple ways. First, we used DT-REFinD: a state-of-the-art tensor image registration algorithm that uses the full diffusion tensor to guide the image registration process [24]. In doing so, DT-REFinD is able to provide a more accurate structural alignment than registration algorithms based only on FA or some feature derived from the tensors [27]. Further, DT-REFinD is a non-linear registration algorithm that allows us to deform and align images with greater freedom than a linear registration algorithm like FSL FLIRT [22, 23]. Even so, results in Figure 10 suggest that registration error can still persist in STEAM, potentially because topological differences may exist between brains that image registration algorithms have difficulty addressing [62]. It is for this reason that we recommend examining the registration accuracy as part of the evaluation of STEAM results, as we did for both case studies in Figures 10 and 11.

To assess the impact of image misalignment on the STEAM abnormality maps, we recommend comparing the resulting abnormality maps (like those shown in Figure 10a) to the blended subject and template images (like those shown in Figure 10c). If the pattern of STEAM-



(a) FA STEAM abnormality map

(b) Blended FA images
(subject in purple, template in green)

Figure 15: An example of image misregistration leading to false positives in the STEAM abnormality maps. In (a), a large region of FA abnormality is identified by STEAM around the right corticospinal tract (shown highlighted by blue arrows). In (b), we see in the same area that the subject's right corticospinal tract (in purple) does not align with the corresponding tract in the template (again highlighted by blue arrows). In this case, the misalignment pattern matches the abnormality pattern, strongly suggesting that these STEAM-detected abnormalities are false positives.

detected abnormalities matches the pattern of misalignment, then it is best to consider those abnormalities to be false positives. One example whose FA results fit this description, and was the worst case in terms of misalignment in our cohort, is shown in Figure 15. Note that the abnormalities around the right corticospinal tract (highlighted by the blue arrows) match the evidence of misalignment in purple in the blended images. In a case like this, it would be appropriate to mark these abnormalities as false positives. On the other hand, if the pattern of STEAM-detected abnormalities extends well beyond the regions of misalignment, as they did in Figure 10, then there is reason to believe that the abnormalities are genuine.

The second criticism regarding VBA has to do with the large number of statistical tests that VBA performs, leading to the concern that a conservative multiple comparison correction scheme will cause some voxels with real diffusion differences to be marked as not being significantly different [61]. While this situation has not changed, we have made an effort to use a less conservative multiple comparison correction scheme in the false discovery rate. Further, we were able to show significant relationships between our VBA results and later neurodevelopmental outcome despite this concern. These results suggest that the large number of statistical tests VBA performs does not hinder our ability to identify abnormalities that could lead to future cognitive or motor impairment.

Additional concerns have been raised about VBA for diffusion tensor imaging, specifically with regards to the amount of image smoothing [20] and what to do with non-normally distributed data [19]. To address the first concern, we created templates of various scales of smoothing (from 0–8 mm) to facilitate a multi-scale VBA technique. We further used all scales in our experiments and aggregated the results from all scales before generating conclusions. We agree with Jones et al. [20] that examining multiple scales is arguably the most consistent way of performing VBA with diffusion tensor data. As for the second concern of non-normally distributed data, we generate p-value images from normalcy tests to allow for the removal of statistics computed at voxels where the assumption of normalcy is not valid. We recognize that removing these voxels from VBA means that there will be certain points in the brain where we would not be able to perform statistical tests. However, we found that over our entire range of templates, only 7.0% percent of the voxels would be removed from a STEAM analysis for not having normally-distributed diffusion data (this percentage decreases to 5.2% if we exclude volume ratio, VR in Figure 5).

Beyond the traditional concerns with the VBA of diffusion tensor images, there remain open questions with regards to template creation. Ideally, we would like to create statistical templates from a control group whose demographics match as closely as possible to the infants who are compared to that template. In that way, we would avoid differences, and template variance, that are due to confounding factors such as sex, GA at birth, or PMA at scan. However, this ideal situation is difficult to meet with a limited number of DTI scans. At the same time, too few scans can also lead to errors when estimating the means and (co-)variances at each voxel in the template. In our experiments, we attempted to trade-off these limitations in two ways. First, we split our control group into four different age windows, allowing us to reduce variance in the templates due to age at scan while still having enough scans from which we could compute template statistics. Second, we modeled the effects of multiple confounding factors in our statistical analyses, specifically sex, birth age, and brain volume. By taking these two steps, we were able to maintain large enough control groups to estimate statistical templates while also mitigating, to some

degree, the impact of confounding factors on our STEAM analysis engine.

Similarly, differences in scanners, imaging protocols, and tensor fitting software can also introduce variability between DTI scans [60]. The presence of these confounding factors makes it unlikely that a single statistical template collection, like the one created here, would properly model a normative population acquired at a different site. In this work, we ensured that the DTI scans used in our templates and in our evaluation of STEAM were all acquired using the same scanner, imaging protocol, and processed using the same software packages (see Section 3.1). By fixing these three factors, we were able to eliminate their impact from our STEAM analysis engine.

Finally, we acknowledge that STEAM is based on the diffusion tensor model, a model that has limitations, most prominently the inability to model more than one fiber tract at a voxel [45]. More recent diffusion modeling techniques, including HARDI (high angular resolution diffusion imaging [63]), DKI (diffusion kurtosis imaging [64]), and DSI (diffusion spectrum imaging [65]), are addressing this concern to the point where a shift away from the diffusion tensor model may be warranted. If such an imaging shift occurs, STEAM will be able to accommodate that change. The two aspects of STEAM that are tensor-specific are the registration technique (DT-REFinD [24]) and the log-Euclidean mapping [26] that maps tensors to a vector space. Any other diffusion model, or even imaging modality, can be used in the STEAM analysis engine by simply swapping out those two pieces of the pipeline (i.e., the registration and vector space mapping) with ones that are specific to the new imaging type. In the case of newer diffusion models, such registration methods and vector space mappings are already being developed [66, 67, 68, 69, 70].

5.2. Future Work

While we have presented STEAM and shown that it can work as a personalized DTI evaluation technique, limitations in STEAM, and a full clinical evaluation of STEAM, remain. As a result, there remains various avenues for future work.

When it comes to the creation of statistical templates, future work will look at generating templates that are less impacted by confounding factors. In the short term, we continue to expand our cohort, which will increase the number of scans we have to generate statistical templates. Increasing the number of scans from our control group will allow us to generate templates along different dimensions, from sex, to birth age, to age at scan. Over the long term, we intend to examine whether scans can be weighted differently in order to account for confounding factors when computing mean and (co-) variance images. For example, Serag et al. employed a weighting technique for atlas creation, albeit with a single covariate (age at scan) [71].

Additional future work can also be done to address confounding factors such as differing scanner types, imaging protocols, and preprocessing software. The impact of different scanner types on a STEAM analysis may be the most limiting of these confounding factors. While a site-specific template can be made to circumvent this issue, such a template requires - and is built from - a normative DTI dataset acquired at that site. Acquiring this dataset may be a challenge for certain research groups, thereby limiting STEAM's usability. Future work will focus on measuring the impact of scanner choice on a STEAM analysis. Further, one may explore the introduction of an intensity normalization step to the preprocessing of diffusion weighted images as a way of addressing these confounding factors. A similar technique is commonly applied to structural MRI studies where images are acquired from multiple sites (e.g., [72]). Alternatively, transfer learning may also be helpful in identifying, through the use of a training set, an intensity mapping that corrects for these confounding factors [73].

When it comes to the voxel-based analysis, future work will focus on improving image registration techniques and highlighting areas of high registration uncertainty (or high predicted error). A recent review of image registration techniques has highlighted various future directions [74], including varying how registration algorithms trade-off image similarity and the rigidity of the deformation (e.g., [75]), as well as merging results from different image registration algorithms to generate a more accurate image alignment (an idea known as meta-registration) (e.g., [76]). Further, more recent works have quantified image registration uncertainty, exposed correlations between uncertainty and error, and used registration confidence (i.e., lack of uncertainty) measures to improve image registration [77, 78]. We intend to follow these developments and examine their impact on the accuracy of the results STEAM generates.

Further, we recognize that the experiments presented here provide merely a proof of concept. We have only shown a sample of the wide variety of experimental groupings, clinical measures, motor, cognitive, behavioral and other neurodevelopmental outcomes which we can examine with STEAM. Also, while we have examined the extent of STEAM-detected abnormalities, other aspects of the VBA results - like magnitude, direction, and abnormality location - may also provide valuable clues regarding a preterm infant's neurodevelopmental outcome. STEAM's VBA-based features provides a wealth of new information that, once fully exploited, may have the potential to predict future neurodevelopmental outcome. We intend to use these new features to complement our recent connectome-based predictor of motor outcome [79] to develop a more holistic predictor of preterm brain health.

While we have presented STEAM in the context of preterm infant DTI analysis, STEAM can be equally applied to other cohorts provided that a normative group of DTI scans are available for template estimation. That said, certain settings in the STEAM analysis engine – particularly the choices of image smoothing, registration tech-

nique, and multiple comparison correction scheme – may be application-specific [19, 20]. Future use of STEAM should include fine tuning of these pipeline settings.

Finally, the number of test scans in our cohort, particularly from infants with abnormal neurodevelopmental outcomes, is modest at this time. As the number of these cases increase, we will be able to say with greater certainty just how important STEAM-detected abnormalities are in the understanding of preterm brain health.

6. Conclusion

STEAM provides a personalized technique for analyzing diffusion measurements at a fine scale over the whole preterm infant brain, something other techniques have yet to claim. The technique is based on generating statistical template images for a normative population, then comparing new scans to those templates using voxel-based analysis. The result of this analysis is an abnormality map that highlights areas of significantly abnormal brain development. We have made the source code for STEAM publicly available as well as statistical templates of the preterm infant brain that were generated in this study[5]. It is our hope that these contributions can further the field's progress towards imaging biomarkers for preterm brain injury and the prediction of neurodevelopmental outcomes.

Acknowledgements

We thank the staff of the Neonatal Follow-Up Program of BC Children's & Women's Hospitals for their valuable contributions in assessing these children in the neonatal ICU and following up on their neurodevelopmental outcomes. We also thank Emma Duerden and Elysia Adams for the white matter lesions segmentations shown in Figures 10b and 11b. Finally, we thank the editors and reviewers of NeuroImage for their valuable and constructive feedback.

CJB and GH were partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN298324, and BGB by NSERC, IODE Canada, and the Government of Alberta. SPM was supported by Canadian Institutes of Health Research (CIHR) operating grant MOP-79262. SPM is currently the Bloorview Children's Hospital Chair in Pediatric Neuroscience and was supported by a Tier 2 Canada Research Chair in Neonatal Neuroscience, and Michael Smith Foundation for Health Research Scholar. REG was supported by Canadian Institutes of Health Research (CIHR) operating grant MOP-86489 and holds a Senior Scientist Award from the Child & Family Research Institute.

[5] `http://www.sfu.ca/~bgb2/steam`

References

[1] Hannah Blencowe, Simon Cousens, Mikkel Z Oestergaard, Doris Chou, Ann-Beth Moller, Rajesh Narwal, Alma Adler, Claudia Vera Garcia, Sarah Rohde, Lale Say, and Joy E Lawn. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*, 379(9832):2162–2172, 2012.

[2] Stephen A. Back and Steven P. Miller. Brain injury in premature neonates: A primary cerebral dysmaturation disorder? *Annals of Neurology*, 75(4):469–486, 2014.

[3] Jeroen Dudink, Jenny L. Kerr, Kathryn Paterson, and Serena J. Counsell. Connecting the developing preterm brain. *Early Human Development*, 84:777–782, 2008.

[4] Alec Aeby, Xavier De Tiège, Marylise Creuzil, Philippe David, Danielle Balériaux, Bart Van Overmeire, Thierry Metens, and Patrick Van Bogaert. Language development at 2 years is correlated to brain microstructure in the left superior temporal gyrus at term equivalent age: A diffusion tensor imaging study. *NeuroImage*, 78:145–151, 2013.

[5] Michelle L. Krishnan, Leigh E. Dyet, James P. Boardman, Olga Kapellou, Joanna M. Allsop, Frances Cowan, A. David Edwards, Mary A. Rutherford, and Serena J. Counsell. Relationship between white matter apparent diffusion coefficients in preterm infants at term-equivalent age and developmental outcome at 2 years. *Pediatrics*, 120:e604–e609, 2007.

[6] Stephen E. Rose, Xanthy Hatzigeorgiou, Mark W. Strudwick, Gail Durbridge, Peter S.W. Davies, and Paul B. Colditz. Altered white matter diffusion anisotropy in normal and preterm infants at term-equivalent age. *Magnetic Resonance in Medicine*, 60:761–767, 2008.

[7] Serena J. Counsell, A. David Edwards, Andrew T M. Chew, Mustafa Anjari, Leigh E. Dyet, Latha Srinivasan, James P. Boardman, Joanna M. Allsop, Joseph V. Hajnal, Mary A. Rutherford, and Frances M. Cowan. Specific relations between neurodevelopmental abilities and white matter microstructure in children born preterm. *Brain*, 131:3201–3208, 2008.

[8] Laura Bassi, Daniela Ricci, Anna Volzone, Joanna M. Allsop, Latha Srinivasan, Aakash Pai, Carmen Ribes, Luca A. Ramenghi, Eugenio Mercuri, Fabio Mosca, A. David Edwards, Frances M. Cowan, Mary A. Rutherford, and Serena J. Counsell. Probabilistic diffusion tractography of the optic radiations and visual function in preterm infants at term equivalent age. *Brain*, 131:573–582, 2008.

[9] Jessica Rose, Erin E. Butler, Lauren E. Lamont, Patrick D. Barnes, Scott W. Atlas, and David K. Stevenson. Neonatal brain structure on MRI and diffusion tensor imaging, sex, and neurodevelopment in very-low-birthweight preterm children. *Developmental Medicine and Child Neurology*, 0:526–535, 2009.

[10] Vann Chau, Anne Synnes, Ruth E Grunau, Kenneth J Poskitt, Rollin Brant, and Steven P Miller. Abnormal brain maturation in preterm neonates associated with adverse developmental outcomes. *Neurology*, 81(24):2082–2089, 2013.

[11] Jessica Rose, Rachel Vassar, Katelyn Cahill-Rowley, Ximena Stecher Guzman, David K. Stevenson, and Naama Barnea-Goraly. Brain microstructural development at near-term age in very-low-birth-weight preterm infants: An atlas-based diffusion imaging study. *NeuroImage*, 86:244–256, 2014.

[12] S.P. Miller, D.M. Ferriero, C. Leonard, R. Piecuch, D.V. Glidden, J.C. Partridge, and et al. Early brain injury in premature newborns detected with magnetic resonance imaging is associated with adverse early neurodevelopmental outcome. *Pediatrics*, 147:609–616, 2005.

[13] L.A. Papile, J. Burstein, R. Burstein, and H. Koffler. Incidence and evolution of subependymal and intraventricular hemorrhage: a study of infants with birth weights less than 1,500 gm. *Pediatrics*, 92:529–534, 1978.

[14] Colin J. Brown, Steven Miller, Brian G. Booth, Shawn Andrews, Vann Chau, Kenneth Poskitt, and Ghassan Hamarneh. Struc-

tural network analysis of brain development in young preterm neonates. *NeuroImage*, 101:667–680, 2014.

[15] Kenichi Oishi, Andreia V. Faria, Shoko Yoshida, Linda Chang, and Susumu Mori. Quantitative evaluation of brain development using anatomical MRI and diffusion tensor imaging. *International Journal of Developmental Neuroscience*, 31(7):512 – 524, 2013.

[16] Jillian Vinall, Ruth E. Grunau, Rollin Brant, Vann Chau, Kenneth J. Poskitt, Anne R. Synnes, and Steven P. Miller. Slower postnatal growth is associated with delayed cerebral cortical maturation in preterm newborns. *Science in Translational Medicine*, 5(168):1–9, 2013.

[17] Mónica Giménez, Maria J. Miranda, A. Peter Born, Zoltan Nagy, Egill Rostrup, and Terry L. Jernigan. Accelerated cerebral white matter development in preterm infants: A voxel-based morphometry study with diffusion tensor MR imaging. *NeuroImage*, 41:728–734, 2008.

[18] John Ashburner and Karl J. Friston. Voxel-based morphometry – the methods. *NeuroImage*, 11:805–821, 2000.

[19] Derek K. Jones and Mara Cercignani. Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR in Biomedicine*, 23:803–820, 2010.

[20] Derek K. Jones, Mark R. Symms, Mara Cercignani, and Robert J. Howard. The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage*, 26:546–554, 2005.

[21] Alexandre Guimond, Jean Meunier, and Jean-Philippe Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77:192–210, 2000.

[22] M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[23] M. Jenkinson and S.M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[24] B.T. Thomas Yeo, Tom Vercauteren, Pierre Fillard, Jean-Marc Peyrat, Xavier Pennec, Polina Golland, Nicholas Ayache, and Olivier Clatz. DT-REFinD: Diffusion tensor registration with exact finite-strain differential. *IEEE Transactions on Medical Imaging*, 28:1914–1928, 2009.

[25] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parameteric image registration. *NeuroImage*, 45:S61–S72, 2009.

[26] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Mag. Res. Med.*, 56:411–421, 2006.

[27] Hae-Jeong Park, Marek Kubicki, Martha E. Shenton, Alexandre Guimond, Robert W. McCarley, Stephen E. Maier, Ron Kikinis, Ferenc A. Jolesz, and Carl-Fredrik Westin. Spatial normalization of diffusion tensor MRI using multiple channels. *NeuroImage*, 20:1995–2009, 2003.

[28] Zhizhou Wang and Baba C. Vemuri. An affine invariant tensor dissimilarity measure and its applications to tensor-valued image segmentation. In *Proceedings of Computer Vision and Pattern Recognition – CVPR 2004*, volume 1, pages 228–233, 2004.

[29] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–30, 2005.

[30] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3617, 1990.

[31] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[32] Brent R. Logan and Daniel B. Rowe. An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, 22:95–108, 2004.

[33] Thomas Nichols. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12(5):419–448, 2003.

[34] C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[35] K.J. Worsley, S. Marrett, P. Neelin, A.C. Vandal, K.J. Friston, and A.C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996.

[36] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

[37] T.E. Nichols and A.P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15:1–25, 2001.

[38] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

[39] C.-F. Westin, S.E. Maier, H. Mamata, A. Nabavi, F.A. Jolesz, and R. Kikinis. Processing and visualization for diffusion tensor mri. *Medical Image Analysis*, 6:93–108, 2002.

[40] H.W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

[41] N. Bayley. *Manual for the Bayley Scales of Infant and Toddler Development*. Harcourt, San Antonio, 3rd edition, 2006.

[42] M. Rhonda Folio and Rebecca R. Fewell. *Peabody Developmental Motor Scales*. PRO-Ed, 2nd edition, 2000.

[43] Lin-Ching Chang, Derek K. Jones, and Carlo Pierpaoli. RESTORE: Robust estimation of tensors by outlier rejection. *Magnetic Resonance in Medicine*, 53:1088–1095, 2005.

[44] J.D. Cardoza, R.B. Goldstein, and R.A. Filly. Exclusion of fetal ventriculomegaly with a single measurement: the width of the lateral ventricular atrium. *Radiology*, 169:711–714, 1988.

[45] Jacques-Donald Tournier, Susumu Mori, and Alexander Leemans. Diffusion tensor imaging and beyond. *Magnetic Resonance in Medicine*, 65:1532–1556, 2011.

[46] Daniel Gallichan, Jan Scholz, Andreas Bartsch, Timothy E. Behrens, Matthew D. Robson, and Karla L. Miller. Addressing a systematic vibration artifact in diffusion-weighted MRI. *Human Brain Mapping*, 31:193–202, 2010.

[47] Savannah C. Partridge, Pratik Mukherjee, Roland G. Henry, Steven P. Miller, Jeffrey I. Berman, Hua Jin, Ying Lu, Orit A. Glenn, Donna M. Ferriero, A. James Barkovich, and Daniel B. Vigneron. Diffusion tensor imaging: serial quantitation of white matter tract maturity in premature newborns. *NeuroImage*, 22:1302–1314, 2004.

[48] Emi Takahashi, Rebecca D. Folkerth, Albert M. Galaburda, and Patricia E. Grant. Emerging cerebral connectivity in the human fetal brain: An MR tractography study. *Cerebral Cortex*, 22(2):455–464, 2012.

[49] Ivica Kostovic, Milos Judas, Marko Rados, and Pero Hrabac. Laminar organization of the human fetal cerebrum revealed by histochemical markers and magnetic resonance imaging. *Cerebral Cortex*, 12:536–544, 2002.

[50] Zoltan Molnar and Mary Rutherford. Brain maturation after preterm birth. *Science and Translational Medicine*, 5(168):168ps2, 2013.

[51] Peter R Huttenlocher and Arun S Dabholkar. Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology*, 387(2):167–178, 1997.

[52] Kate Teffer, Daniel P. Buxhoeveden, Cheryl D. Stimpson, Archibald J. Fobbs, Steven J. Schapiro, Wallace B. Baze, Mark J. McArthur, William D. Hopkins, Patrick R. Hof, and Chet C. Sherwood et al. Developmental changes in the spatial organization of neurons in the neocortex of humans and common chimpanzees. *Journal of Comparative Neurology*, 521(18):4249–4259, 2013.

[53] Malcolm Battin and Mary A Rutherford. *MRI of the Neonatal Brain*, chapter 3 - Magnetic resonance imaging of the brain in preterm infants: 24 weeks gestation to term. Saunders Ltd.,

London, 2002.

[54] M. Ono, S. Kubick, and C.D. Abernathey. *Atlas of the cerebral sulci*. Thieme Medical Publishers, New York, 1990.

[55] B.J.M. van Kooij, L.S. de Vries, G. Ball, I.C. van Haastert, M.J.N.L. Benders, F. Groenendaal, and S.J. Counsell. Neonatal tract-based spatial statistics findings and outcome in preterm infants. *American Journal of Neuroradiology*, 33:188–194, 2012.

[56] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19(0):4 – 19, 2014.

[57] Jeroen Dudink, Maarten Lequin, Carola van Pul, Jan Buijs, Nikk Conneman, Johannes van Goudoever, and Paul Govaert. Fractional anisotropy in white matter tracts of very-low-birth-weight infants. *Pediatric Radiology*, 37:1216–1223, 2007.

[58] Tatsuji Hasegawa, Kei Yamada, Masafumi Morimoto, Shigemi Morioka, Takenori Tozawa, Kenichi Isoda, Aki Murakami, Tomohiro Chiyonobu, Sachiko Tokuda, Akira Nishimura, Tsunehiko Nishimura, and Hajime Hosoi. Development of corpus callosum in preterm infants is affected by the prematurity: In vivo assessment of diffusion tensor imaging at term-equivalent age. *Pediatric Research*, 69(3):249–254, 2011.

[59] Gareth Ball, James P. Boardman, Daniel Rueckert, Paul Aljabar, Tomoki Arichi, Nazakat Merchant, Ioannis S. Gousias, A. David Edwards, and Serena J. Counsell. The effect of preterm birth on thalamic and cortical development. *Cerebral Cortex*, 22(5):1016–1024, 2012.

[60] Tong Zhu, Rui Hu, Xing Qiu, Michael Taylor, Yuen Tso, Constantin Yiannoutsos, Bradford Navia, Susumu Mori, Sven Ekholm, Giovanni Schifitto, and Jianhui Zhong. Quantification of accuracy and precision of multi-center dti measurements: A diffusion phantom and human brain study. *NeuroImage*, 56:1398–1411, 2011.

[61] Stephen M. Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins, Olga Ciccarelli, M. Zaheer Cader, Paul M. Matthews, and Timothy E.J. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31:1487–1505, 2006.

[62] Alan C. Evans, Andrew L. Janke, D. Louis Collins, and Sylvain Baillet. Brain templates and atlases. *NeuroImage*, 62(2):911–922, 2012.

[63] David S. Tuch, Timothy G. Reese, Mette R. Wiegell, Nikos Makris, John W. Belliveau, and Van J. Wedeen. High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine*, 48:577–582, 2002.

[64] Jens H. Jensen, Joseph A. Helpern, Anita Ramani, Hanzhang Lu, and Kyle Kaczynski. Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine*, 53:1432–1440, 2005.

[65] V. Wedeen, T. Reese, D. Tuch, M. Wiegel, J.-G. Dou, R. Weiskoff, and D. Chessler. Mapping fiber orientation spectra in cerebral white matter with fourier-transform diffusion MRI. In *Proceedings of the International Society of Magnetic Resonance in Medicine*, page 82, 2000.

[66] Jia Du, Alvina Goh, and Anqi Qiu. Diffeomorphic metric mapping of high angular resolution diffusion imaging based on riemannian structure of orientation distribution functions. *IEEE Transactions on Medical Imaging*, 31(5):1021–1033, 2012.

[67] Xiujuan Geng, Thomas J. Ross, Hong Gu, Wanyong Shin, Wang Zhan, Yi-Ping Chao, Ching-Po Lin, Norbert Schuff, and Yihong Yang. Diffeomorphic image registration of diffusion MRI using spherical harmonics. *IEEE Transactions on Medical Imaging*, 30(3):747–758, 2011.

[68] Yung-Chin Hsu, Ching-Han Hsu, and Wen-Yih Isaac Tseng. A large deformation diffeomorphic metric mapping solution for diffusion spectrum imaging datasets. *NeuroImage*, 63:818–834, 2012.

[69] David Raffelt, J-Donald Tournier, Jurgen Fripp, Stuart Crozier, Alan Connelly, and Olivier Salvado. Symmetric diffeomorphic registration of fibre orientation distributions. *NeuroImage*, 56:1171–1180, 2011.

[70] Guang Cheng, Baba C. Vemuri, Paul R. Carney, and Thomas H. Mareci. Non-rigid registration of high angular resolution diffusion images represented by gaussian mixture fields. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, volume 5761 of *LNCS*, pages 190–197. Springer, 2009.

[71] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J. Counsell, James P. Boardman, Mary A. Rutherford, A. David Edwards, Joseph V. Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3):2255–2265, 2012.

[72] Kelvin K. Leung, Matthew J. Clarkson, Jonathan W. Bartlett, Shona Clegg, Clifford R. Jack Jr., Michael W. Weiner, Nick C. Fox, and Sbastien Ourselin. Robust atrophy rate measurement in alzheimer's disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection. *NeuroImage*, 50(2):516–523, 2010.

[73] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[74] Yangming Ou, Hamed Akbari, Michel Bilello, Xiao Da, and Christos Davatzikos. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights. *IEEE Transactions on Medical Imaging*, 33(10):2039–2065, 2014.

[75] Lisa Tang, Ghassan Hamarneh, and Rafeef Abugharbieh. Reliability-driven, spatially-adaptive regularization for deformable registration. In *Workshop on Biomedical Image Registration (WBIR)*, pages 173–185, 2010.

[76] Sharmishtaa Seshamani, Purnima Rajan, Rajesh Kumar, Hani Girgis, Themos Dassopoulos, Gerard Mullin, and Gregory Hager. A meta registration framework for lesion matching. In *Medical Image Computing and Computer-Assisted Intervention MICCAI*, pages 582–589. Springer, 2009.

[77] Lisa Tang and Ghassan Hamarneh. Random walks with efficient search and contextually adapted image similarity for deformable registration. In *Medical Image Computing and Computer-Assisted Intervention MICCAI*, pages 43–50. Springer, 2013.

[78] Tayebeh Lotfi, Lisa Tang, Shawn Andrews, and Ghassan Hamarneh. Improving probabilistic image registration via reinforcement learning and uncertainty evaluation. In *MICCAI Workshop on Machine Learning in Medical Imaging (MLMI)*, pages 188–195, 2013.

[79] Colin J Brown, Steven P Miller, Brian G Booth, Kenneth J Poskitt, Vann Chau, Anne R Synnes, Jill G Zwicker, Ruth E Grunau, and Ghassan Hamarneh. Prediction of motor function in very preterm infants using connectome features and local synthetic instances. In *Medical Image Computing and Computer-Assisted Intervention MICCAI*, pages 69–76. Springer, 2015.