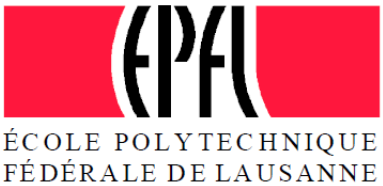
ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
SCHOOL OF LIFE SCIENCES



Master's project in Life Sciences and Technology

Using medical images acquired during routine clinical care for research purposes: a comprehensive medical informatics approach applied to the study of brain development from birth through 2 years of age as indexed by diffusion MRI.

Carried out in the Laboratory for Neuroimaging Applications to Pain, Acupuncture and Placebo Research
at Athinoula A. Martinos Center for Biomedical Imaging, Boston, MA, USA
Under the supervision of Randy L. Gollub, MD, PhD

Done by

VINCENT ROCH

Under the direction of

PROF. Nouchine Hadjikhani

In the laboratory of Social Cognitive Neurosciences
EPFL

LAUSANNE, EPFL 2011

Contact information

Vincent Roch
Route de la Borette 14
1890 St-Maurice
Valais
Switzerland

Swiss phone : +41 79 785 02 62
US pone : +1 857 264 99 20

e-mail addresses : vincent.roch@epfl.ch
vincentr@nmr.mgh.harvard.edu

Boston, 4th of August 2011 Vincent Roch

Acknowledgments

I would like to sincerely thank Pr. Nouchine Hadjikhani for accepting to be my master project advisor at EPFL, for following me throughout this project, and for her availability and support. She kindly introduced me to Dr. Randy Gollub who generously accepted me in her laboratory where I carried out this work for the last year. Despite her very busy schedule, Dr. Randy Gollub was always available to lend her guidance when I needed help and she always encouraged me, especially in difficult times. She notably allowed me to accomplish this work by introducing me to knowledgeable researchers who were invaluable all the way through this project.

My acknowledgments go to Dr. Ellen Grant, initiator of the Baby Brain project, whose well-informed expertise was crucial for the success of this project. I want to also thank Lilla Zollei and Rudolph Piennar who kindly and regularly helped me with processing tools and script coding.

I am grateful to the mi2b2 team and its Principal Investigators, Dr. Shawn Murphy and Dr. Randy Gollub for allowing me to participate in their weekly meetings and for allowing me the honor of using the first version of the prototype of mi2b2 software. I want to particularly thank Bill Wang, Chris Herrick and David Wang from the mi2b2 team for helping me retrieve all the medical images using the newly developed software.

I also would like to express my gratitude to all my lab colleagues who helped me and supported me in different ways. Specifically, Rita Loiotile for Unix scripting, Karin Jensen and Marco Loggia for statistical analyses, and Rosa Spaeth for the final revision of the thesis. My particular gratitude goes to Alexandra Cheetham who took care of all administrative procedures and who carefully revised the first draft of my thesis with helpful suggestions.

Finally, I would like to express my gratitude to all my friends, relatives and close family who have constantly supported me in my study, with a special thanks to my Parents, Esther and Jean-Didier Roch, for their unconditional encouragements.

This work was supported by National Institutes of Health, National Center for Research Resources (NIH-NCRR) as an ARRA supplement to Harvard Medical School's Clinical Translational Science Award (the Harvard Catalyst) under the "Medical Imaging Informatics Bench to Bedside (mi2b2) grant NCRR (1 UL1 RR 025758-01) and the Biomedical Informatics Research Network (1 U24 RR 025736-01).

Summary

Biologic variability and dramatic changes of brain development in children aged 0 to 2 years make it challenging to accurately detect subtle abnormalities in single Magnetic Resonance Imaging (MRI) scans. Diffusion MRI (dMRI) indices such as Apparent Diffusion Coefficient (ADC) are reliable measures of water content in the brain and thus an excellent surrogate marker for brain development. Developing robust age-specific diffusion biomarkers for quantitative measurement of normative brain evolution would enhance our ability to detect subtle alterations due to tissue injuries or neuropathological disorders. Obtaining significant numbers of normative MRI scans for this age group means redirecting clinical data from hospital databases for research purposes.

Therefore, this pilot project demonstrates the feasibility of identifying, retrieving and analyzing pediatric clinical dMRI data to investigate normal brain development from birth to 2 years.

Research Patient Data Registry (RPDR) at Massachusetts General Hospital (MGH) was used to collect patient medical information and identify healthy children according to radiology reports. Corresponding MRI data were retrieved from MGH Picture Archiving and Communication System (PACS) using the prototype of Medical Imaging Informatics Bench to Bedside (mi2b2) software. A specific pipeline was created to handle the volume of studies and extract technical scan information used to identify comparable diffusion series; 193 studies were used for analysis.

Two markers, whole brain average of ADC and Fractional Anisotropy (FA) values (WBA_{ADC} and WBA_{FA}), were computed for each patient, their age-evolution across patients was investigated with different models. WBA_{ADC} and WBA_{FA} seem to exhibit biexponential decay and increase respectively and might be gender-specific.

These results have clinical implications for potentially determining the health status of an unknown individual, and research utility for continued development of these tools.

Table of Contents

1. INTRODUCTION **1**

2. BACKGROUND **4**

2.1. DIFFUSION MAGNETIC RESONANCE IMAGING (DMRI) 4
2.2. DIFFUSION MRI OF BRAIN DEVELOPMENT 7
2.3. BRAIN ATLASES OF HEALTHY BABIES AND CHILDREN 10
2.4. CLINICAL APPLICATIONS OF DIFFUSION IMAGING 11

3. MATERIALS AND METHODS **14**

3.1. INSTITUTIONAL REVIEW BOARD (IRB) APPROVAL 14
3.2. DATABASE MINING (MEDICAL INFORMATICS) 14
3.2.1. RPDR query *15*
3.2.2. Microsoft Access Database queries *18*
3.2.2.1. Number of brain MRI scans for each age range 20
3.2.2.2. Proportion of MRIs that include diffusion imaging 21
3.2.2.3. How old are the scans? 22
3.2.2.4. "Normal" or "Abnormal" brain MRI scan? 22
3.2.2.5. Longitudinal data 25
3.2.3. MRI data retrieval from PACS – mi2b2 software *25*
3.2.3.1. PACS 25
3.2.3.2. mi2b2 software 26
3.2.4. Automated MRI data organization: the "BB-pipeline" *28*
3.2.5. Selecting data of interest *33*
3.3. MRI DATA ANALYSIS 34
3.3.1. Average calculation of whole brain diffusion indices *35*
3.3.2. Diffusion indices time evolution analysis *37*

4. RESULTS **39**

4.1. DATABASE MINING (MEDICAL INFORMATICS) 39
4.1.1. RPDR query *39*
4.1.2. Microsoft Access Database queries *39*
4.1.2.1. Number of brain MRI scan for each age range 40
4.1.2.2. Proportion of MRIs that include diffusion imaging 41
4.1.2.3. How old are the scans? 41
4.1.2.4. "Normal" or "Abnormal" brain MRI scan? 42
4.1.2.5. Longitudinal data 42
4.1.3. PACS retrieval *44*
4.1.4. Data of interest *44*
4.2. TIME EVOLUTION OF DIFFUSION INDICES ACROSS AGES 49
4.2.1. Data processing *49*
4.2.2. Curve fitting analysis *51*
4.2.3. GLM analysis *54*
4.2.4. Within patient WBA ADC/FA time evolution controls *55*

5. DISCUSSION **58**

5.1. DATABASE MINING 58
5.2. DATA ANALYSIS 60

6. CONCLUSION **64**

7. REFERENCES 65

7.1. PAPERS 65
7.2. BOOKS 68
7.3. COURSES 68
7.4. WEB SITES 68
7.5. OTHER 69

8. APPENDIX 70

8.1. ABBREVIATIONS 70
8.2. FIGURES AND TABLES 72
8.2.1. Figures 72
8.2.2. Tables 74
8.3. PEOPLE 75
8.4. BRAIN MRI PROCEDURES IN RPDR 77
8.5. HIV RELATED DIAGNOSIS 78
8.6. COMPLETE LIST OF TABLES RECEIVED FROM RPDR 79
8.7. IMPRESSION NOTES ASSESSING "NORMAL" BRAINS 80
8.8. DICOM HEADERS 81
8.9. ESTIMATED PARAMETERS 81

1. Introduction

This particular work using clinical data to study newborn and infant brain development emerged a few months ago, resulting from the convergence of two complementary on-going research projects. Specifically, the Medical Imaging Informatics Bench to Bedside (mi2b2) project implemented at MGH and Dr. Ellen Grant's work on pediatric data (see People section, part 8.3). My work has been made possible by this union because the mi2b2 project facilitates retrieval of medical images, while Dr. Ellen Grant's work justifies the usefulness of mi2b2. Let me explain:

Hospital databases are real gold mines in terms of the amount, quality and specificity of biomedical data they contain. For the last decade, medical images acquired during routine clinical care often equal the quality of the best research imaging data sets available worldwide. Most importantly, the plethora of clinical scans *vastly* exceeds the scope of research data and includes much larger and more diverse patient populations, including normal cohorts who are imaged to rule out pathological conditions. However, until recently, accessibility to and secondary use of clinical imaging data for research purposes has been severely limited due to complex administrative, legal and technical reasons (Gollub and Turner, 2010). Recently, several projects aiming to facilitate the transition of clinical data to translational research have been initiated. Notably, the Informatics for Integrating Biology and the Bedside (i2b2) project whose software, when implemented in hospitals, can be used to find sets of wanted patients from electronic patient medical record data, while preserving patient privacy through a query tool interface (Murphy et al., 2007 ; Murphy et al., 2010). Similar tools have also been specifically developed in certain institutions, like Partners Research Patient Data Registry (RPDR) tool (see part 3.2.1), enabling affiliated and authorized researchers to identify and access medical information of any patients of interest within the institution. Despite the increasingly widespread use of such tools, neither i2b2 nor RPDR managed medical images. With funds from an American Recovery and Reinvestment Act (ARRA) Administrative Supplement to the Harvard Catalyst, a prototype of the mi2b2 software has been deployed by a devoted team headed by Dr. Randy Gollub and Dr. Shawn Murphy (see People section, part 8.3). This software provides the infrastructure and regulatory policies necessary to locate and retrieve medical images from clinical repositories known as Picture Archiving and Communication Systems (PACS) (Murphy, Marcus et al, 2011)[1] (see part 3.2.3).

Soon that the infrastructure to identify and retrieve medical imaging data will be ready to deliver to affiliated researchers (actually, mi2b2 software is not publically available yet, I had the honor of using the first prototype version). A critical need during this development process was to have a prototype project requiring medical images: this is where Dr. Ellen Grant's work comes in. She proposed that the use of such medical data would be of particular value and interest in regards to Magnetic Resonance (MR) neuroimaging in pediatric patients. In this population, age related changes in MR contrast properties are so dramatic that they compromise interpretation of pathological changes (Sagar and Grant, 2006 ; Rodrigues and Ellen Grant, 2011 ; Utsunomiya, 2011) (see part 2.4). Investigating normal pediatric brain development and extracting quantitative measurements would support radiologists in their ability to detect subtle alterations present in

[1] Murphy SN, Marcus D, et al. New Tools for Integrating Clinical Images into Research Studies. Presented at Society for Imaging Informatics in Medicine. 2011.

conditions such as metabolic disorders or tissue injury (ischemic, necrotic, gliosis, edema). These subtle differences, if detected, would greatly improve radiologists' ability to diagnose such conditions based on clinical imaging, monitor treatment responses and hopefully improve outcomes[2]. Research has shown (see part 2.2) that myelination undergoes dramatic changes in children from birth to two years of age (Deoni et al., 2011). Apparent Diffusion Coefficient (ADC) maps and other Diffusion Weighted Imaging (DWI) indices are reliable measures of water content in the brain and thus an excellent surrogate marker for development of myelination over time within a subject (see part 2.1). Diffusion images, and their corresponding volumetric T1 and isotropic DWI data, are already available from the clinical archives in our institution (MGH) for newborns and young children. These images are of good enough quality for this project. The quantity of data available in MGH PACS is significantly larger than what has been available so far in research due to legal and technical issues in scanning newborns and infants for research purposes only. This highlights the importance and the usefulness of using medical imaging data for our purposes and in this context.

In this project, I am going to demonstrate the feasibility of identifying, retrieving, and analyzing pediatric clinical multimodal MRI data, using the available tools and the prototype mi2b2 software. The aim of this project is to study normal human brain development from birth through 2 years of age as indexed by diffusion MRI.

The overall workflow of this project is displayed in Figure 1. In order to begin this project, I first had to obtain the necessary authorization from the Institutional Review Board (IRB) (see chapter 3.1) to collect MRI data from MGH PACS (2) and corresponding medical information (3). The RPDR tool available at MGH was used in order to request medical information (4) for any patients aged 0-2 years who had a clinical brain MRI scan acquired after the year 2000 (when DWI scans were added to the standard acquisition sequences see part 3.2.1). Mining this medical information, which was retrieved in Microsoft Access database format, enabled me to refine the query and individually identify patients of interest who met the selection criteria (5) (see 3.2.2). MRI data from the selected patients was requested from PACS through the newly developed prototype mi2b2 software (6) (see 3.2.3) and sent through a pipeline (developed by myself) that automatically reformats, organizes, converts (7) and extracts additional information from the newly acquired imaging data (8) (see 3.2.4). From there, MRI images needed for the developmental study were identified according to modality (e.g. diffusion images) and sequence parameters to ensure their comparability (9) (see 3.2.5). Finally, I used diffusion indices (i.e. ADC and FA) to further analyze the selected data in order to investigate age-specific evolution of normal human development from birth through 2 years of age (10) (see 3.3).

I will start this work with a quick overview of the current knowledge in the field of diffusion imaging (part 2.1), its application to brain development (part 2.2 and 2.3) and its clinical relevance (part 2.4). A step-by-step description of the materials and methods outlined above will then be thoroughly presented in part 3 followed by results from database mining and data analysis. Finally, I will finish this thesis by discussing the obtained results before concluding (parts 5 and 6 respectively).

[2] Project's IRB application

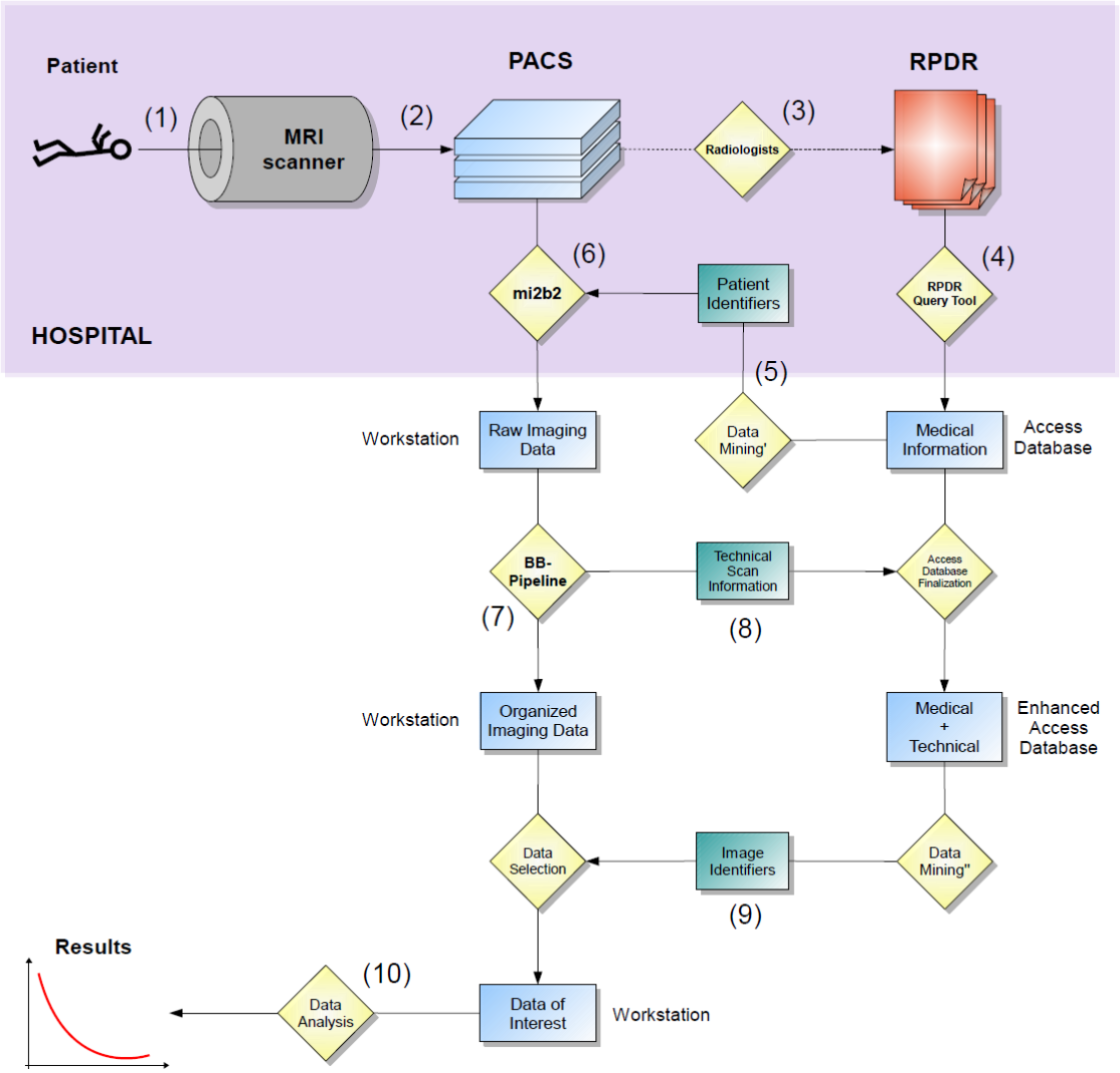


Figure 1: Patients are scanned in the hospital (1) and MRI data are stored in PACS (2). Radiologists access the images and their resulting observations (as well as all medical information regarding patients) are stored and organized in RPDR (3). The RPDR query allows the retrieval of this medical information in Microsoft Access database format (4). This database is used to identify patients meeting certain criteria (5) and the corresponding list of identifiers is sent to mi2b2 software in order to retrieve the related images (6). The BB-pipeline then automatically organizes the data (7) and extracts technical information that is added to the Access database (8). The resulting database is used to precisely identify and select comparable images with modalities of interest (9) that are used for further data analysis (10).

The current project is the culmination of my master's work; it has also been integral to both the mi2b2 project and Harvard Catalyst. The results that I have obtained, as well as the outcomes of troubleshooting the problems that I encountered with the software, have helped the mi2b2 team to improve the utility and the impact of their new software for retrieving clinical data to improve translational investigations. Additionally, my results have been included as preliminary data in a five year RO1 grant application submitted in June 2011 requesting funds to further for the development of mi2b2 and to develop a novel Harvard Catalyst Radiological Decision Support (RDS) Toolkit. Specifically, the grant proposes to develop an extensible framework to support a specific Pediatric MRI module based on neurodevelopmental MRI atlases for structural (T1 and T2) and diffusion tensor imaging (DTI) data from brain scans of patients without neuropathology.

2. Background

2.1. Diffusion Magnetic Resonance Imaging (dMRI)

dMRI is one of many imaging modalities obtainable using a standard MRI scanner. This technique is based on the attenuation of the MR signal due to water motion (diffusion) within different parts of the brain.

In order to provoke this attenuation, an additional gradient, called a Diffusion Weighted (DW) gradient is applied during a typical gradient or spin-echo pulse sequence (see Figure 2, taken from Mori and Zhang, 2006).

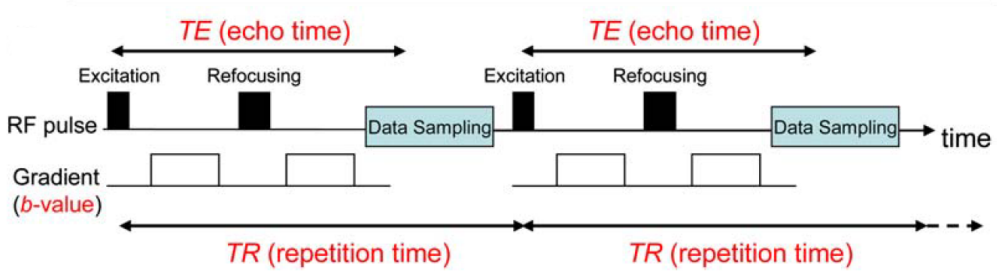


Figure 2: Standard spin-echo sequence with the additional gradient (the other gradients along the x, y and z directions for spatial encoding are omitted for simplification). The b-value is an experimental parameter which depends on the length, height and timing of the DW gradient.

Water molecules that diffuse along the direction of this DW gradient within the time between the excitation and the data sampling induce a loss in phase coherence, leading to the attenuation of the signal (see Figure 3, taken from Mori and Zhang, 2006). As a result, the more water molecules that are able to diffuse along the direction of the gradient, the more phase coherence loss will be induced, and the more signal loss will be observed.

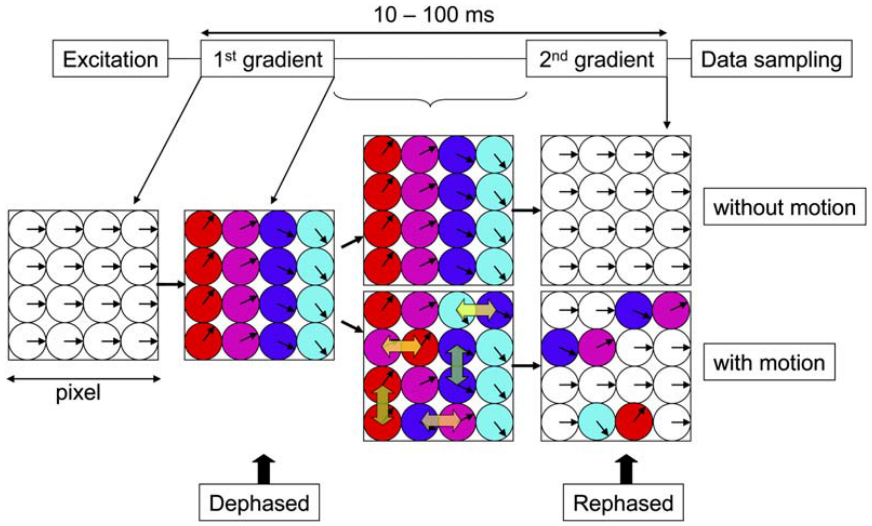


Figure 3: When molecules of water diffuse along the direction of the gradient (horizontal yellow arrows), a loss in phase coherence is noticed after the rephasing step.

For instance, the MR signal of a voxel containing white matter tracts that are parallel to the DW gradient will be lower than the signal of voxel containing white matter tracts running perpendicular to the gradient. In fact, in the former case, water molecules can easily diffuse in the direction of the gradient. However in the latter, water diffusion in the direction of the gradient will be limited by the cell membrane resulting in less signal attenuation (see Figure 4, taken from Hagmann et al., 2006).

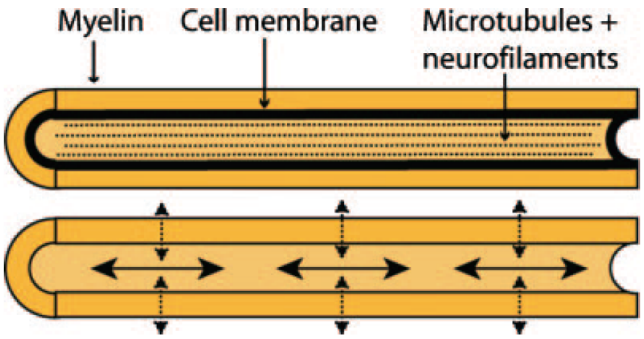


Figure 4: Myelin, cell membrane and microtubules/neurofilaments restrict water diffusion in perpendicular directions leading to less signal attenuation.

The attenuation of the signal can be modeled by the following equation:

$$S = S_0 e^{-bD} \tag{1}$$

where S is the DW signal, S_0 is the signal without any DW gradient, b is the parameters of the DW gradient and D is the ADC.

D corresponds to the diffusivity of water molecules along the direction of the DW gradient, thus D is the value we want to obtain. Since we have one only equation but two unknowns (S_0 and D) for one pulse sequence with one b value, we need a second pulse sequence with different b-value in order to calculate D which is constant across sequences (D is an intrinsic property of the brain for a determined volume). Now we have two equations:

$$S_1 = S_0 e^{-b_1 D} \text{ and } S_2 = S_0 e^{-b_2 D} \tag{2) and (3}$$

and by simply dividing equation (2) by equation (3), taking the natural logarithm and rearranging the resulting expression, we can calculate D:

$$D = \frac{\ln(S_1) - \ln(S_2)}{b_2 - b_1} \tag{4}$$

If we calculate this diffusion coefficient at each voxel, we can calculate a map of the diffusion coefficient, the so-called ADC map, in which the intensity of each voxel is proportional to the extent of diffusion (Mori and Zhang, 2006). One of the two images is often referred as the reference image and has in general a lower b-value; this b-value is commonly referred as b0, b-zero or low-b.

At this stage it is worthwhile to emphasize that D gives us information about the diffusivity of water in **one** direction only, the direction of the DW gradient. In order to get a more precise and complete estimation of the water diffusion information, these steps have to be replicated several times changing the direction of the DW gradient each time. Doing so, we can obtain a specific D for each direction. From these D-values (at least six from linearly independent directions) a 3x3 tensor matrix can be estimated using multiple linear least squares methods (Basser et al., 1994) or non-linear modeling (Alexander et al., 2007). This matrix can be represented by a tensor (an ellipsoid) calculating its eigenvectors and the corresponding eigenvalues (λ_1, λ_2, λ_3) which give an indication about the directionality of the diffusion; this method gives rise to Diffusion Tensor Imaging (DTI) (Le Bihan et al., 2001).

From all these values, different calculations can be computed in order to get some specific information about the diffusivity of the water molecules. Fractional Anisotropy (FA) is among the most commonly used calculations:

$$FA = \frac{\sqrt{\left((\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2\right)}}{\sqrt{2\left(\lambda_1^{\,2} + \lambda_2^{\,2} + \lambda_3^{\,2}\right)}} \tag{5}$$

The FA ranges from 0 to 1 which gives an index of the strength of the directionality: 1 being an anisotropic diffusion (water molecules tend to diffuse in one specific direction) and 0 being an isotropic one (water molecules move in any direction indiscriminately, i.e. $\lambda_1 = \lambda_2 = \lambda_3$). Additional values that can be calculated to obtain more information about the diffusion include the Mean Diffusivity (MD, D_{ave}, ADC_{ave}, $\bar{D}$ or Trace/3) Axial Diffusivity (AD), and Radial diffusivity (RD) corresponding respectively to the average of the eigenvalues (($\lambda_1 + \lambda_2 + \lambda_3$)/3), the largest eigenvalue (λ_1), and the average of the smaller two eigenvalues (($\lambda_2 + \lambda_3$)/2). These calculated values give an indication of the properties of water diffusion within the brain, which, in turn, reflect underlying biophysical characteristics of the brain (myelination, axonal density, etc.).

Some confusion can emerge using the term ADC; depending on authors, ADC can either make reference to the ADC map (see above) or to MD. To avoid such misunderstandings, I will use MD, D_{ave}, ADC_{ave}, $\bar{D}$ or Trace/3 interchangeably in this thesis, and when referring to images corresponding to diffusion coefficients in only one direction, I will use the term “ADC map”; additionally, D in equation (1) to (4) will be referred to as “ADC” only.

One last point is important to understand. We saw earlier that D in equation (1), calculated in equation (4) is constant across sequences because it is an intrinsic property of the brain for a determined volume, namely, the properties of water diffusion occurring within a particular voxel. However, evidence from Ogura et al., showed that ADC value measurements actually depend on the choice of b-values by applying different b-values to different control mediums (see Figure 5, taken from Ogura et al., 2011). Ogura et al. also emphasized that choosing long TR and short TE was effective for accurate measurement of ADC (Ogura et al., 2011).

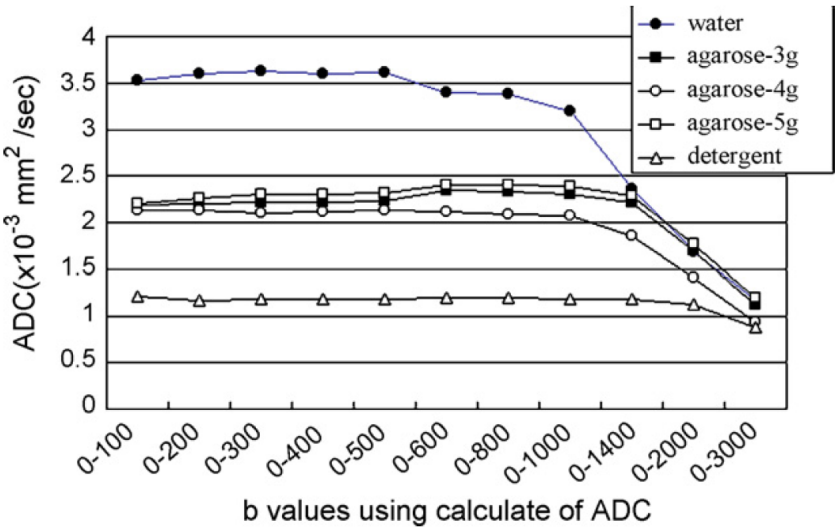


Figure 5: ADC values are calculated for different b-values. Higher b-values have more influence on ADC than lower b-values

Thus, this issue has to be taken into account when comparing data acquired with different parameters, and this will be of special concern when selecting our diffusion data of interest (see part 3.2.5).

2.2. Diffusion MRI of brain development

As we have seen, diffusion imaging is sensitive to the motion of water molecules within the brain, called diffusivity. Diffusivity depends on the internal structure and composition of the brain, such as axonal organization or myelination, and thus makes this modality especially useful for exploring maturation of the brain. Indeed, certain diffusion indices seem to be even more sensitive to internal brain changes than conventional MRI (T1- or T2-weighted imaging) in the context of either maturation processes or brain damage (Hüppi and Dubois, 2006). For example, an early study by Miller et al. emphasized the advantage of using ADC and diffusion anisotropy over the conventional MRI techniques, showing that diffusion imaging could be more objective and sensitive to detect subtle developmental changes. However, their study was based on only a few individual subjects from 26 weeks to 6 years of age (Miller et al., 2003).

The combined improvements in diffusion image acquisition parameters and MRI scanner hardware during this last decade have enabled more sensitive image data to be collected in less time and have allowed researchers to envision using diffusion MRI more systematically in developmental research. As this modality have become more reliable, available and standardized, an increasing number of research groups have started to investigate brain development in terms of the evolution of diffusion indices (mainly MD, FA and eigenvalues) in different Regions of Interest (ROIs). However, age ranges, numbers of subjects, diffusion indices, and experimental protocols vary greatly across studies. For example, one study established ADC values in normal fetal brains (22-35 weeks) *in utero* in 15 fetuses (Righini et al., 2003), whereas another study attempted to create an MRI/clinical/behavioral database from 500 children aged 7 days to 18 years using several

modalities and indices (structural/volumetric analysis, MR spectroscopy, DTI among others) (Evans and , 2006 ; Almli et al., 2007).

Most of the studies have tried to establish a time-related evolution of different diffusion indices throughout life with different models. Westlye et al. aimed to outline age trajectories of DTI indices within 8-85 years in 430 healthy subjects using nonparametric regressions. The study concluded that after accelerated changes during childhood and early adulthood, DTI indices reach a plateau in the early 30s, followed by a slow change until middle-age and ended by a rapid change in the latest part of life (Westlye et al., 2010). This same age related pattern of change was evident across all their ROIs. Other groups restricted their studies to more specific age ranges using different models. The group of Asato et al., for example, investigated 114 subjects ranging from 8 to 28 years in order to characterize specific white matter integrity changes during adolescence using DTI; RD seemed to decrease across different age groups (from childhood to adolescence, and from adolescence to young adulthood) in distinct brain regions (Asato et al., 2010). Snook et al. also focused their attention on the regional changes in the maturation of the brain from childhood (8-13 years) to young adulthood (21-27 years) in 60 subjects. They highlighted either increases or decreases in MD or FA indices depending on the brain region (see Figure 6 for an example of negative correlation trends of MD in 6 different ROIs, taken from Snook et al., 2005). Their findings suggested that the microstructural development of the brain persists throughout adolescence, something that would have been much more difficult to observe with conventional T1-weighted MRI (Snook et al., 2005).

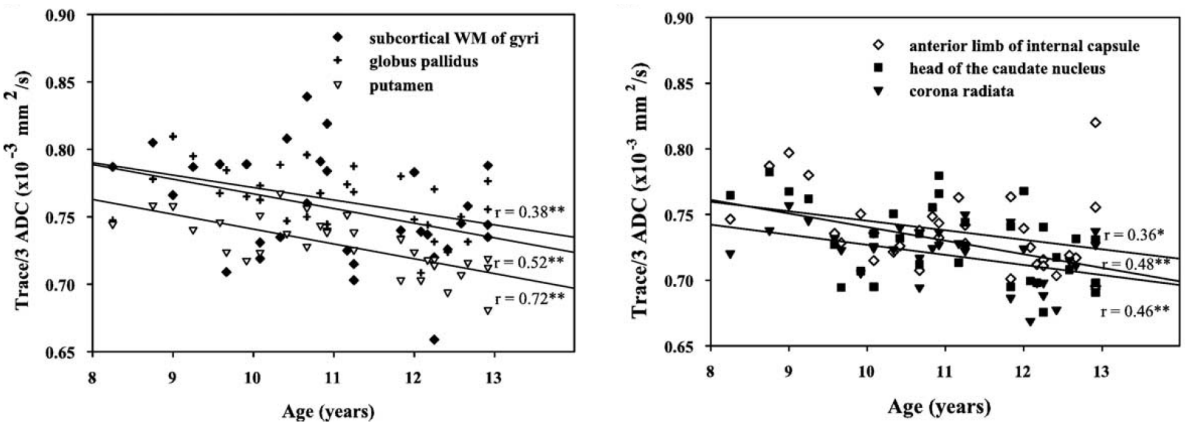


Figure 6: Correlation plot between the ages and the MD values in 6 different ROIs (Trace/3 ADC is equivalent to the MD).

Three other papers deserve to be mentioned. The early study by Murkherjee et al. tracked the time course of D_{ave} and FA values in 153 subjects (age range, 1 day to 11 years) in different ROIs and observed a biexponential decay of D_{ave} (see Figure 7, taken from Mukherjee et al., 2001) and a more complex increase in FA with age in gray and white matter (Mukherjee et al., 2001).

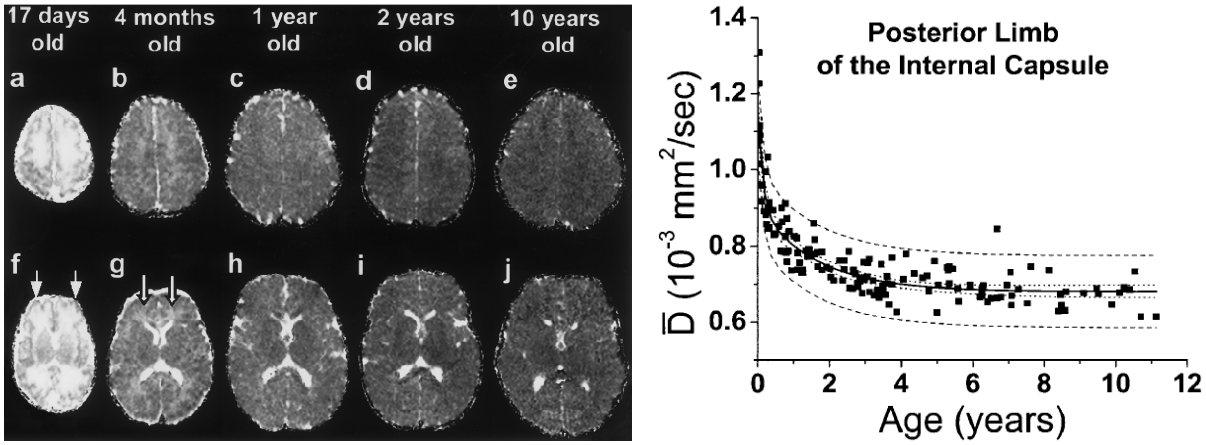


Figure 7: Transverse MR images in five children 17 days to 10 years old at 2 different levels (left picture) and the time course plot of D_{ave} in the Posterior Limb of the Internal Capsule (PLIC). The decay of D_{ave} in early developmental stage can clearly be noticed in both figures (e.g. change in brightness from very bright to dark in left picture).

Later work by Hermoye et al. in a more restricted age range (23 patients, aged 0-54 months) obtained similar results (see Figure 8A, taken from Hermoye et al., 2006). Once again, these results confirm that diffusion imaging is of special relevance when studying brain development in the early stages of life, since its different indices (D_{ave} and FA) reveal the greatest changes within this period. Subsequently, Löbel et al. observed similar changes for ADC and FA investigating 72 patients aged 3 weeks to 19 years retrospectively. They found that logarithmic functions best described the data (see Figure 8B taken from Löbel et al., 2009).

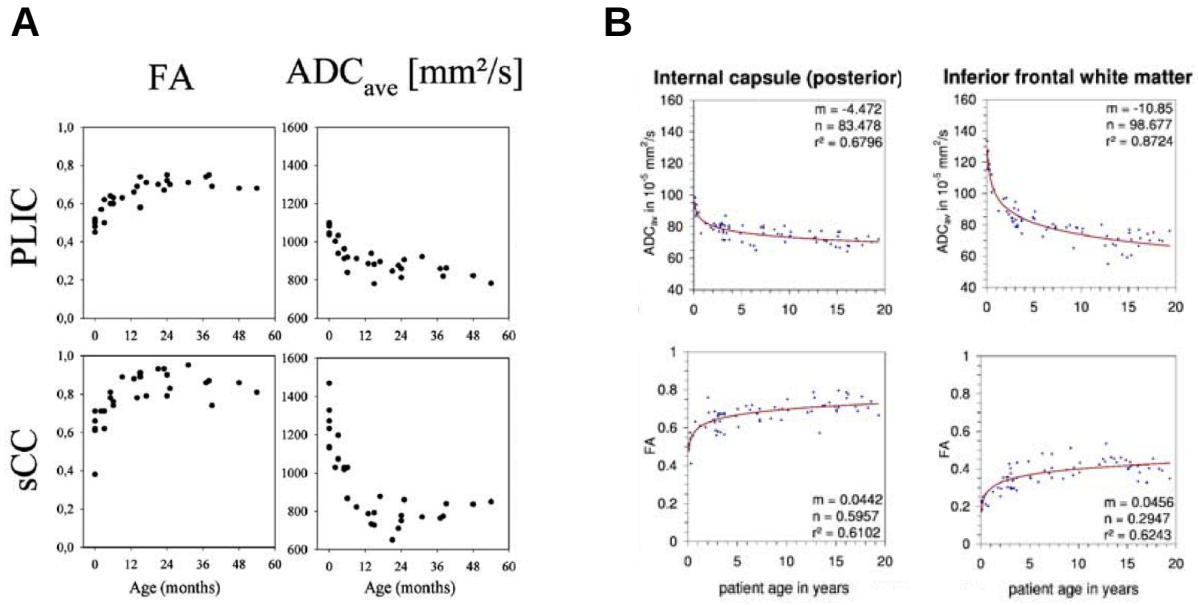


Figure 8: Time courses for FA and D_{ave} in the Posterior Limb of the Internal Capsule (PLIC) and the splenium of the Corpus Callosum (sCC) as shown by Hermoye et al. (2006) (A). Similar time course in Löbel et al. (2009) papers for the PLIC and the inferior frontal white matter (B); note the logarithmic fitting curve in red. Comparing the D_{ave} time evolution with the one of Figure 7, we note a similar pattern for the PLIC ROI.

Finally, it is important to remember that these diffusion indices only give information about the diffusivity of water molecules within the brain, and even though diffusivity depends on the underlying tissue structure (myelination, axonal density, etc.), we should be careful not to interpret the results as direct measurements of underlying biophysical properties (Wheeler-Kingshott and Cercignani, 2009) as may be the case in some studies.

2.3. Brain atlases of healthy babies and children

Before considering methods for the creation of atlases for diffusion indices, crucial pre-processing image analysis steps must be addressed. One of them, brain normalization, is of key importance when investigating different subjects within a cohort. In fact, by observing raw images from different subjects, inter-individual differences in terms of brain size, shape or internal structure are evident even to the eye. These individual differences render the data challenging to use to make meaningful comparisons. The normalization step consists of transforming the MRI data from an individual subject to match the spatial properties of a standardized image, such as an averaged brain derived from a sample of many individuals[3].

In this context, research groups have tended to use common normalization schemes called stereotaxic spaces in order to facilitate inter-laboratory communications. The most widely used stereotaxic spaces are the Talairach space (derived from a single brain of an elderly woman) or the Montreal Neurological Institute (MNI) space (derived from the average of MRI structural images from more than 100 hundred subjects). However, most of the tools used to transform brains in a standard frame have only been developed to process adult brains and thus cannot directly be used with data from babies or children.

Recently, different groups have tried to tackle this problem in order to be able to investigate the brains of babies and children in a more robust way. In 2002, Wilke et al. compared the linear scaling parameters and the deformations from the non-linear spatial normalization obtained from both a standard adult and a custom pediatric template with T1-weighted images, pointing out that caution should be used when describing functional activations in children on the basis of adult data such as Talairach coordinates (Wilke et al., 2002). The next year, Wilke et al. came to the same conclusion comparing their newly created pediatric templates and *a priori* brain tissue data from 148 healthy children (age range 5-19 years) with standard adult data available within SPM software[4] (SPM99): concluding again that caution should be used when analyzing pediatric brain data using adult *a priori* information (Wilke et al., 2003).

Further investigation by Kazemi et al. was done on younger subjects using an updated version of SPM (SPM2). Seven newborns (gestational age range 39-42 weeks) were used to create a neonatal atlas template. The following is a summary of their protocol used to generate a brain template of newborns. They first selected a reference image and positioned it in a standard way (anterior commissure at the origin of the line connecting the anterior with the posterior commissure in the horizontal plane). Subsequently, an affine registration was applied to all the other images in the data set with respect to the reference image in order to correct the position and global shape

[3] Huettel, Song & Mcarthy (2009) Functional Magnetic Resonance Imaging, Sinauer Associates (Sunderland, MA)
[4] Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK.

differences. These registered images were then nonlinearly normalized to the reference image[5] and the average deformation was applied to them. At this point, the template was generated by averaging these newly obtained images. The last step was to use this generated template as the new reference and the entire process was replicated in order to reduce the bias induced by the first reference image. They came to the conclusion that using a data set-specific template for alignment of neonatal images taken from this data set resulted in an improved normalization process, compared with the results obtained with the use of an *a priori* adult template or even with the use of an *a priori* pediatric template such as the Cincinnati Children's Hospital Medical center (CCHMC) Pediatric Brain Template (Kazemi et al., 2007). In 2008, Altaye et al. completed the same kind of study with different processing steps and used a larger cohort of 76 infants ranging in age from 9 to 15 months using SPM5. They used their results not only for normalization, but also for segmentation of the infant brain (Altaye et al., 2008). Meanwhile, the same group developed a toolbox for creating customized pediatric templates called "Template-O-Matic" for the SPM5 image data processing suite. They came to the conclusion that their tool was of special utility for customized reference data generation and for image processing of an "unusual sample", such as children or elderly subjects (Wilke et al., 2008).

All of the works described above were done on T1-weighted images and were not applied to DWI or DTI. To date, I have not come across any papers working on the creation of diffusion atlases for babies/children that simultaneously assess the efficiency of the method used to generate the template, as was the case in the studies presented above. Current diffusion studies (Bartha et al., 2007 ; Faria et al., 2010) have been more interested in investigating the time course of different diffusion indices (see 2.2) than in the generation of a DWI/DTI template. In order to normalize their data, they used already available tools such as the Automated Image Registration (AIR) software with an *a priori* reference template such as the ICBM-DTI-81 template, a template that is based on probabilistic tensor maps obtained from 81 normal adult subjects acquired under an initiative of the International Consortium of Brain Mapping (ICBM)[6]. As indicated in the previous paragraph, this might not be the best solution since the *a priori* data set used to normalize the brains of babies/children in these studies originated from adult subjects.

2.4. Clinical applications of diffusion imaging

Diffusion imaging is not only of special interest and sensitivity for detecting and observing subtle changes during brain development, as indicated above, but also in detecting and observing general changes within the brain. This means that changes caused by strokes, traumatic brain injuries, abnormal development or diseases affecting brain integrity could also be detected with the help of diffusion measurements. Actually, these different measurements are currently used in Radiology departments to detect brain abnormalities in adults as well as in babies and children. I had the chance to spend one afternoon with Dr. Grant in within the Pediatric Radiology Center at Children's Hospital Boston to observe how they analyze the medical imaging data to arrive at a diagnosis. I can confirm that diffusion data are used for diagnostic purposes. Currently, this modality is used as a complement to other modalities (e.g. T1 and T2 weighted MRI).

[5] Using the method presented by (Ashburner and Friston, 1999)
[6] http://www.loni.ucla.edu/Atlases/Atlas_Detail.jsp?atlas_id=15

It seems that diffusion imaging could possibly be a more appropriate tool to detect early effects of brain injuries than conventional MRI since water diffusion changes are an early indicator of cellular injury. This could be of particular importance in infants in the context of administration of neuroprotective therapies (Hüppi and Dubois, 2006). In their 2006 review, Pallavi Sagar and Ellen Grant described the known pathophysiologic processes linked with changes in ADCs (increase or decrease) and described the pattern and time course of DWI and ADC findings in a number of pediatric disorders (Sagar and Grant, 2006). In their 2011 papers, Katyucia Rodrigues and Ellen Grant emphasized the importance of using DWI and ADC images concurrently with conventional MR images in order to facilitate an accurate diagnosis. This is even more relevant for immature brains where incomplete myelination can render injury detection more difficult if only screened on T2-weighted images (Rodrigues and Ellen Grant, 2011).

Unpublished notes from Dr. Ellen Grant point out some limitations on visually detecting brain injuries in early development without quantitative normative data for comparison. For example, *gestalt* visual diagnosis of brain injury is difficult in the developing brain due to the rapidly changing appearance of normal, making it difficult for even experienced neuro-radiologists to detect subtle abnormalities when relying on visual interpretation alone. In the case of neonatal hypoxic ischemic brain injury, subtle injuries can be missed on DTI because areas prone to injury normally have lower diffusivity than adjacent areas due to developing myelination (Figure 9 a, d). Injury causes a further decrease in diffusivity (Figure 9 e) and therefore knowing when the diffusivity is too low can be difficult and can lead to variability in clinical reads.

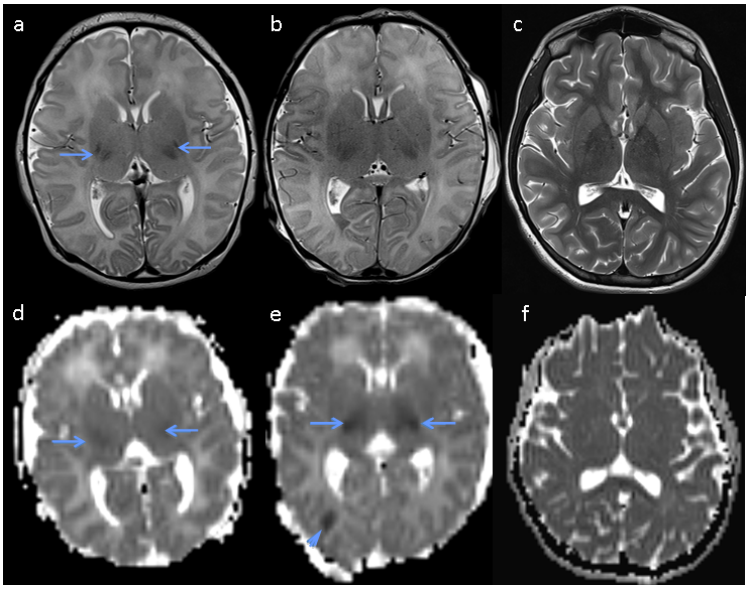


Figure 9: Top Row T2 weighted images. Bottom row Apparent Diffusion Coefficient (ADC) Maps. Normal neonates have lower T2 signal and ADC in regions that are undergoing myelination (arrows in a, d respectively). Neonates with hypoxic Ischemic Injury have lower ADC in the same regions (arrows in e). In normal older children with fully myelinated brain regions (c, f) the ADC image is homogeneous (f) as compared to the normal neonate (d) making areas of abnormally decreased ADC easier to detect in older children because fully myelinated brain regions have uniform diffusivity. Therefore injuries are more easily detected by visual inspection alone as they appear as deviations on a homogeneous background.[7]

[7] Unpublished figure from Dr. Ellen Grant

However, when quantitative DTI data is compared between normal and those diagnosed clinically with hypoxic ischemic injury, significant differences are obtained and these differences correlate with outcome providing potential prognostic significance (Zarifi et al., 2002 ; Vermeulen et al., 2008)".[8]

In summary, diffusion imaging may be very well suited to the investigation of early brain development and to the construction of age-specific markers/atlases for newborns or infant, since the diffusion indices are sensitive to dramatic changes during this period. In addition, the sensitivity of diffusion indices such as MD or FA to detect brain "abnormalities" makes this modality particularly suitable in this context if quantitative normative data can be obtained and used as a support tool to help radiologist in detecting brain injuries in young patients.

[8] Unpublished notes from Dr. Ellen Grant

3. Materials and methods

3.1. Institutional Review Board (IRB) approval

Prior to any research studies on human subjects, official authorization given by an IRB has to be obtained to ensure the protection of the rights and welfare of any research subjects (Hart and Belotto, 2010). In the United States, this is a federal regulation that any research institution conducting human subjects research must comply with. This committee is chosen by each institution and has to follow certain guidelines notably dictated by the Food and Drug Administration (FDA).

Within MGH, the IRB is known as the Partners Human Research Committee (PHRC). "The PHRC must approve all human-subject research conducted by a Partners-affiliated investigator. Human-subject research is a systematic investigation designed to develop or contribute to generalizable knowledge where an investigator obtains data on individuals either through direct intervention/interaction or through the use of identifiable private information (medical records) or specimens"[9]. Under these conditions, the current project falls under IRB approval since data on individuals through the use of identifiable private information needs to be collected.

This project was approved in November 2010 under the title: "Diffusion ADC atlas of normal development in children (ages birth-2 years and 2-20 years)". It is sponsored by the mi2b2 grant from the National Institutes of Health - Center for Research and Resources (NIH-NCRR) as a supplement to Harvard Medical School's Clinical Translational Science Award (the Harvard Catalyst). Our IRB application briefly describes the purpose of the research and its significance, the type of data to be collected (in our case, medical record review including images of any children (aged 0-20 years) who had an MR image data set collected at MGH from 2000 to the present (rolling forward) until IRB expires), the security measures taken to protect patient personal information, and finally the study staff approved to have access to this data. Additionally, everyone on the IRB has to be CITI[10]-certified meaning that they have passed specific training modules related to research on humans and its legal implications.

3.2. Database mining (Medical Informatics)

Retrieving appropriate medical imaging data from hospital databases requires patience and multiple steps through which I was able to gradually refine selection criteria. Each step, from retrieving the scans for a patient of interest, to finding specific MRI series for the proposed research, will be thoroughly described in the following chapters. Working with large amounts of data requires some organization to avoid the risk of drowning in the huge amount of available data. I have completed each phase of the process described below, and I designed each step specifically to facilitate the search for information, automating as many processing steps as possible. Here is a brief step-by-step overview:

[9] http://healthcare.partners.org/phsirb/abouthrc.htm
[10] Collaborative Institutional Training Initiative (CITI), https://www.citiprogram.org/

The first step required the use of the Research Patient Data Registry (RPDR), a database specific to MGH, BWH and several other Partners institutions. RPDR stores all patient information gathered at one of these hospitals. The RPDR web tool allows the request of this information for any patients who fulfill specific predefined criteria (see part 3.2.1).

Patient information was then retrieved in Microsoft Access Database format that enabled the refining of patient criteria selection. At this point, individual patient information was available, allowing the identification of patients of interest whose images needed to be retrieved (3.2.2).

After having designated the patients of interest, I was able to use their Medical Record Number (MRN, unique hospital-specific identifier for each patient) to query the imaging data from the Picture Archiving and Communication System (PACS) of the chosen institution, using the mi2b2 software prototype currently under development (3.2.3).

The results of this query provided information about the images and their corresponding detailed specifications (modalities used, pulse sequences, etc.) that are stored in Digital Imaging and Communications in Medicine (DICOM) format. DICOM is the universal format used in PACS image storage, which contains not only the images, but also technical acquisition information about these images. The raw data retrieved from PACS first had to be reorganized (in terms of file/directory names, directory structure, etc.) in order to facilitate further work on them (3.2.4). At the same time, I automatically extracted MRI image details (sequence parameters, image resolution, etc.) from the DICOM headers and added them to the Microsoft Access Database to complete it with technical information that was not available from the RPDR query (3.2.5).

Once the data were well organized, including all pertinent information, I was able to, in couple of mouse clicks, instantly execute a request such as the following: “Give me the ADC volumes of any patients who were scanned at the age of 30 to 60 days whose corresponding radiology reports don’t mention any major brain abnormality”. This type of request allowed further research to be conducted on specific data modalities from specific populations (3.2.5).

Let’s now have a look into each step in more detail. Every step is described in light of the specific aims of my project, namely to find and retrieve all the data from patients who had an MRI scan of the brain acquired on or after the year 2000, who were aged 0 to 2 years at the time of the scan and whose corresponding radiology reports reveal no major brain abnormality (we limited our query to only those scans collected after the year 2000 when we know from our collaborator, Dr. Ellen Grant, the standard acquisition protocols began to include diffusion weighted images).

3.2.1. RPDR query

RPDR is a database specific to MGH, BWH, and several other Partners institutions that stores all visits, procedures, diagnoses, and patient information gathered at any one of these hospitals. It notably contains more than 5 million Partners Healthcare patients. RPDR allows users to perform IRB approved queries and returns all relevant records (including demographic data, medical records, and accession numbers identifying medical images) for patients who match the specified search criteria (Query Items). It is a Microsoft SQL database sitting on a Windows server that can

be accessed from Partners workstations via the online query tool[11] (see Figure 10 for an explanatory scheme of the query structure and process).

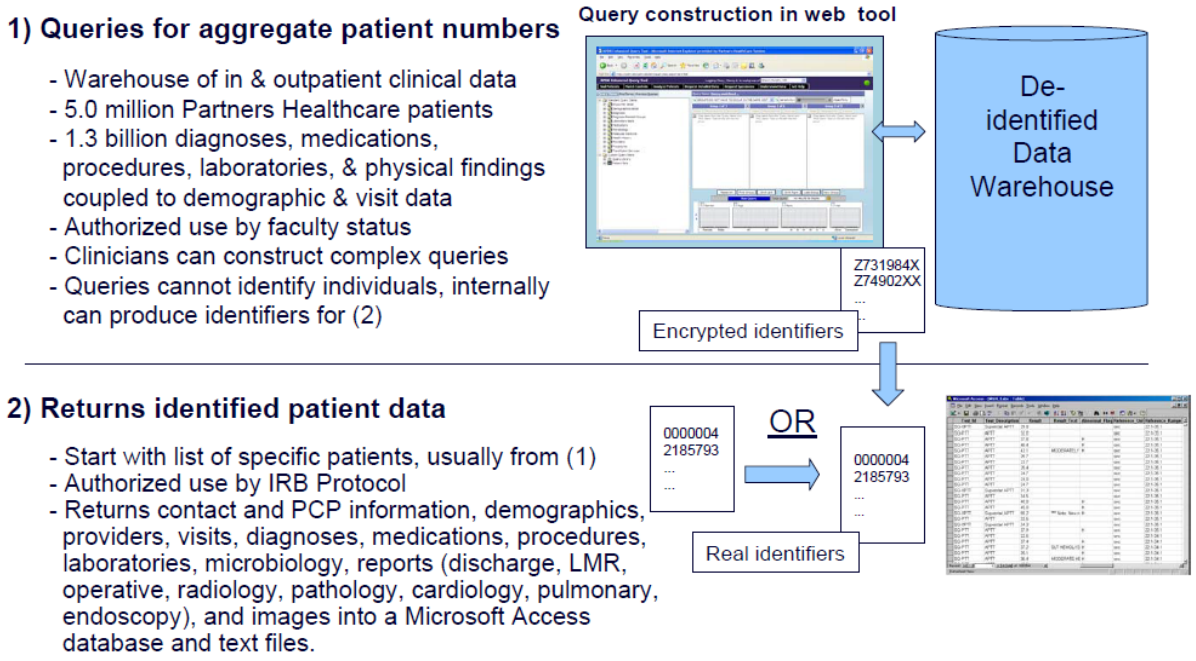


Figure 10: Scheme of RPDR query[12]

The first step consists of constructing a query that allows the user to select a population of interest without accessing any patient identifiers. Therefore this step does not require an IRB approved protocol but nonetheless, it requires being member of an RPDR registered Partners Faculty Sponsored Workgroup (Figure 10.1). At this level, only the aggregate number of patients who fit the selection criteria are available to the user. Other aggregate statistics such as gender, age, and ethnicity are also available.

The second step is the patient information request. Since this includes personal information, it requires an IRB approved protocol in order to be launched. This can be requested either from step 1 or from a pre-defined patient list. All patient information is then sent to the user in different text files in a pre-defined database format that can be handled by Microsoft Access software. Completing this step takes 1 to 3 weeks (Figure 10.2).

The RPDR Query Tool web interface is shown in Figure 11. Any *Query Items*, such as encounter details, demographics (Age, Country, Gender, etc.), diagnosis or procedures, can be found in the left column of the page either by looking for the items of interest going through the folder hierarchy or by using the *Find Terms* tool provided that performs a keyword search.

[11] http://www.partners.org/rescomputing/template.asp?pageid=99&ArticleTitle=RPDR&level1ID=9&tocID=9&articleSubPage=true

[12] Shawn Murphy MD, Ph.D., "Research Patient Data Registry (RPDR) at Partners Healthcare" presentation, January 31, 2011

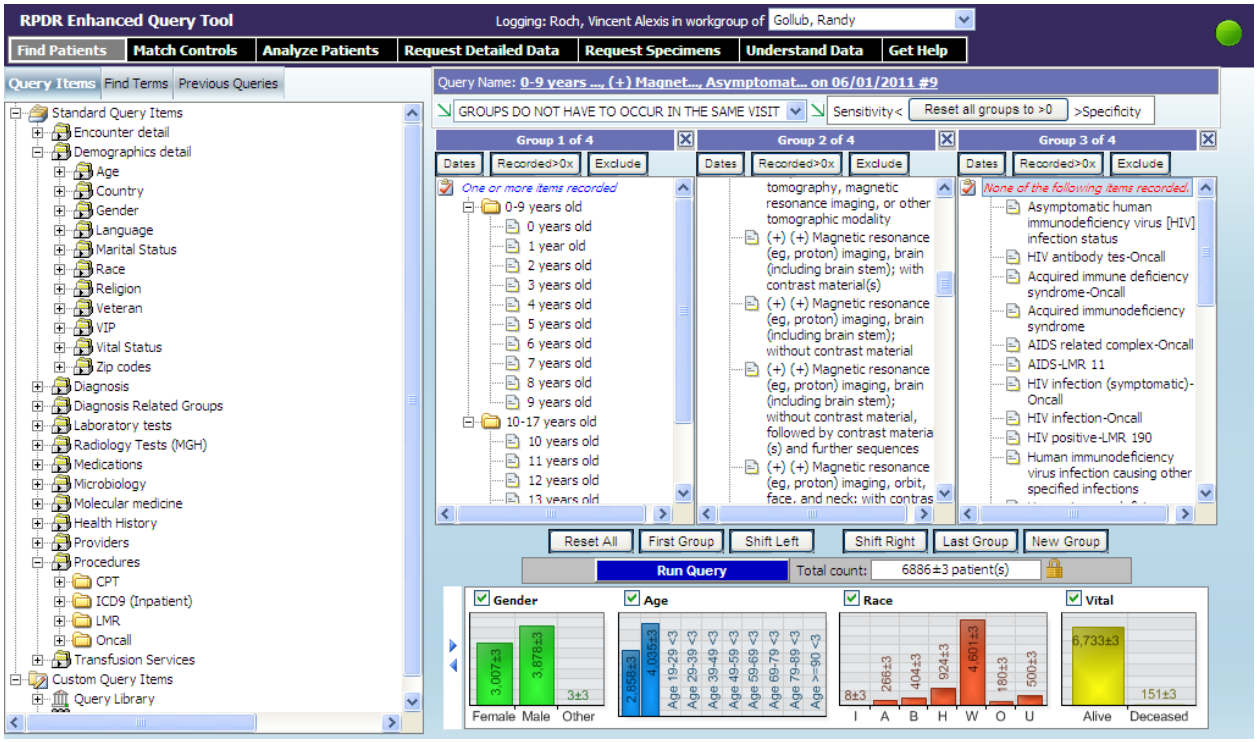


Figure 11: RPDR Enhanced Query Tool web interface

When the desired item (criteria) is found, it is dragged to one of the columns that act like logical operators, either to the column where items are being linked by an **OR** operator or to the column linked by an **AND** operator as appropriate. Some other options are also available for each column such as *Dates* and *Exclude*. The former is used to restrict the dates at which the items in this column were performed and the latter excluding any patients who are related to the items included in this column (acting as an **AND NOT** logical operator).

In this project, we wanted to obtain all the medical imaging data acquired at MGH since the year 2000 from any patients who had a brain MRI scan when they were 0 to 2 years old. In reality, I queried patients from 0 to 6 years old for the purpose of further investigation that will be done by others after I have completed this master project. The RPDR query tool has limited criteria and for example, there is no option to select the age of the patients *at the time of the scan*, but only their current age. Thus, to be certain not to miss any data, I selected a broader age range (0-17 years) in order to get at least all the data from 2000 of patients who could have been 0 to 6 years old at the time of the scan. From this, data from patients older than 6 years at the time of the scan were naturally acquired, but I was able to discard them during the next step (see part 3.2.2).

For the first criterion, Age*, the* items from *0* to *17 years old* were simply dragged to the first column. This indicated that all the patients whose ages were 0 or 1 or 2 or … or 17 years old were included in the search.

The next criterion included only patients who had a brain MRI scan; this was not as straightforward as it appears. Indeed, this information has to be retrieved from the procedure description field of the records, which is not standardized and thus makes the criteria selection difficult. For example, MRI scan can be found under the following descriptions: "*Brain imaging, complete study; static*", "*Cerebral scan*" or "*Magnetic resonance (eg, proton) imaging, brain (including brain stem); with contrast agent*" among others. Thus, any procedures which included

keywords related to brain MRI such as “MRI”, “Magnetic” or “Brain” using the *Find Terms* tool were manually selected and dragged to the second column (for the complete list of selected procedures items, see Appendix 8.4). For this criterion (column), the *Dates* option was used to restrain the query only to patients whose procedures were performed after 2000.

In addition, we wanted to avoid patients with a history of human immunodeficiency virus (HIV), thus any patients with any diagnosis related to HIV were discarded (for the complete list of selected diagnosis items, see Appendix 8.5). Similarly, I found HIV related diagnosis using keywords and the *Find Terms* tool, dragged them to the third column and used the *Exclude* option this time to remove patients with HIV.

Finally, MGH criterion was included in the fourth column to only retrieve patients imaged at MGH.

In Boolean operators, the query criteria looked like this: “age from 0 to 17” **AND** “procedures field contain any term related to brain MRI” **AND NOT** “HIV diagnosis history” **AND** “MGH patients”.

At this point, the query was run and the aggregate number of patients who met the selection criteria was displayed in the user interface as well as the other statistics (see results in part 4.1.1).

The related identified patient data could then be requested (Figure 10, 2) and the query was reviewed to verify its compliance with the corresponding IRB approved protocol. If each step had been completed correctly, patient information was sent to the user.

3.2.2. Microsoft Access Database queries

From the RPDR query requested in the previous chapter (part 3.2.1), identified patient data were received in a pre-defined database format that can be handled by Microsoft Access software. It is a relational database (formatted for Microsoft Access) which is organized into different sets of tables, each containing a key or unique relationship that allows linking between tables.

In this case, the database contained all the medical information (ranging from the dates of birth to physicians' hand written notes) concerning the patients. This information is organized into the following different tables:

- the *Demographics* table contains the following fields for each patient:

  - **EMPI** (Enterprise Master Patient Index number): unique identifier for each patient across the different hospitals within Partners
  - **MRN** (Medical Record Number): unique identifiers for each patient within one particular hospital (e.g. within MGH)
  - **MRN_Type**: identifier of the institution associated with the MRN (e.g. MGH)
  - **Gender**
  - **Date_Of_Birth**
  - **Age**
  - **Language**
  - and other information related to demographic data (e.g. race, marital status, etc.)

- the *Procedures* table:

  - **EMPI**, **MRN**, **MRN_Type**
  - **Date** (of procedure)
  - **Procedure_Name**
  - **Code_Type** and **Code**: standardized system of charge codes for clinical procedures
  - **Encounter_Number**: unique identifier of the record/visit
  - and other information related to the procedure (e.g. the provider, the clinic, etc.)

- the *Encounter* table:

  - **EMPI**, **MRN**, **MRN_Type**
  - **Admit_Date**
  - **Encounter_Number**
  - **Principal_Diagnosis** : condition established after study to be chiefly responsible for the admission of the patient to the hospital for care
  - and other information related to the visit/record (e.g. the physician making the diagnosis, etc.)

- the *Radiology* table:

  - **EMPI**, **MRN**, **MRN_Type**
  - **Report_Number**
  - **Report_Date_Time**
  - **Report_Text**: containing information about the scan (written by the radiologist). It notably includes technical details and medical descriptions about the images as well as supposed diagnosis.
  - and other information related to the radiology visit/record.

- Other tables containing additional information: for a complete list of all the tables available, see Appendix 8.6. (Details for each of these tables are available on the Partners RPDR website[13])

As listed above, all of the tables have several fields in common, thus allowing different tables to be linked in order to retrieve the information of interest. For example, if we are looking for the age of the patients at the time of their procedures: we need the date of birth of the patients (in Demographics table) and the date of the procedures (in Procedures table). By linking the two tables by their MRN and using a predefined function available in Microsoft Access software, we obtain the time elapsed between two dates, namely the date of birth and the date of the procedure. By performing such a query, we obtain a new table containing any fields we want from Demographics with any corresponding fields from Procedures for each patient. In Structured Query Language (SQL), this looks like the following expression:

[13] http://rpdrweb/partners/datainfo/datainfo.htm (only accessible through Partners network)

```
"SELECT Demographics.EMPI, Demographics.MRN_Type, Demographics.MRN,
Demographics.Date_Of_Birth, Procedures.Date, Procedures.Procedure_Name, Procedures.Code_Type,
Procedures.Code, Procedures.Encounter_Number,
DateDiff("m",Demographics.Date_Of_Birth,Procedures.Date) AS Age
FROM Demographics INNER JOIN Procedures ON Demographics.MRN = Procedures.MRN;"
```

meaning that we want the query to return a table containing the following fields (cf. "SELECT") for each patient: EMPI, MRN Type, MRN from Demographics (or from *Procedures* since they have these fields in common), the date of Birth from *Demographics*, the procedure date, name and the corresponding encounter number from *Procedures*, and the Age of the patient (cf. "AS Age") at the time of the procedure using the function "DateDiff" which returns in months ("m") the time elapsed between the two dates. The last line defines the link between the two tables: in that case, we link Demographics and Procedures by their common MRN field ("Demographics.MRN = Procedures.MRN") with an "inner join" meaning that the resulting table will only include information (rows) where the joined fields (MRN in our case) from both tables are equal.

The Microsoft Access software (also called "pseudo-relational database management system") also allows us to filter the resulting table with certain criteria. We can for instance apply logical operators (=,>, <) in order to select patients in a specific age range at the time of their procedures. We can also use text operators such as "like" in order to filter data containing specific terms. Combining any of these methods together allows the construction of more complex queries satisfying all needs.

In this project, I wanted to first determine and identify how many patients in each age range had a brain MRI scan (e.g. how many patients had a MRI scan when they were 12 months old) (see part 3.2.2.1) and then more specifically, how many of them had a diffusion MRI (3.2.2.2). The distribution of the scans according to the year in which they were completed was also investigated (3.2.2.3). Moreover, we were interested in the health history of the patients, namely whether they had been diagnosed as having any particular diseases. This is an important step since we only wanted to include patients with "normal" brain imaging data (i.e. no diagnosed brain abnormalities in their clinical reports) in the context of this project (3.2.2.4). In addition, longitudinal information about the patients was also investigated in order to know if a particular patient had had other MRI scans at a different age. This information could be very valuable for investigations about early brain development or time course of disease evolution (3.2.2.5). Finally, I searched for brain MRI scans having corresponding Computed Tomography (CT) scans, since they could be potentially valuable for subsequent investigations (3.2.2.5). All the results related to these specific queries can be found in the results chapter (see part 4.1.2).

3.2.2.1. Number of brain MRI scans for each age range

Given the querying methods described above, I could start to complete the first query where I wanted to know how many patients from 0 to 6 years old (0 to 72 months old) had had a brain MRI scan for each age range (in months). The first step in this query was to determine the patients who underwent a brain MRI scan. As mentioned in part 3.2.1, the information "brain MRI scan" in the *Procedure* table can be found in different forms. When I completed the online RPDR query, I actually included 51 different types of procedures as inclusion criteria that could potentially

include brain MRI (see Appendix 8.4). Within the Access database, all these different procedures can be found under the Procedures_Name field of the *Procedure* table or under the corresponding Code field (see Appendix 8.4). After having explored the information related to these 51 procedures in the database and under the direction of Dr. Ellen Grant, I refined the query to 12 different procedures that are the most likely to include a brain MRI (see Table 1).

Code	Procedure_Name
70552	Magnetic resonance (eg, proton) imaging, brain (including brain stem); with contrast material(s)
70551	Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material
70553	Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material, followed by contrast material(s) and further sequences
70542	Magnetic resonance (eg, proton) imaging, orbit, face, and neck; with contrast material(s)
70540	Magnetic resonance (eg, proton) imaging, orbit, face, and neck; without contrast material(s)
70543	Magnetic resonance (eg, proton) imaging, orbit, face, and neck; without contrast material(s), followed by contrast material(s) and further sequences
70541	Magnetic resonance angiography, head and/or neck with or without contrast material(s)
70545	Magnetic resonance angiography, head; with contrast material(s)
70544	Magnetic resonance angiography, head; without contrast material(s)
70546	Magnetic resonance angiography, head; without contrast material(s), followed by contrast material(s) and further sequences
88.91	Magnetic resonance imaging of brain and brain stem
76390	Magnetic resonance spectroscopy

Table 1: List of the 12 procedures that are the most likely to include brain MRI

In order to obtain the age of the patients at the time of these procedures, I had to link the *Demographics* and *Procedures* tables by the MRN and include the "Datediff" function to display the age, adding two selection criteria, one for the age (>= 0 and <=72 months) and the other for procedures codes (i.e. only procedures containing one of the 12 codes listed just above were included).

I realized afterward that the "Datediff" function didn't return the age in months in the sense I understood it ("real age" of the patient according to the number of days between the date of birth and the date of the procedure). In fact this function simply calculates the difference between the month of birth and the month of procedure meaning that a person who was born on 11/01/2000 and had an MRI scan on 12/31/2000 (so aged 12-11 = 1 month) would have been the same age at someone who was born on 11/30/2000 and had an MRI scan on 12/01/2000 even though the former was 60 days old and the later 1 day old and thus didn't have the same "real age" in months. To avoid any further source of confusion, I always used ages in days and consider age in months according the time elapse in days (0 month old being aged from 0 to 29 days, 1 month old from 30 to 60 days, 2 months old from 61 to 90 days and so on); the age in days can be retrieved from Microsoft Access using the same "Datediff" function but replacing the "m" (month) option by the "d" (day) option.

3.2.2.2. Proportion of MRIs that include diffusion imaging

The next step was to refine the query by specifying the proportion of these scans supposedly containing diffusion MRI data. Two major difficulties emerged: first, no specific fields in *Procedures*

mentioned whether the procedures include diffusion data. This information has to be retrieved from the *Radiology* table under the Report_Text field stated in a non-standardized manner by physicians. As a result, I had to filter the radiology reports not according to standardized codes, but according to keywords such as “diffusion”, “DWI”, “ADC”, “DTI” or “diffusivity”. In other words, the patients whose Report_Text fields contained at least one of the previous diffusion-related terms were included. The second problem was that the Report_Text fields from *Radiology* could not be directly linked to their corresponding procedures by a common encounter number for instance. I had to link them indirectly according to their dates: for each patient brain MRI scan (procedure) found in the previous query, I looked for a corresponding Radiology_Text written within 2 days of the procedure for the same patient

3.2.2.3. How old are the scans?

A specific query was also constructed in order to know the distribution of the brain MRI scans according to the date (in years) when the scans were executed. This information was useful when selecting the most interesting scans for the project because the quality and validity of our results partly depended on their similarities in term of modalities and MRI sequence parameters. Hence, the closer in time the scans were run, the more likely they would be “technically” comparable.

3.2.2.4. “Normal” or “Abnormal” brain MRI scan?

The next step was to investigate more precisely the health history of the patients and their related diagnoses. The purpose of this step was to identify patients whose brain MRI scan seemed to be “normal” (i.e. without apparent major abnormalities in their brain images). This information can be retrieved, at least partially, under two different fields: Report_Text field in *Radiology* and Principal_Diagnosis field in *Encounter*. Report_text fields notably contain written reports of MRI data where physicians assess health states of patients describing their brain images. Let’s have a look at the following Report_text taken from a patient (exam number, Date/Time and names had been removed for confidentiality reasons):

Exam Number: Report Status: Final
Type: MRI BRAIN -
Date/Time:
Exam Code: MRBRN/SED
Ordering Provider:

HISTORY:
eval for structural abnormality
PEDI SEIZURE PROTOCOL, SUSCEPTIBILITY. JH....

DIAGNOSIS:
new onset seizures

REPORT:
This study was reviewed with Dr. X

History of seizure like episodes with clenching of teeth and crossing legs.
Brain MRI without gadolinium enhancement according to pediatric seizure protocol and susceptibility axial images. There are no prior studies available for comparison.
There is no evidence of abnormal susceptibility artifact suggest blood products or calcification.
The brain is normal in signal intensity and morphology. There is no evidence of cortical dysplasia or gray matter heterotopia.
There is no evidence of an infarct by diffusion weighted imaging.
There is no evidence of intracranial hemorrhage. The size of the sulci, ventricles, and cisterns are age appropriate. There is no extra-axial fluid collections. The visualized soft tissues are unremarkable. The globes are unremarkable. There is mild mucosal thickening within the maxillary sinuses.

IMPRESSION
Unremarkable brain MRI with no identifiable structural abnormality..

RADIOLOGISTS: SIGNATURES:

Even though this patient was admitted for a "new onset seizures" diagnosis, his overall brain imaging data seem to be normal, and thus could be included as a "normal" subject in our cohort study. Let's take another example:

Exam Number: Report Status: Final
Type: MRI,Brain W/O
Date/Time:
Exam Code: 580/NE
Ordering Provider:

HISTORY:
33 WK PREEMIE WITH GRADE 4 IVH, DEPRESSED NEUROLOGICAL STATUS ALSO INTUBATED.

REPORT:
This examination was reviewed with Dr. X
An MRI of the brain was performed including sagittal T1 weighted images, axial T1, FLAIR, T2 and diffusion weighted images.
Comparison is to the CT scan dated ...
There is evidence for a left germinal matrix hematoma that has ruptured into the left ventricle. The ventricles are massively dilated and filled with hemorrhage, left greater than right with blood/fluid levels. There is parenchymal extension of the hemorrhage through the corpus callosum into the left cingulated gyrus and the left centrum semiovale. The parenchymal hemorrhagic extension extends to involve the posterior left thalamus. These findings are not significantly changed from the CT scan.
There is evidence for subarachnoid hemorrhage in the sulci of the right frontal and parietal lobes which may represent recirculation of the subarachnoid hemorrhage present in the ventricles.
There is extensive T2 white matter hyperintensity present, much of which may be due to immaturity. However, there is more intense signal in the parietal lobes bilaterally which may represent superimposed edema.

There are several foci of hyperintensity on the diffusion weighted images in the medial parietal lobes bilaterally. However, there is no ADC restriction in this location and acute infarction is unlikely.
There is low T2 signal in the thalami bilaterally and posterior lentiform nuclei bilaterally as well as the posterior pons along the floor of the fourth ventricle. These areas may represent petechial hemorrhage versus myelin breakdown products.

IMPRESSION
1. EVIDENCE OF A LARGE LEFT GERMINAL MATRIX HEMORRHAGE WITH RUPTURE INTO MASSIVELY DILATED LATERAL VENTRICLES AND PARENCHYMAL EXTENSION AS DESCRIBED ABOVE. THIS APPEARANCE IS NOT SIGNIFICANTLY CHANGED FROM THE PRIOR CT SCAN.
2. THERE IS LOW T2 SIGNAL IN THE THALAMI BILATERALLY AND POSTERIOR LENTIFORM NUCLEI BILATERALLY AS WELL AS THE POSTERIOR PONS ALONG THE FLOOR OF THE FOURTH VENTRICLE. THESE AREAS MAY REPRESENT PETECHIAL HEMORRHAGE VERSUS MYELIN BREAKDOWN PRODUCTS.
3. THERE IS EXTENSIVE T2 WHITE MATTER HYPERINTENSITY PRESENT, MUCH OF WHICH MAY BE DUE TO IMMATURITY. HOWEVER, THERE IS MORE INTENSE SIGNAL IN THE PARIETAL LOBES BILATERALLY WHICH MAY REPRESENT SUPERIMPOSED EDEMA.

RADIOLOGISTS: SIGNATURES:
[report_end]

In this case, it is obvious that the patient suffers from major abnormalities affecting the integrity of the brain and thus such a patient had to be discarded from the “normal” patient cohort.

This information can only be retrieved by individually reading through each report, consequently, the prospect of building an automated method/query that could filter patients according to their status (“normal” versus ”abnormal”) was not feasible. Nonetheless, one part of the radiology reports shared by all is worthy of special attention, namely the “IMPRESSION” section. Indeed, it contains a final note summarizing the reports that can facilitate the evaluation of the patient’s status and optimize the time needed to read through hundreds or even thousands of reports. After spending dozens of hours reading through hundreds of radiology reports, I quickly realized that key sentences were used by different radiologists to describe a brain as “normal” in this section; I recorded them and a non-exhaustive list is available in Appendix 8.7. This list could be used towards the goal of developing automated ways of selecting patients of interest for further investigations using for example machine learning methods with the complete list of key sentences as training data (see Discussion part 5.1).

We have to keep in mind that radiology reports are not always as easily classified as the two examples provided above, their interpretation could be subject to personal judgments that would render their automated classification even more challenging. Practically speaking, when such an ambiguous case was encountered (i.e. when a radiology report could not be unequivocally classified as “normal” or “abnormal”, according to my current medical knowledge), the patient was classified as “undefined” and set aside for further investigations by more knowledgeable people in the field in order to be correctly classified.

In order to keep track of this classification, I created another table in the Access database which contained two fields: MRN and Status. After each radiology report was reviewed, the corresponding patient was tagged as “Normal”, ”Abnormal” or “Undefined” in this table. This new table could be linked to the other ones and used as explained in previous chapters in order to refine the queries. Therefore this table enabled the selection of patients not only according to their ages or procedures but also according to their brain image health status.

3.2.2.5. Longitudinal data

The last part of this database mining was to investigate the availability of longitudinal brain MRI datasets. In other words, I wanted to know whether some patients had multiple brain MRI scans at different ages. Basically, the same protocol explained in part 3.2.2.1 was used in order to obtain the ages at which they had a scan for each patient. The resulting table displayed a list of MRNs (i.e. a patients’ list) with the corresponding ages at the time of scan (2 column table). In order to present these results in a more intuitive way, I wrote a small script in Visual Basic (VB) programming language to first transform this two-column table into a more comprehensive one and then include additional information.

3.2.3. MRI data retrieval from PACS – mi2b2 software

The last chapters were dedicated to the search of patients of interest by using different appropriate selection criteria. Once the patient cohort and the related MRI procedures had been determined, the corresponding data needed to be retrieved out of MGH PACS.

3.2.3.1. PACS

In the medical field, PACS, as its name suggests, is a system that provides hospitals with short and long term digital storage, rapid retrieval, management, within and between site distribution and presentation of images of various modalities including Ultrasound (US), Computed Tomography (CT), Positron Emission Tomography (PET) and MRI, among others. It was first developed to replace hard copies such as film archives or hard copy reports by digital supports, saving, as a result, a considerable amount of space, money and greatly facilitating its use. Beside its archive storage utility it also provides a secured network for the transmission of patient information as well as workstations for viewing, processing and interpreting images (Huang, 2011).

This system gave us access to the medical data stored in Digital Imaging and Communications in Medicine (DICOM) format, which is the universal format used in PACS image storage and transfer. A DICOM file not only contains images (pixel data), but also other information such as patient identifiers, demographics, and technical information regarding the scanning sequence parameters

As researchers, we don’t have direct access to this system which is a real gold mine in terms of the quality and amount of medical data it contains. This is the reason why the Medical Imaging informatics Bench to Bedside (mi2b2) team has put all its efforts into facilitating access to PACS and promoting translational investigations.

3.2.3.2. mi2b2 software

As briefly described in the introduction, the mi2b2 project is being developed and deployed by a dedicated team with the aim of providing the infrastructure and regulatory policies necessary to locate and retrieve medical images from PACS[14]. I had the chance to be the first external user working with a prototype of this software. As such, I was able to provide the mi2b2 team with feedback either by directly using the software and reporting errors due to faults or missing elements in the programming and by giving comments about the user interface and functionalities.

Technically speaking, the mi2b2 software provides a way to obtain images from a clinical PACS, but it does so in a way that enables research use of the image data stored in the PACS. Although many research-oriented viewing workstations (OSIRIX[15]) can connect to a PACS, the clinical Radiology departments need assurance that retrieving images for research will not interfere with the clinical mission. In a typical research project, many thousands of images may be obtained, and often in usage patterns that are quite different from conventional clinical workflow (research requests for images often delve deep into imaging archives). Furthermore, audit logs for clinical PACS that must comply with Health Insurance Portability and Accountability Act (HIPAA)[16] and IRB regulations that need to be followed when obtaining images for research are not easily accessible.[17]

The graphical user interface of the mi2b2 Workbench shown in Figure 12 is designed for automating and streamlining the process of retrieving images from the clinical PACS for research purposes (Murphy, Marcus et al, 2011)[18]. The Workbench is a Java Eclipse plug-in that leverages the existing i2b2 open source framework (see Introduction) to communicate through RESTful web services with a corresponding mi2b2 server-side web application. The mi2b2 server-side cell is also built with Java and the i2b2 framework, but additionally utilizes the dcm4che toolkit[19] to find and move DICOM images from the clinical PACS and store them in a local cache. The resulting workflow is designed to preserve clinical performance of the PACS while efficiently retrieving only the desired medical images in a secure way.

Briefly described, the interactive steps are as follows. Once the set of patients has been determined (see parts 3.2.1 and 3.2.2), the list of MRNs is submitted to the mi2b2 Workbench to retrieve information on which imaging studies exist for each patient. In the future, queries can also be made with the accession numbers that are specific to the particular imaging study[20]. Unfortunately, we could not use accession numbers for this project because they were not yet included in the Procedures table we obtained from RPDR. This feature has only been added in the past two months. Mi2b2 queries the PACS and returns the complete list of available studies for each patient. We then browse through the list and select the imaging studies to be downloaded. Because there may be hundreds of studies returned, only a subset of which are our target images,

[14] See mi2b2 wiki page: http://www.na-mic.org/Wiki/index.php/CTSC:ARRA_supplement
[15] http://www.osirix-viewer.com/, ClearCanvas, http://www.clearcanvas.ca/dnn/
[16] http://www.cms.gov/HIPAAGenInfo/02_TheHIPAALawandRelated%20Information.asp#TopOfPage
[17] Adapted from the RO1 grant application (see end of the introduction)
[18] Murphy SN, Marcus D, et al. New Tools for Integrating Clinical Images into Research Studies. Presented at Society for Imaging Informatics in Medicine. 2011.
[19] http://www.dcm4che.org/
[20] A *study*, in this context, is defined as one data set for one particular patient at a specific scanning visit; the term *series* is applied for on particular sequence (T1-, T2-weighted images, DWI images, FLAIR images, etc.) within one study.

one or more filtering and sorting criteria can be used to quickly refine the search list to include only those target images. Once a final study list has been determined, this list is submitted to mi2b2 for download along with the location of a directory where the images should be copied (Figure 12, upper image).

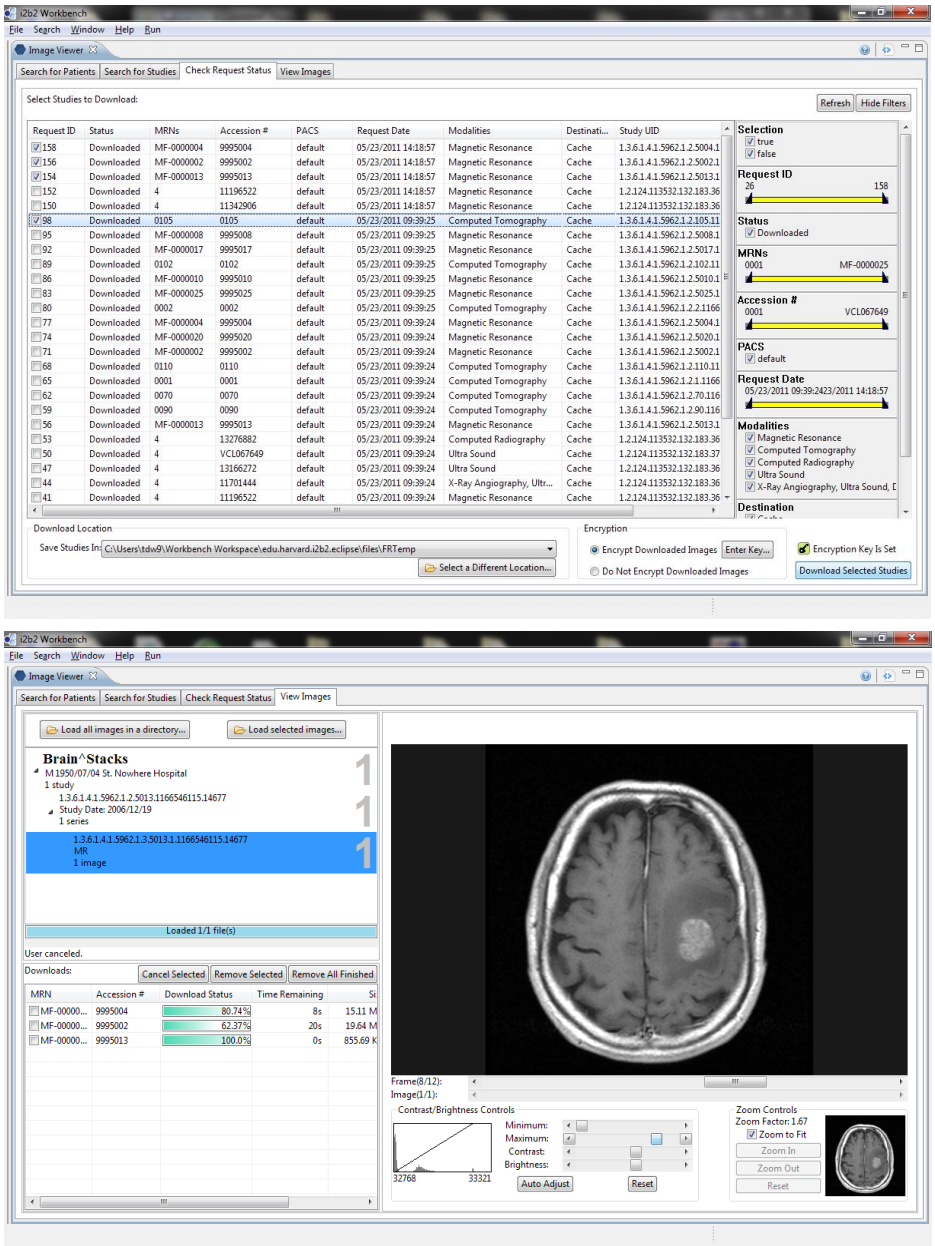


Figure 12: Mi2b2 workbench user interface. It queries the PACS for selected MRNs and returns the complete list of available studies for each patient. One or more filtering and sorting criteria can be used to quickly refine the search list (upper image). An additional tab (which was not available in the prototype I used) will be used to view image files directly from the mi2b2 cache before transferring them to the users own disk space (lower image)[21].

[21] Adapted from Gollub RL, Roch V et al.. Developmental brain ADC atlas creation from clinical images. Displayed at the Organization for Human Brain Mapping, Canada, June 2011.

The mi2b2 server will then queue these requests and download the images based on the rules governing the interaction with each individual institutions' PACS. The rules limit the time, rate, and number of images that may be downloaded in parallel. They are set for each institution so as to not interfere with the clinical workflow. Users can use the mi2b2 Workbench to check the status of their image downloads at any time. Once mi2b2 downloads the images, the status of the request is updated and the user can find their images in their specified directory location. Users also have the ability to use the mi2b2 Workbench to view the image files directly from the mi2b2 cache before they are transferred[22] (Figure 12, lower panel).

In the prototype version that I used, the viewing tab was not yet available and images requested were first downloaded to mi2b2 team members' workstation (Bill Wand, Chris Herrick and David Wang, see People Section 8.3) and then to a secured Windows Share I had access to through my secure Partners logon. The Production version of the software will be released to the Partners user community in the Fall of 2011.

3.2.4. Automated MRI data organization: the "BB-pipeline"

When patient studies are retrieved from PACS through the mi2b2 software, they come in a format that is not well designed for dealing with a huge amount of imaging data (more than 30,000 different volumes retrieved). I built the BB-pipeline for use in a Bash shell on my Linux workstation with the following guidelines in mind: reducing to a minimum any manual steps and facilitating as much as possible further data findings. This pipeline had been improved as I was progressing with this project to fit the special needs and the increasing number of data sets retrieved from PACS through the development of the mi2b2 workbench. Not all studies were requested and retrieved from PACS at once; rather they had been progressively pulled out since February 2011. The aims of this pipeline were to 1) transform "poor" directory hierarchy and name formats into a categorized hierarchy with meaningful name formats, 2) convert DICOM images into NIfTI[23] format, 3) extract technical information about scans to add to the Access database, and 4) to keep track of any studies requested and received from PACS.

Figure 13 delineates the steps required to achieve these aims. Here is a description of the automated BB-pipeline through which all the data sets were sent after PACS retrieval:

(1)

All retrieved studies are transferred from the shared space (see part 3.2.3.2) to our local workstation and decompressed ("unzipped") (Studies are compressed by mi2b2 software when retrieved from PACS).

[22] Adapted from the RO1 grant application (see end of the introduction)
[23] See Neuroimaging Informatics Technology Initiative web site: http://nifti.nimh.nih.gov/nifti-1/

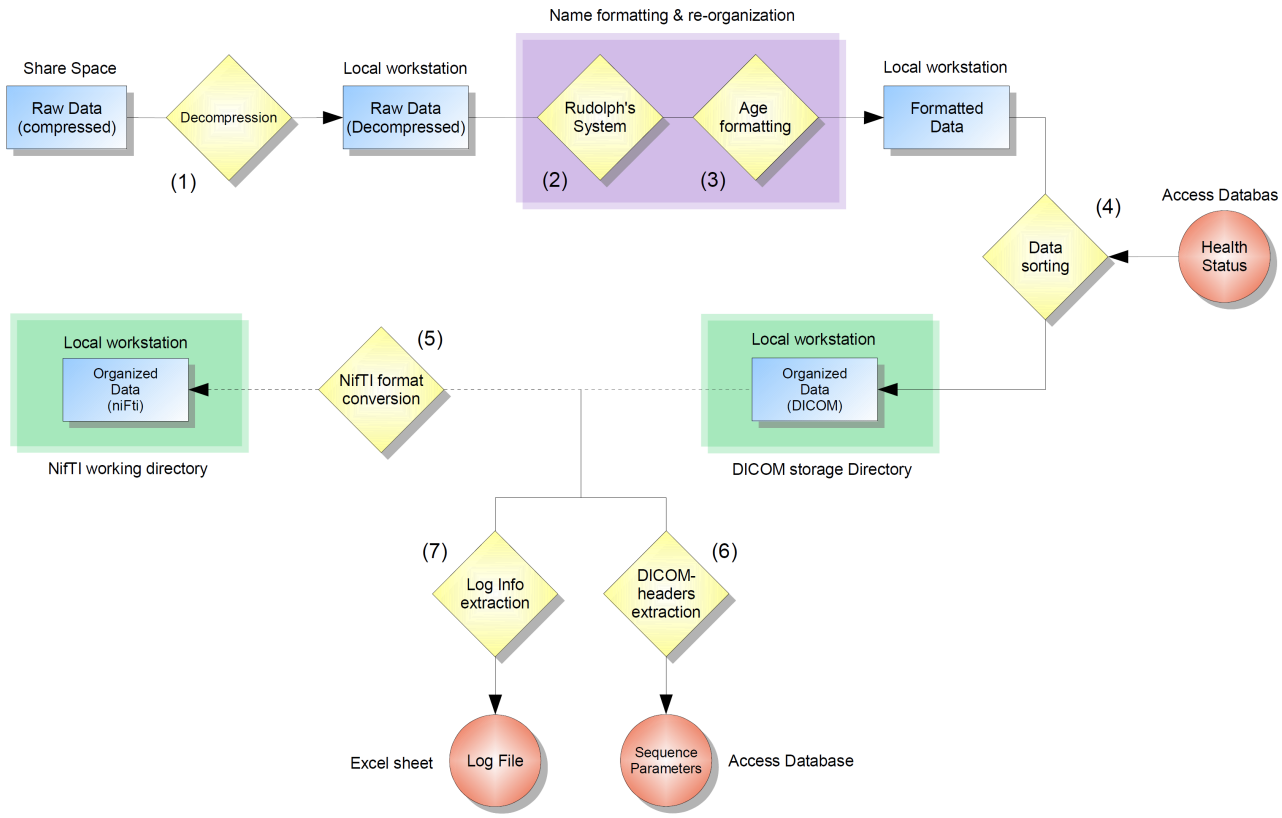


Figure 13: BB-pipline workflow. Data received from PACS are first transferred and decompressed to the local workstation (1). They are then sent through Rudolph's pipeline and age calculator that rename and pre-organize the raw data (2)-(3). Using Access information input, the studies are then sorted into different directories according to their age and health status (4). DICOM files are converted into NIfTI fomat and sent to the working directory (5). DICOM headers are extracted and added to the Access database (6), and log file is updated (7).

(2)

Data are then automatically sent through a pipeline (which has been developed by Rudolph Pienaar, see People section 8.3) that renames and pre-organizes the data.

This system is essentially a DICOM listener/unpacker that receives DICOM user SCU/SCP (Service Class User / Service Class Provider)[24] requests/services. It is built off a DICOM toolkit (DCMTK) set of applications[25]. Upon receipt of DICOM data, several other scripts parse the incoming data for information such as MRNs, age, scan time, machine type and machine ID from the DICOM headers. All images are packed into a directory with a name based on these tags (i.e. <MRN>-<AGE>-<SCANTIME>.... etc.). Images are renamed according to their series numbers which correspond to particular series types (T1-, T2-weighted images, DWI images, FLAIR images, etc.).

Once the final transmission has been received, a termination script analyzes all the received images and builds a table-of-contents text file that is stored in the directory. This file can be quickly consulted in order to get an idea of what the study contains:

[24] For further additional information, see Oleg S. Pianykh (2008), *Digital Imaging and Communications in Medicine (DICOM) : A Practical Introduction and Survival Guide*, Springer-Verlag (Berlin Heidelberg)
[25] http://dicom.offis.de/dcmtk.php.en

```
                  Patient ID          7928364
                Patient Name          TART^EMPION
                 Patient Age          002Y
                 Patient Sex          M
              Patient Birthday        20030901
               Image Scan-Date        20050931
          Scanner Manufacturer        GE MEDICAL SYSTEMS
                 Scanner Model        GENESIS_SIGNA
                  Software Ver        08

Scan 0-000001-000001.dcm              SAG T1
Scan 0-000003-000001.dcm              AX FLAIR
Scan 0-000004-000001.dcm              AX T2 FSE
Scan 0-000005-000001.dcm              AX SPGR 3D
Scan 0-000006-000001.dcm              COR FSE T2
Scan 0-000007-000001.dcm              SAG FSE T2
Scan 0-000100-000001.dcm              DWI
Scan 0-000101-000001.dcm              ADC
Scan 0-000102-000001.dcm              LOWB
Scan 0-000103-000001.dcm              EXP
Scan 0-000104-000001.dcm              FA
```

(3)

Once the study has passed through Rudolph Pienaar's (See People Section 8.3) system, the BB-pipeline further modifies the directory name format because the <AGE> fields extracted from DICOM headers are not in a suitable format for further automated age-specific analysis. Indeed, patient ages are displayed in days, weeks, months or years depending on their age range. In consequence, I wrote a script that renames each study directory with a unified age format (in days) by extracting the date of birth and the procedure date of each patient study from the DICOM header, and calculating the time elapse between these two dates.

At this point, we have transformed inconvenient name formats (see Figure 14) into meaningful ones that are easy to handle (see Figure 15, Studies and Images).

(4)

This step consists of sorting the data according to the age of the patients and their health status (see part 3.2.2.4). To do so, information from the Access database has to be accessible to obtain the health status of the patient. Since the Access software is not available from the workstation and is not convenient to work with in terms of compatibility when using Bash scripts, a manual step is needed to retrieve individual patient health status and send it as input for the BB-pipeline. Basically, this information is exported from the Access database to a two column text file containing the list of MRNs with the corresponding health status. This text file has to be sent manually to the workstation in a dedicated directory to which the BB-pipeline has access.

From here, the pipeline reads through the text file and sorts the data accordingly (Figure 15, Health status). The data that are retrieved from PACS but are not yet classified are sent to a *temporary* directory. Simultaneously, a function calculates the age of the patient in months from the age in days and sorts the studies accordingly (Figure 15, Age, Health status).

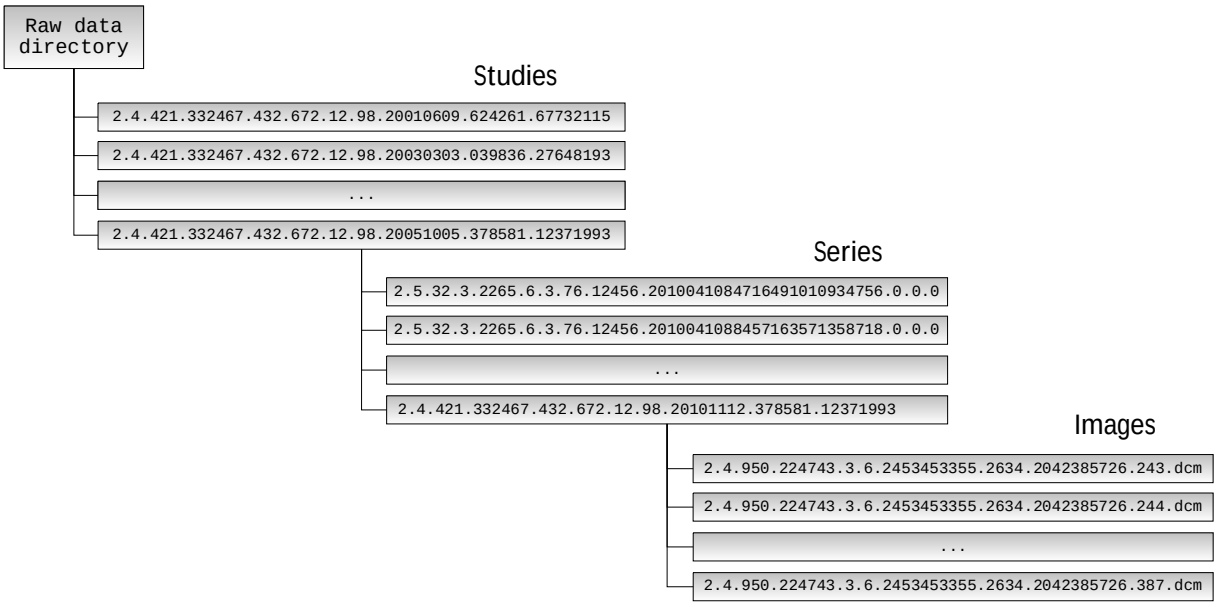


Figure 14: Directory hierarchy and name format as retrieved from PACS. Directory and file names are a combination of institution, manufacturer, machine model, and randomly generated numbers based on date and time. According to DICOM, they are globally unique identifiers.

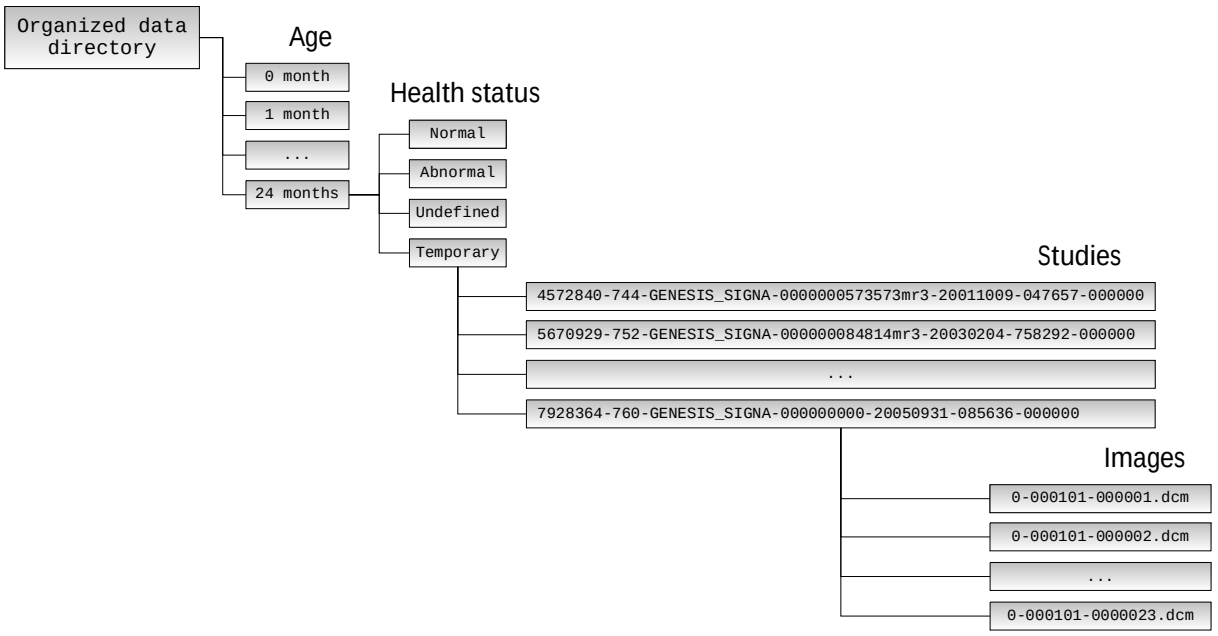


Figure 15: Directory hierarchy after data have been sent through BB-pipeline. They are now sorted according to the age in months, health status, names of directories, and files have been transformed meaningfully: the study name now contains patient MRN, age in days, study date, and other scanner information. The image name includes the series numbers corresponding to a particular series type (T1-, T2-weighted images, DWI images, FLAIR images, etc.).

(5)

After the names and structure of the results have been transformed, DICOM images (.dcm) that are needed for further data analysis are converted into NIfTI format (.nii) which is easier and faster to work with. Indeed, each whole volume in DICOM format is separated in several files, one for each slice. For instance, an ADC volume is found as:

```
0-000101-000001.dcm
0-000101-000002.dcm
0-000101-000003.dcm
…
0-000101-0000023.dcm
```

101 being the series number corresponding to ADC images and *1, 2, 3, …, 23* corresponding to the slices. Furthermore, as mentioned earlier, DICOM files contain a lot of additional information about patients and scan parameters that are not directly needed when working with the images only. This excess information also slows the process. Thus, converting DICOM images into NIfTI format stacks separate slices into one single file and eliminates patient information, resulting in only the useful information regarding the scan volume (orientation, number of slices, pixel dimensions, etc.).

The pipeline automatically passes through all the DICOM files, looks into the corresponding headers ("series description") for key words related to the series we are interested in (DWI, ADC, FA, LOWB, AX-T2 for instance), converts images into NIfTI file with the following name format convenient to work with:

<MRN> - <Age> - <Scan_Date> - <Series_Number> - <Series_Name>

and finally sorts the data according to the same steps as explained in (4). All the DICOM files are kept intact in a storage directory.

(6)

This step consists of extracting additional technical information about retrieved scans and including them in the Access database. By doing so, the completed database not only contains all the information about patient procedures, radiology reports, and diagnosis, but also all the related technical scan details that will be used to select the data of interest (see part 3.2.5).

The pipeline goes through all the series and extracts any chosen related DICOM headers. The extracted information contains mainly details about scan sequences such as TR, TE or the name of the series (T1-, T2-weighted, ADC, FA, …), but also some information about the resolution of the images (see Appendix 8.8). All these DICOM headers are extracted from each series to a text file (in a specific format) that is manually included as a new table in the Access database (now called "enhanced Access database").

(7)

Finally, a text file, containing the list of MRNs and the corresponding scan date of each study that has been sent through the pipeline, is generated and used to update an excel sheet log file. This log file contains, for each study, the MRN and corresponding study date as well as the date it was requested from PACS through mi2b2. A VB script takes the previously generated file and updates the log file with retrieval dates (see Table 2).

MRN	Study Date	Request #	Request Date	Status	Retrieval Date	Age in days	>= 2scans
4357692	3/7/2002	6	3/9/2011	done	3/15/2011	65	
4568929	4/21/2003	12	4/5/2011	done	4/11/2011	171	
4798324	4/7/2003	7	3/15/2011	done	3/18/2011	79	x
4798324	4/17/2003	7	3/15/2011	done	3/18/2011	89	x
5678989	5/3/2004	12	4/5/2011	pending		599	
5839587	2/23/2005	5	3/7/2011	done	3/8/2011	185	x
6738598	11/14/2005	11	3/28/2011	done	4/8/2011	449	x
6738598	6/29/2009	11	3/28/2011	pending		1072	x
6937563	1/5/2006	10	3/21/2011	done	3/28/2011	15	
7984745	7/21/2006	25	5/13/2011	pending		872	
8364634	3/1/2007	10	3/21/2011	done	3/28/2011	9	x
8573645	2/2/2007	10	3/21/2011	done	3/28/2011	12	x
…	…	…	…	…	…	…	…
8736545	1/26/2007	10	3/21/2011	done	3/28/2011	34	
8936745	1/27/2007	19	4/29/2011	done	4/30/2011	3	
9164653	2/27/2008	19	4/29/2011	not found		2	
9465257	1/23/2008	12	4/5/2011	done	4/14/2011	218	
9672356	1/17/2009	19	4/29/2011	done	4/30/2011	10	x
9672356	1/22/2009	19	4/29/2011	done	4/30/2011	15	x
9672356	1/23/2009	19	4/29/2011	done	5/3/2011	16	x
9832545	2/11/2010	12	4/5/2011	done	4/14/2011	491	

Table 2: The log file contains MRNs with their corresponding study date, request number (equivalent to a request ID) with its request date, the status of the request, the date at which the study has been retrieved and finally other useful information concerning the study (Age in days and whether the patient has had several scans). The study is flagged as "done", "pending" and "not found" depending on whether the study has been successfully retrieved, not yet retrieved or was not found in the PACS respectively. A VB script automatically updates this datasheet from the text file generated by the BB-pipeline.

This log file was useful in working with the development mi2b2 prototype in order to give feedback to the mi2b2 team regarding the number of PACS requested studies and the number actually retrieved, and to re-request only the data that were not retrieved. It was also helpful when requesting new data that had not been already requested in previous queries.

Finally, another script was used to automatically update the DICOM storage directory and the NIfTI working directory after the Access database was modified in terms of health status in order to sort the studies sitting in the *temporary* directories (see step (4)) or to redirect studies whose health status had been modified.

3.2.5. Selecting data of interest

At this stage, all the information (medical information and technical scan parameters) concerning retrieved studies was available in the enhanced Access database allowing, as a result, the execution of the previously mentioned request example (see part 3.2, end of the overview):

"Give me the ADC volume of any patients who were scanned at the age of 30 to 60 days whose corresponding radiology reports don't mention any major brain abnormality"

It was necessary to link the new "technical" table with the *Demographics* table to calculate the age in days of the patient at the time of the scans using an age restriction (>= 30 and <= 60) and a criteria for the sequence name (Like "ADC"), and finally include a criteria on the "status flag" table

to only query "normal" patients (namely those whose radiology reports mentioned no major abnormality). Querying any combination of Access tables for specific criteria, at that point, allowed the retrieval of any information of interest and an easy access to the corresponding sorted images either in the DICOM storage directory or the corresponding NIfTI working directory.

With the framework in place to be able to selectively and repeatedly access images that fulfill specific criteria, the next challenge was to determine whether the image acquisition parameters of these scans were sufficiently comparable to enable directly pooled analyses.

When working with diffusion images, the first image acquisition parameter that has to be taken into account is the b-value as explained in part 2.1. In theory, D values are understood to be an intrinsic property of the brain for a determined volume, but the reliability of the D values partially depends on b-values specified by the diffusion scan sequences (Ogura et al., 2011). TE, TR as well as other parameters have less of an influence on D values, but rather these specifications affect the quality of the images, Signal to Noise Ratio (SNR), and accuracy of measured D values[26].

Another concern that was raised by reviewing the retrieved data that was collected between 2000 and 2010 and confirmed by Dr. Ellen Grant, was the complete change in the pediatric acquisition protocol for MR diffusion imaging sequences at MGH between the years 2005 and 2006. At that time, under her supervision they completely changed the software and the sequence parameters used for diffusion MRI in order to improve the resolution and the quality of the images. Unfortunately, we noticed that a lot of diffusion volumes acquired using the new protocol (2006-2010) contained artifacts and were less consistent across patients. Dr. Ellen Grant confirmed that clinicians had encountered some difficulties with the new protocol which directly affected the diffusion scans. For these reasons, and for the limited amount of time that I had at my disposal to analyze the data, we decided to focus further investigation using only the images with the "old" sequence parameters (before the year 2006) when the diffusion scans were more consistent in terms of the quality and sequence parameters across patients. This insured a higher comparability across different patients at the cost of lower resolution and less accurate diffusion indices. Future work of the team will focus on developing image processing and/or statistical modeling approaches to enable use of the newer data for the analyses done for this thesis.

All the details concerning the choices of diffusion data can be found in the results section part 4.1.4.

3.3. MRI Data Analysis

This last method section is dedicated to MRI data analysis per se, using the data that were retrieved and met all the selection criteria described in the previous chapters. Despite filtering out a considerable number of patients, I still ended up with several hundred studies that could be used for data analysis. Taking into account the limited amount of time left at this point, any manual processing (e.g. ROI delineation) or new algorithm development was not feasible. Consequently,

[26] Adapted from a discussion with Simon K. Warfield, Ph.D., Professor of Radiology, Harvard Medical School, Director of Radiology Research, Director Computational Radiology Laboratory, Department of Radiology, Children's Hospital Boston.

we decided to investigate time evolution of diffusion indices across ages using already available tools and simple processing methods that can be automated as much as possible.

We focused our attention on two diffusion indices, namely ADC and FA. In brief, for each selected volume we removed any non-brain parenchyma regions and calculated the average ADC and FA values for the whole brain. This gave two values (two markers: whole brain ADC and FA average) for each volume which was compared across patients at different time points (for sake of simplicity, these two markers were called WBA_{ADC} and WBA_{FA} values, WBA = Whole Brain Average)

3.3.1. Average calculation of whole brain diffusion indices

I specifically developed this brain extraction pipeline for the sake of this project (with advice from Dr. Ellen Grant, Lilla Zollei and Rudolph Pienaar), trying to find the best methods that did not require any manual editing of images. Several attempts were necessary to reach this goal and this process gave me my first insight in brain imaging tools use and script designs. This pipeline is displayed in Figure 16.

The first processing step to remove non-brain regions is a brain extraction procedure carried out by the Brain Extraction Tool (BET)[27] as part of the FMRIB Software Library (FSL)[28]. This method is based on a deformable model which evolves to fit the surface of the brain by applying a set of locally adaptive model forces (Smith, 2002). It finds the surface of the brain (including brain stem and cerebellum) and removes external non-brain regions. I tried to use BET directly on ADC or FA volumes but since this tool was initially developed to be used on T1- or T2-weighted volumes, it did not work properly. Hence, for each ADC/FA volume I used the corresponding LOWB volume (reference volume, which is actually a T2-weighted volume) in order to effectively run BET (Figure 16, (1)).

The default parameters were used except for the Fractional Intensity Threshold (FIT)[29] that was set at 0.4. After testing different FIT values, we determined that this value worked the best on the LOWB volumes. Using the output of BET from the LOWB volumes, a mask was generated and applied to the corresponding ADC volumes (Figure 16, (2)).

BET does not remove Cerebrospinal Fluid (CSF). CSF needs to be excluded from the ADC and FA average calculations. CSF is not brain tissue, varies in size widely across individuals, and has markedly higher ADC and lower FA values thus introducing a spurious confound into WBA_{ADC} and WBA_{FA} measures. Therefore, a different way to remove CSF had to be found. Water molecules in CSF are known to have a nearly free diffusion property and thus should appear as voxels with highest value in ADC volumes (which represent the average water diffusion rate for each voxel, see part 2.1). As a result, a voxel-wise upper threshold (threshold = 2.05×10^{-3} [s/mm^2]) was applied on previously brain extracted ADC volumes in order to get rid of CSF regions (Figure 16, (3)). Several threshold values were tested and the resulting volumes were visually inspected in order to find the best threshold value.

[27] http://www.fmrib.ox.ac.uk/fsl/bet2/index.html
[28] http://www.fmrib.ox.ac.uk/fsl/
[29] The default FIT is 0.5, by setting it under 0.5, the overall segmented brain becomes larger (<0.5) and is less restrictive.

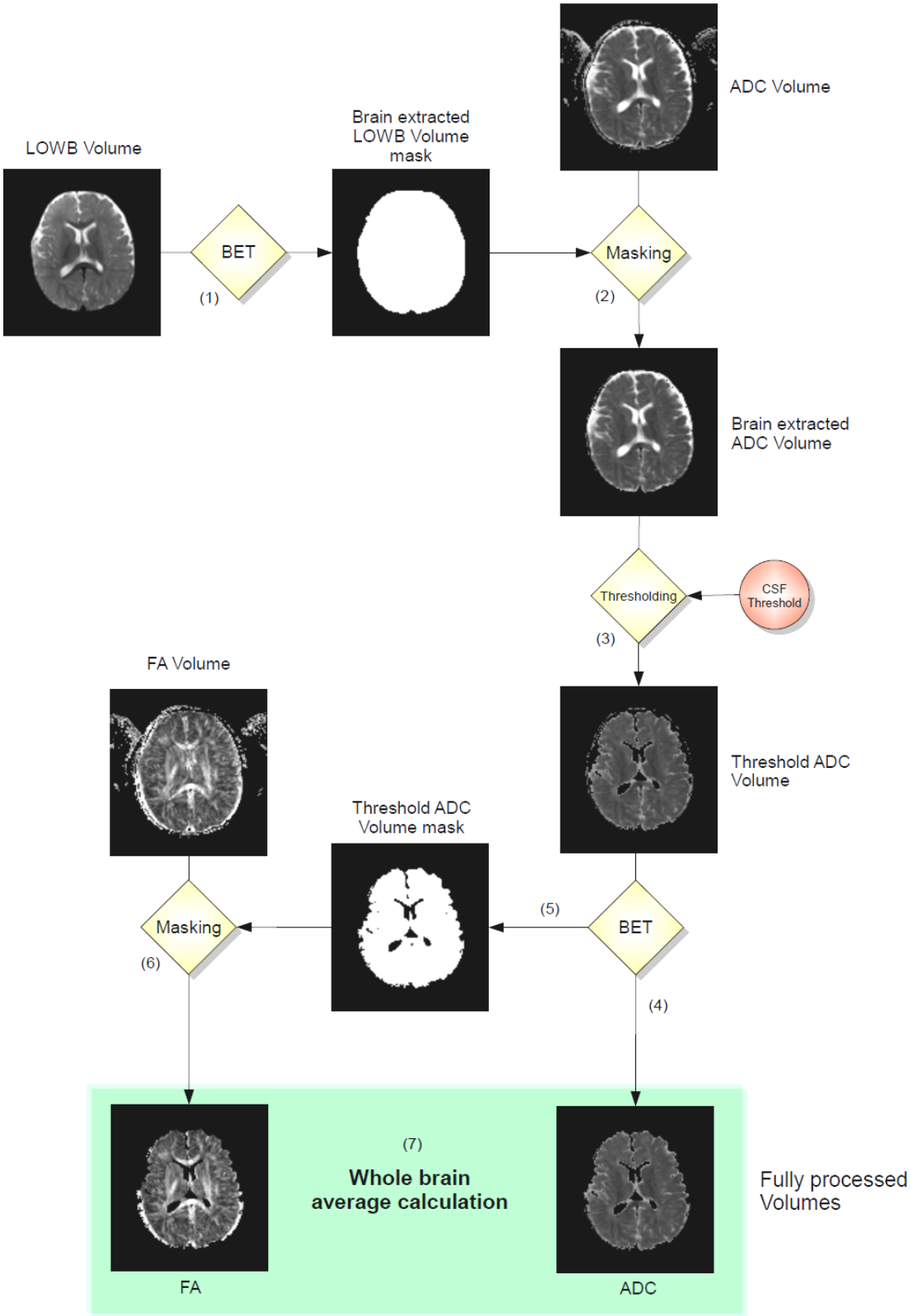


Figure 16: Data processing workflow for the calculation of WBA ADC and FA. The LOWB volume is brain extracted through BET (1) and applied as a mask to the ADC volume (2). Then, an up-threshold is applied to this volume in order to remove CSF regions (3) and the resulting ADC volume is sent through BET (4) to obtain the fully processed ADC volume. At the same time, a mask is generated (5) and applied to the corresponding FA volume (6) to obtain the fully processed FA volume.

By visually inspecting the outputs of this last processing step, I noticed that some non-brain tissue remained at the periphery of the brain (e.g. fragments of skull, dura, etc). I found that applying BET (FIT = 0.4) directly on these ADC volumes was effective in removing these remaining non-brain regions at the periphery without affecting any other parts of the brain (Figure 16, (4)). In this way, the fully processed ADC volumes were obtained and used for average calculations.

The non-zero voxels in these newly obtained ADC volumes for each individual were then used as a mask (Figure 16, (5)) for their corresponding FA volumes (Figure 16, (6)). Thus, fully processed FA volumes used for average calculations were obtained.

Finally, whole brain averages were calculated from non-zero voxels of fully processed ADC and FA volumes (Figure 16, (7)) using MatLab (The MathWorks, Natick, MA, USA).

3.3.2. Diffusion indices time evolution analysis

The final aim of this project was to use the previously calculated WBA_{ADC} and WBA_{FA} values in order to investigate the time evolution of these markers across ages.

I first looked at the effect of the data processing methods used in section 3.3.1 on the distribution of WBA_{ADC} and WBA_{FA} values across ages. These average values were plotted for each patient according to their age in days after each processing step. The patients were then grouped according to their age in months (0 month, 1-2 months, 3-4 months, etc.) and the standard deviation of WBA_{ADC} and WBA_{FA} for each group was calculated. A decrease in standard deviations, before and after the data were processed, was expected by reducing the variability across patients due to non-brain tissue and artifacts. According to the literature (see part 2.2), an overall decrease of WBA_{ADC} values and an overall increase of WBA_{FA} values across ages were also expected (see results part 4.2.1).

Subsequently, I tried to mathematically describe more precisely the time evolution profile of these indices, by applying different curve fitting methods to the data in order to see which function describes the data best.

In this context, three different models were used (x = age, f(x) = estimated WBA_{ADC} or WBA_{FA}):

(1)	Linear:	$f(x) = h + m \cdot x$	(h, m constants)
(2)	Logarithmic:	$f(x) = a + b \cdot \ln(x)$	(a, b constants)
(3)	Biexponential:	$f(x) = p \cdot e^{-x/s} + q \cdot e^{-x/t}$	(p, s, q, t constants)

I decided to use the logarithmic and biexponential models as suggested in previous studies (Löbel et al., 2009 and Muhkerjee et al., 2001 respectively). All parametric fits were performed by the Curve Fitting Toolbox[30] provided by Matlab. A standard linear least square method was used to estimate the parameters in linear equations and a nonlinear least square method to estimate the parameters in logarithmic and biexponential equations.

[30] See Curve Fitting Toolbox manual in pdf format :
http://hug.phys.huji.ac.il/PHYS_HUG/MAABADA/Mabada_b/curve%20fitting%20in%20Matlab.pdf

R-square (R^2) fit statistics were used to compare the models and determine the best one. This statistic measures how successful a fit is in explaining the variation of the data. Resulting R^2 values fall between 0 and 1, a value closer to 1 indicating a better fit. Let's say that we obtain an $R^2 = 0.81$ for a particular fit, this means that this fit explains 81% of the total variation in the data around the average.

Prediction bounds for the fitted function and for new observations were also included. To understand these two concepts, we can take the example of a curve that fit the WBA_{ADC} values across ages. The 95% confidence prediction bounds *for the fitted function* denote the range of WBA_{ADC} values at each age where the true mean WBA_{ADC} value for that age range should be within 95% certainty. The 95% confidence predictions bounds *for new observations* represent the range of WBA_{ADC} values at each age within which the WBA_{ADC} value for a new individual of that age, taken from this population, can be determined with 95% certainty. In other words, this interval indicates that we have 95% chance that the new observation is actually contained within these lower and upper prediction bounds (see results part 4.2.2).

Then, a General Linear Model (GLM) including Age (days from birth) as a continuous predictor and Gender (Male/Female) as a categorical predictor was used in order to examine the effect of these variables and their interaction on WBA_{ADC} and WBA_{FA} (see results, 4.2.3). As time evolution of WBA_{ADC} and WBA_{FA} values did not seem to follow a linear model, but rather a logarithmic or biexponential model, the GLM was applied after transforming Age and WBA values using natural logarithms (i.e., ln(Age) and ln(WBA)) (see results, 4.2.3). These analyses were performed using Statistica v.10 software (StatSoft, Tulsa, OK, USA).

Finally, as mentioned previously, I also looked for longitudinal data of patients with multiple scans at different ages and they were used as controls for within subject time evolution of the diffusion indices. Four patients with multiple scans were found and highlighted in the WBA_{ADC} and WBA_{FA} age evolution plot (see results part 4.2.4).

4. Results

In this results section, I will first present the results acquired from the step-by-step database mining process resulting from the RPDR query (see part 4.1.1), then discuss the corresponding Access database medical information (see part 4.1.2), and finally the scan parameters of the data retrieved from PACS (see parts 4.1.3 & 4.1.4).

I will then focus my attention on the results obtained from the data analysis investigating cross-patient time evolution of different diffusion indices between the ages of 0 and 2 years (see part 4.2).

4.1. Database mining (Medical informatics)

4.1.1. RPDR query

The RPDR query described in part 3.2.1 was run on the 5th of November 2010 and the resulting aggregate numbers are shown in the bottom right quadrant of Figure 11 (the screenshot displayed in Figure 11 is not the actual query requested at that time (note the dates do not match), but rather the query was run again with the same criteria and parameters for the sake of this report to obtain the screen shot.

Note the aggregate number of patients satisfying our selection criteria, namely patients who had a brain MRI scan at the age of 0 to 6 years since 2000 and who don't have any HIV history. We obtained 6886 ± 3 patients who met these criteria (restraining this query to patients from 0 to 2 years of age, we obtained 5278 patients). This number was a rough estimate of how much data might be of interest in the context of this project. We have to keep in mind that it also included patients who were older than the age range of interest and that some patients may not have undergone an MRI scan that included the modality we were interested in (i.e. diffusion imaging). This number gradually decreased as we refined the query in the following steps.

The distribution of gender in this population was 44% female and 56% male with the following ethnic distribution: 66.8% White, 13.4% Hispanic, 5.9% Black, 3.9% Asian and 10% Other/Unknown. Finally, at the time of the query, 98% of the patients were still alive.

4.1.2. Microsoft Access Database queries

As explained in part 3.2.2, the Access Database retrieved from RPDR enables us to refine our queries after obtaining additional medical information regarding the patients. This information was extracted before any data were retrieved from PACS. Therefore, it gave us some crucial information about the data potentially available from PACS and helped focus our attention on the most valuable and interesting studies.

4.1.2.1. Number of brain MRI scan for each age range

In the first step, I limited the query to patients aged 0 to 6 years (0 to 72 months) at the time of the scan and whose procedures were included in Table 1 using SQL methods explained in part 3.2.2. Doing so, the aggregate number of unique patients obtained dropped from 6886 to 3044. However, when the total number of scans that met the criteria was considered (i.e. taking into account that some of these patients had several scans at different ages), I found a total number of 4272 brain MRI scans that met the inclusion criteria.

Looking at the distribution of the scans according to the age of the patients at the time of the scan, I obtained the histogram displayed in Figure 17A.

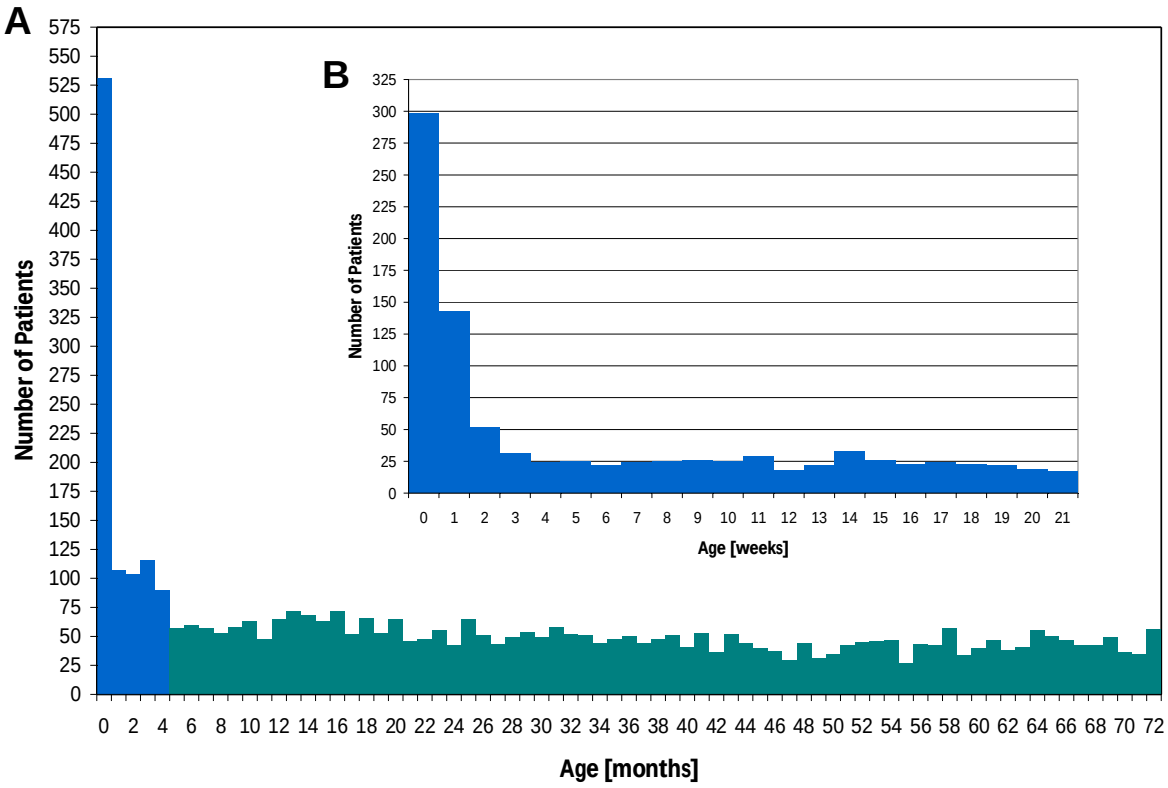


Figure 17: Number of patients who had a brain MRI scan at a particular age in months from 0 to 6 years (A) and in weeks from 0 to 4 months (B).

From this figure, it is clear that the distribution of the number of patients with a brain MRI scan is on average homogenous across different ages except from 0 to 4 months where there are more scans. After 4 months, the average number of brain MRI scans per month remains constant (mean = 49, std = 10.02) (data from 0 to 4 months is not included in this average). By looking more precisely (in weeks) at the distribution of scans between 0 and 4 months (see Figure 17B), we notice that a predominant number of scans were made in newborns within the first days of life (first 2 weeks). This is expected given that the predominant indication for pediatric neuroimaging is complications at the time of birth (e.g. traumatic birth with suspected ischemia or seizures or some abnormality in the neonates' anatomy or behavior).

4.1.2.2. Proportion of MRIs that include diffusion imaging

Using the keyword criteria selection, using words related to diffusion images used by physicians in radiology reports (see part 3.2.2.2), I was able to approximate the proportion of these scans that likely included diffusion images. Figure 18 displays the proportion of brain MRI scans that contain diffusion imaging (in blue) and the proportion of undetermined scans (in orange).

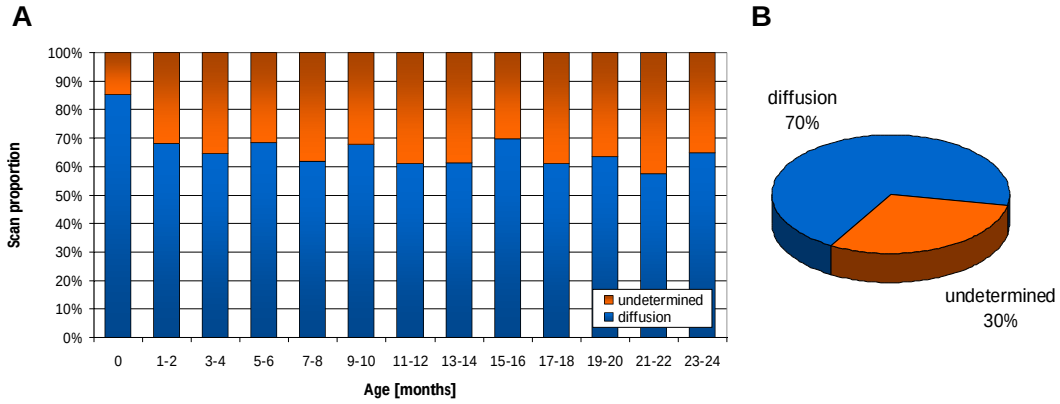


Figure 18: Proportion of Diffusion brain MRI scans from 0 to 24 months. The blue portion represents the scans that include diffusion criteria and the orange portion represents the remaining brain MRI scans (A); average proportion of diffusion and undetermined scans over 0 to 24 months (B).

The age-range from 0 to 24 months is shown for simplicity, but this proportion stays more or less constant across different age ranges, with an average of 70% diffusion scans. However, the fact that no diffusion-related term is used by the physicians in a radiology report description doesn't necessarily imply that this scan doesn't contain diffusion imaging; the percentage of MRI data containing diffusion imaging may be higher than expected as we will see in part 4.1.4.

4.1.2.3. How old are the scans?

Another interesting information that can be queried from the Microsoft Access database is the proportion of the scans acquired at different time periods (years) for the reasons explained in part 3.2.2.3 & 3.2.5. Figure 19A displays the proportion of scans that were completed before 2006 and after 2006 for each age range (from 0 to 24 months). Figure 19B shows the overall distribution of these scans for each year from 2000 to 2010.

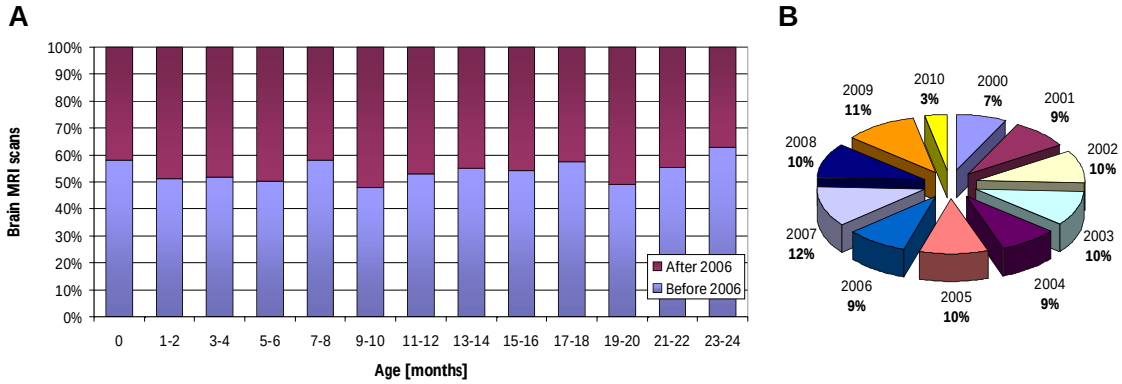


Figure 19: Proportion of scans acquired before the year 2006 (blue) and after 2006 (magenta) (A) and overall distribution of these scans for each year from 2000 to 2010.

It is clear that the scans are fairly evenly distributed across the age range within different years. The proportion of 2010 scans relative to the total number of scans is fewer in comparison to other years since our database contains only scans completed before August 2010.

In addition, a large proportion of the scans (55%) were acquired before the year 2006, where we decided to focus our attention.

4.1.2.4. "Normal" or "Abnormal" brain MRI scan?

At the time of this report, I manually reviewed more than 1500 radiology reports (out of 2111) from patients aged 0 to 24 months at the time of the scan and tagged them according to criteria explained in part 3.2.2.4. So far, all the patients from 0 to 12 months old have been classified and their proportion are displayed in Figure 20A; the average proportion of "normal", "abnormal" and "undefined" patients from 0 to 12 months is displayed in Figure 20B.

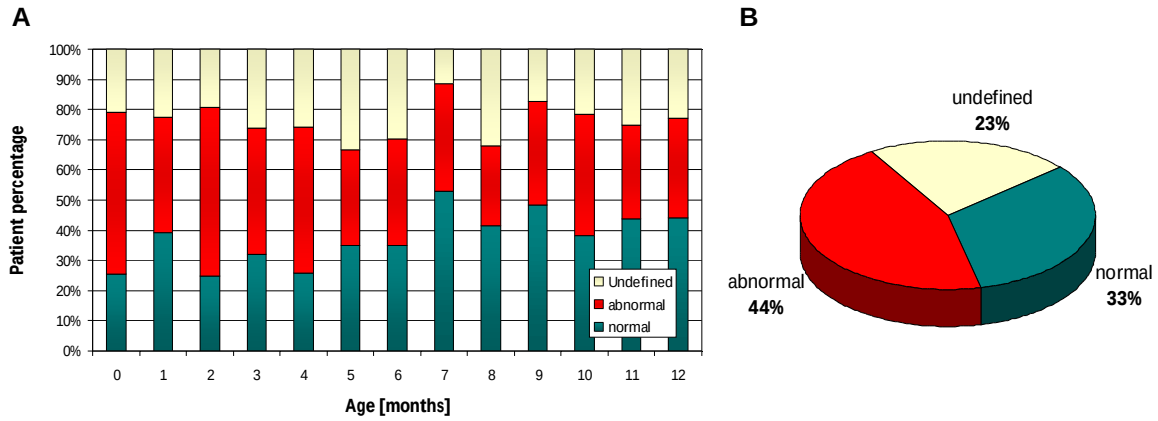


Figure 20: Distribution (A) and average proportion (B) of "normal", "abnormal" and "undefined" patients across the ages of 0 to 12 months.

We noticed that on average 33% of patients have "normal" outcomes from their radiology report meaning that no major brain abnormalities were detected by physicians for these patients.

4.1.2.5. Longitudinal data

We were interested in investigating patients with multiple MRI scans at different time points for longitudinal investigations. As explained in part 3.2.2.5, we used a VB script to present the data in a more comprehensive way that includes additional information about the scans; we looked at patients between the ages of 0 and 6 years (see Table 3).

Looking at these results, we note that 493 patients had more than one scan (ranging from 2 to 17 scans). As mentioned in part 3.2.2, a query of CT scans was also constructed in order to obtain supplementary information about the MRI. A table only including CT-related procedures (cf. method explained in part 3.2.2.1 for brain MRI-related procedures) was generated using 15 different CT-related Codes as criteria. Then, by linking the brain MRI-related and CT-related tables by their Encounter_Number and their MRN field, a list of patients having both brain MRI and CT procedures during the same encounter was obtained. From this list, and using the VB script, this information could be integrated in Table 3 by highlighting in yellow brain MRI scans with a

corresponding CT scan. This additional information could be useful either for registration purposes or further investigations in the future.

MRN	Age [months]																							...												# of scans
493	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24		62	63	64	65	66	67	68	69	70	71	72	1479
7080388		o	o																																		2
5788567		o	o																																		2
8565235					o		o																														2
9265863			o																x																		2
⋮																																					⋮
9057163									x				o		o																						4
8496818	o												o				o																				4
1256323	o			o						o																x											4
6234246							o						o		o				x																		4
2545663																																				o	4
6725462	o	o		o	o																																4
⋮																																					⋮
8452562			o										x	o						x																	5
7647524														o						o													x				6
2334522														o	x			x							o												6
6234562														o											o										o		6
8475624								o													o										o						6
⋮																																					⋮
3456894				o			x	o		o				o				x						o													9
8356129															o	o			o				o														9
5825483																											o					o					9
1856038																											o					x					10
9039845						o			x			o			o			o			o																	10
⋮																																					⋮
3759913	o																										o			o			x			o	11
9934724																							x	o													11
3586753															x			x	o	x				o									x				12
8365403																									x				x			x					x	12
9039403																	o	x			o			x					x									13
8563564																														x	o		o		o			14
2351385																											o				x							14
3560199																x	o	o			x			x							o							15
3536623																											x			x			x			x		17

Table 3: This table displays a subset of the data because the complete data set could not fit on a regular page. The first column contains the MRN. The furthest right column indicates the number of brain MRI scans each patient had and the bottom rows partially delineate the age range from 0 to 72 months. The circles specify the scans containing diffusion imaging and the crosses are the undetermined brain MRI scans (see part 4.1.2.2). The yellow highlighting indicates MRI scans that have a corresponding CT scan. Finally, the number in the lower left corner corresponds to the total number of patients having more than one scan and the lower right corner corresponds to the total number of scans for these patients. Note that due to HIPAA regulations, the MRNs have been altered for the purposes of this Figure.

Thus, the data we have at our disposal contains a considerable amount of longitudinal information, which could be valuable for further analysis. Notably, I found and used some of this longitudinal data as a within-subjects control for our time-evolution analysis of diffusion indices (see part 4.2.4).

From the RPDR Access database alone, I was able to roughly estimate the number of potential studies we were interested in, to wit, studies from normal patients aged 0 to 2 years with diffusion scans acquired from the year 2000 to 2005. Out of 2111 studies identified for that age range, at least 70% contained diffusion series, 55% were acquired before the year 2006 and 33% were defined as normal. This left us with an estimation of 268 studies meeting the selection criteria.

4.1.3. PACS retrieval

All previous results were based only on the output from the RPDR query with no additional information from the medical images themselves or from an interaction with the PACS. The next steps were performed with access to the imaging data that enabled me to refine the criteria because I knew exactly what series each study contained, rather than speculating based on the incomplete information in the previous Access database.

To date we have retrieved 1646 studies out of the 2111 studies I identified with the Access database retrieved from RPDR. These studies correspond to patients aged 0 to 24 months at the time of the scan and were retrieved from the MGH PACS. All of the studies from patients aged 0 to 12 months were requested and all the remaining studies are currently being requested through the mi2b2 prototype software. Thus, the following results are based on patients aged 0 to 12 months (Figure 21), but they are also representative of the data we will be able to obtain for full age range (0 to 24 months).

Age [months]	Total # of studies	Retrieved from PACS	Not available
0	531	515	16
1	107	104	3
2	104	101	3
3	116	105	11
4	90	85	5
5	57	55	2
6	60	57	3
7	57	53	4
8	53	48	5
9	58	53	5
10	63	58	5
11	48	44	4
12	65	62	3
Total	**1409**	**1340**	**69**

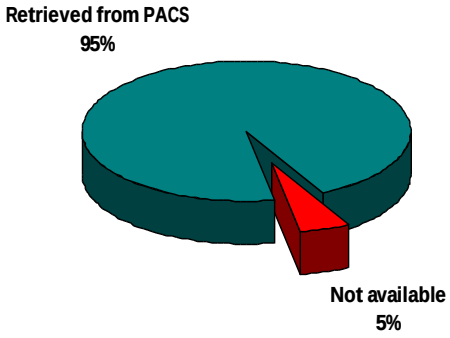


Figure 21: Table summarizes, the total number of studies identified, the corresponding number of studies successfully retrieved from PACS and the number of studies not available, for each age range. The average proportion is displayed in the pie chart (right).

Using the prototype of mi2b2 software, 95% of the studies present in the Access database were successfully retrieved. The remaining 5% were either not found in PACS through the mi2b2 prototype, or failed in the downloading process due to technical issues. Studies that failed to download from PACS were re-requested up to three times, after which point they were considered as not available for this thesis. This retrieval failure was at least in part if not totally due to task and cache management limitations of the mi2b2 software that govern the interaction between the PACS and mi2b2 cell. This project highlighted this critical limitation of the mi2b2 software design and resulted in significant improvements in the subsequent version of the software that is just now being deployed and will be tested later this summer.

4.1.4. Data of interest

By extracting the technical information from the PACS retrieved images and adding them to the Access database as described in part 3.2.4, I was able to refine the query to its final stage. This

allowed the identification of the diffusion imaging data actually used for the time-evolution analysis of the diffusion indices across ages.

First, we examine the actual proportion of studies including diffusion scans to see whether the estimation stated in part 4.1.2.2 was accurate. Using the technical scan information (e.g. "series description") from the studies retrieved from PACS and the Access database, I could precisely determine the percentage of studies that actually contained diffusion images (see Figure 22).

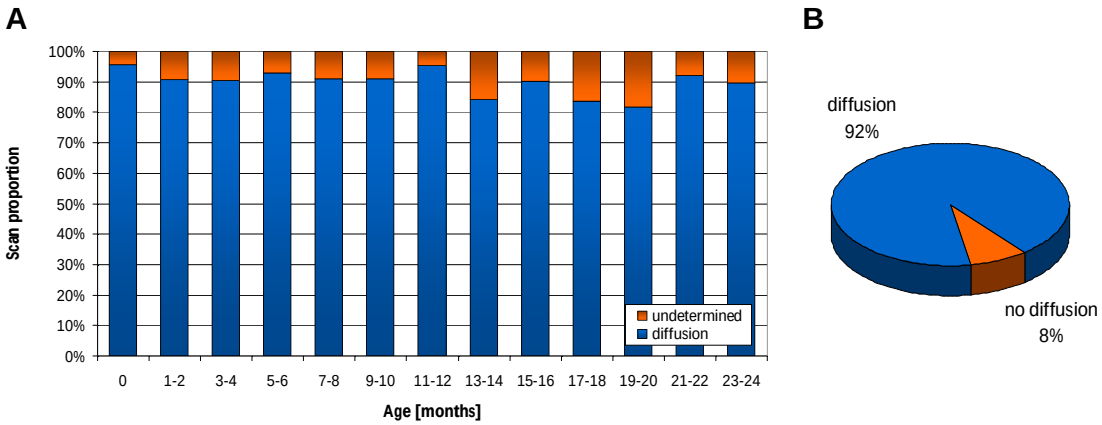


Figure 22: Actual proportion of Diffusion brain MRI scans from 0 to 24 months. The blue portion represents the scans that include diffusion criteria and the orange portion represents the remaining brain MRI scans (A); average proportion of diffusion and no diffusion scans over 0 to 24 months (B).

We note that nearly all of the studies collected (92%, 1515 studies out of the 1646 retrieved from PACS) actually contained diffusion images, which is exactly what our collaborator Ellen Grant predicted based on her intimate knowledge of the clinical scan acquisition protocols; however, this proportion is much greater than was expected based only on the medical information available in RPDR (see part 4.1.2.2, Figure 18). This limitation of the information available in RPDR for this level of detail of the scan acquisition is not easy to address given the current system of medical informatics at our institutions, but this project will help move the system in the right direction.

Having confirmed that most of the imaging studies include diffusion scans, the selection criteria had to be constrained even more at that point. The desired search criteria included:

"Any *comparable* diffusion imaging data of patients between the ages of 0 to 24 months who don't present any major brain abnormality"

The age range and the health status could easily be identified using methods explained in the previous chapters. Then, linking the procedure table with the newly added table containing scan details (see part 3.2.4, step 6) by their MRNs and study dates, I was able to select only information about patients whose corresponding images were retrieved from PACS.

The "comparable diffusion imaging data" criterion, on the other hand, was more complicated to define. As explained in part 3.2.5, we decided to focus our attention on diffusion data acquired before the year 2006 when diffusion scans were more consistent in terms of quality and sequence parameters across patients. However, even within this subset of data, I encountered some variation in sequence parameters. The following is an example of a study with different set of diffusion parameters (Table 4):

MRN	Date	Manufacturer	Magnetic_Field_Strength	Series_Number	Series_Description	Repetition_Time	Echo_Time	Rows	Columns	Pixel_Spacing	Slice_Thickness	Pixel_Band_Width
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	1	Sag T1 FSE	450	9.592	256	256	0.937500\0.937500	5	162.760422
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	2	AX T2	6016.664	110.543999	512	512	0.390625\0.390625	4	44.389202
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	11	AX T2	6016.664	110.543999	512	512	0.390625\0.390625	4	44.389202
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	3	3D SPGR AX-30 DEGREE	30	8	256	256	0.859375\0.859375	1.2	81.40625
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	4	3D SPGR AX-10 DEGREE	30	8	256	256	0.859375\0.859375	1.2	81.40625
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	7	DWI raw b=1000	7499.996	101.300003	128	512	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	100	DWI	7499.996	101.300003	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	101	ADC	7499.996	101.300003	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	102	LOWB	7499.996	101.300003	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	103	EXP	7499.996	101.300003	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	104	FA	7499.996	101.300003	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	8	DWI raw b=700	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	105	DWI	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	106	ADC	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	107	LOWB	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	108	EXP	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	109	FA	7499.996	93.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	9	DWI raw b=1500	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	110	DWI	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	111	ADC	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	112	LOWB	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	113	EXP	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75
5636754	4/22/2004	GE MEDICAL SYSTEMS	15000	114	FA	7499.996	111.400002	128	128	1.562500\1.562500	4	1218.75

Table 4: Scan details about a particular study. Magnetic field strength is displayed in mT, TR and TE in ms, rows, columns, pixel spacing, slice thickness in mm and bandwidth in Hz/pixel. Series number is specific for each series within a study (see part 3.2.4 (2)). Uncolored rows represent non-diffusion data and colored ones indicate diffusion sequences.

Two main problems arose when I looked more precisely at the diffusion data stored in the PACS. First, and most importantly, in most studies, the raw diffusion data (from which the diffusion tensor matrix and the related diffusion indices are computed, see part 2.1) were not available, unlike the example presented in Table 4 (e.g. “DWI raw b=1000”). The decision in the radiology department was to store only volumes from the already computed diffusion indices (DWI average, ADC average and FA, see Table 4). Thus we could not compute these indices with the algorithms of our choice.

The other issue was that b-value information for each sampled direction was unfortunately not available from the DICOM headers. To date, the leadership within the DICOM standards consortium has not yet led the field to include this important image acquisition parameter in the DICOM header, as it is still considered a proprietary field of information. As explained earlier, the main criterion, in terms of comparability of diffusion scan measures, was to have similar b-values. Thus I had to find an alternative way to figure out what the b-values were for each series. Dr. Ellen Grant informed me that they used 3 different sets of values at that time, the longer the TE, the higher the b-value. This information could me confirmed looking at some studies whose raw diffusion data were available and whose b-values were included in the series description name (Table 4). In this particular example, 3 different diffusion sequences were acquired (cf. 3 colors) with the 3 different b-values and the corresponding TEs: b = 700 [s/mm^2], TE = 93.4 [ms]; b = 1000

[s/mm^2], TE = 101.3 [ms]; b = 1500 [s/mm^2], TE = 111.4 [ms]. Using this information, I could determine which TE was used the most and subsequently infer its corresponding b-value (Figure 23).

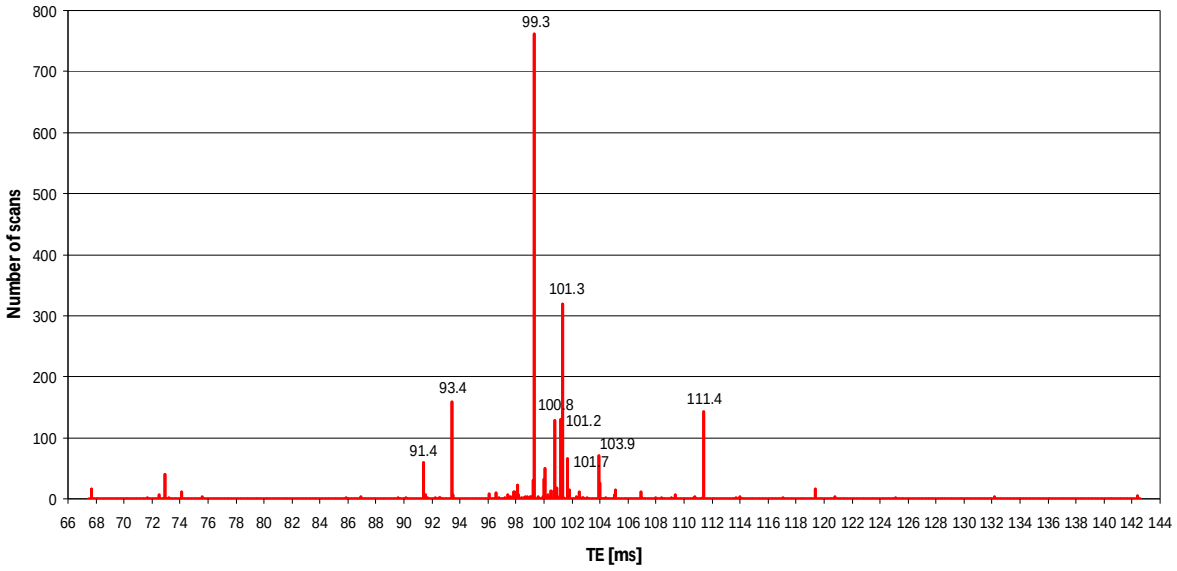


Figure 23: Distribution of diffusion scans according to their TE in ms. This graph includes all of the diffusion data we had at our disposal that were acquired with the "old" (prior to 2006) diffusion protocol (n = 2371).

As expected, we notice three main different TE clusters, one around 93 [ms], one around 100 [ms] and one around 111 [ms] with corresponding b-values of 700, 1000 and 1500 [s/mm^2] respectively. We also note that most diffusion scans (72%) were acquired with a TE near 100±3 [ms]. Looking at all of the raw diffusion data with b-values included in the name of the series, I could conclude with almost 100 percent confidence that all diffusion data with TE >= 97.4 [ms] and TE <= 108 [ms] were acquired with the same b-value of 1000 [s/mm^2]. Therefore, I decided to constrain our diffusion data of interest to those within this TE range, where most of the studies are found. As a result, the b-value comparability criterion can be satisfied and minor differences in TEs, as explained previously, should only have minimal influence on calculated diffusion indices (see part 3.2.5).

One last detail has to be mentioned. I also came across studies with two different diffusion scans acquired with the same TE. Inspecting the corresponding volumes, I noticed that the volume with the lowest series number always had important artifacts that compromised the readability, and that the one with the highest series number was of better quality. Dr. Ellen Grant later confirmed that a second diffusion scan with same sequence parameters as a previous one could have been acquired when the first one was of poor quality due to patient motion or other causes. For that reason, I decided to systematically select the second series (i.e. the one with the higher series number), when two series were acquired with similar parameters within one study.

To meet the goals of this project, only scans that included previously computed ADC, LOWB (b0 reference) and FA values were investigated. Unique series numbers, representative of the "old" scan parameters, were used to filter the data to only include scan obtained using the "old" diffusion protocol. These series were selected using the following series numbers:

```
ADC   : 101, 106, 111, etc. (+ multiple of 5)
LOWB  : 102, 107, 112, etc.
FA    : 104, 109, 114, etc.
```

After these considerations, I was finally able to identify and select all of the scans needed for further data analysis that fell under the final and most restricted selection criteria:

"Comparable (97.4 [ms] <= TE <= 108 [ms]) diffusion imaging data (ADC, LOWB and FA with corresponding specific highest series numbers) of patients between the ages of 0 to 24 months who don't present any major brain abnormality (i.e. whose corresponding health status = *normal*)"

This can be done using the enhanced Access database containing the health status tags and the series sequence parameters by linking *Demographic*, *Procedures*, *Scan Parameters* and *Health Status* tables and constraining the search with the criteria of interest. 255 different series fulfilling these criteria were obtained (see distribution in Figure 24). This Access database query generated the list of MRNs with corresponding dates and series numbers that was used to automatically retrieve the data from the NIfTI working directory and copy them to a separate directory used for data analysis.

All ADC and FA volumes, analyzed together for each subject, were visually inspected using the Freeview[31] visualization tool. The FA index, as opposed to the ADC index, is more sensitive to artifacts (such as motion) due to its directionality specificity (in contrast, the ADC index is simply the average diffusion in any direction). Thus, the FA volumes were first inspected, followed by the corresponding ADC volume. Volumes were discarded when structures usually seen on FA images could not be distinguished (due to excess blurring) or when important changes in brightness from slice to slice were noticed (probably due to motion artifacts). The corresponding ADC volumes in these cases were also often blurry (to a lesser extent for the reasons explained above), and were also discarded.

After this visual inspection, 61 series (FA and corresponding ADC volumes) had to be discarded due to insufficient quality or major artifacts. This resulted in a total number of 193 useable series (193 ADC volumes and 193 corresponding FA volumes) for our further analysis (see distribution Figure 24).

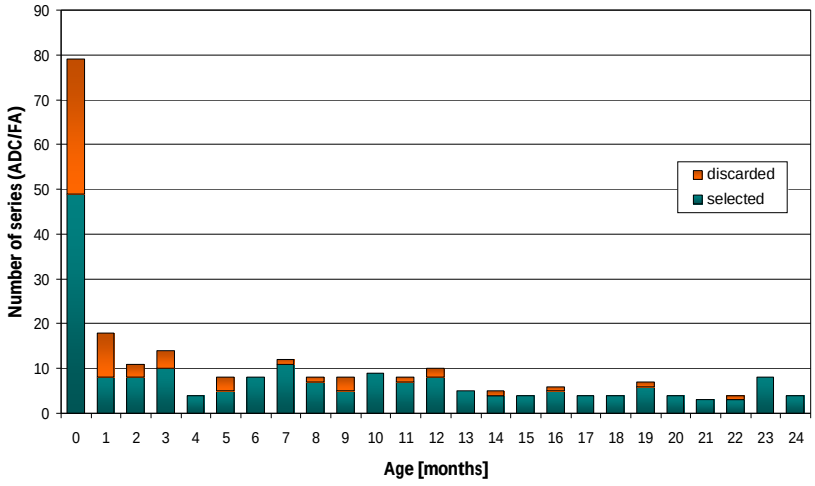


Figure 24: Total number of studies falling under data of interest criteria (selected + discarded) for each age range from 0 to 24 months. In orange, the number of series discarded for insufficient quality and in green the number of study selected for further analysis.

[31] http://surfer.nmr.mgh.harvard.edu/fswiki/FreeviewGuide/FreeviewIntroduction

A larger proportion of studies were discarded in the age ranges of 0 and 1 month. This was due to difficulties (patient motion, very small head size) encountered when scanning very young patients (newborns). Many studies were not available or not yet classified after the age of 12 months, explaining the reduction of the number of series from 13 to 24 months.

Finally, all these selected MR diffusion scans were acquired at MGH on a 1.5 Tesla scanner (GE Medical System) with the standard scanning sequence and with the following parameters: six different encoding directions with b-values of 0 and 1000 [s/mm^2]; TE = 97.4 - 108 [ms]; TR = 6000 or 7500 [ms]; matrix 128×128 [mm]; voxel size = 1.72×1.72, 1.56×1.56 [mm] or 1.41×1.41 [mm]. 23 slices (a few between 17 and 22) were acquired across the brain with thicknesses of 3-6 [mm] and gap of 1 [mm]. Calculations of eigenvalues, diffusion tensor matrices and corresponding ADC and FA were performed by the proprietary scan vendor software as part of the clinical image processing done by the department of radiology.

The slight differences in acquisition parameters across patients were due to specific adjustments made by the pediatric neuroradiology department faculty at MGH to optimize the quality of the MR images according to each patient. The variability was further exacerbated by the inevitable and desired changes due to scanner hardware and software upgrades.

Of the 2111 studies identified in the Access database from patients aged 0 to 24 months, we successfully retrieved 1646 studies from PACS using the mi2b2 software. Of these, 476 were categorized as normal. Selecting the data acquired with the old diffusion sequence parameters further reduced the number of studies to 273. Finally, the TEs selection criteria reduced this cohort to 255 comparable diffusion studies and 62 studies had to be discarded for quality issues. Thus our final data set was comprised of 193 comparable studies.

4.2. Time evolution of diffusion indices across ages

4.2.1. Data processing

To investigate the effect and effectiveness of the data processing (see part 3.3.1, Figure 16) in reducing artifacts and removing non-brain tissue on the distribution of WBA_{ADC} and WBA_{FA} values across ages, the WBA_{ADC} and WBA_{FA} values were plotted after each stage of image processing (Figure 25 & Figure 26, respectively).

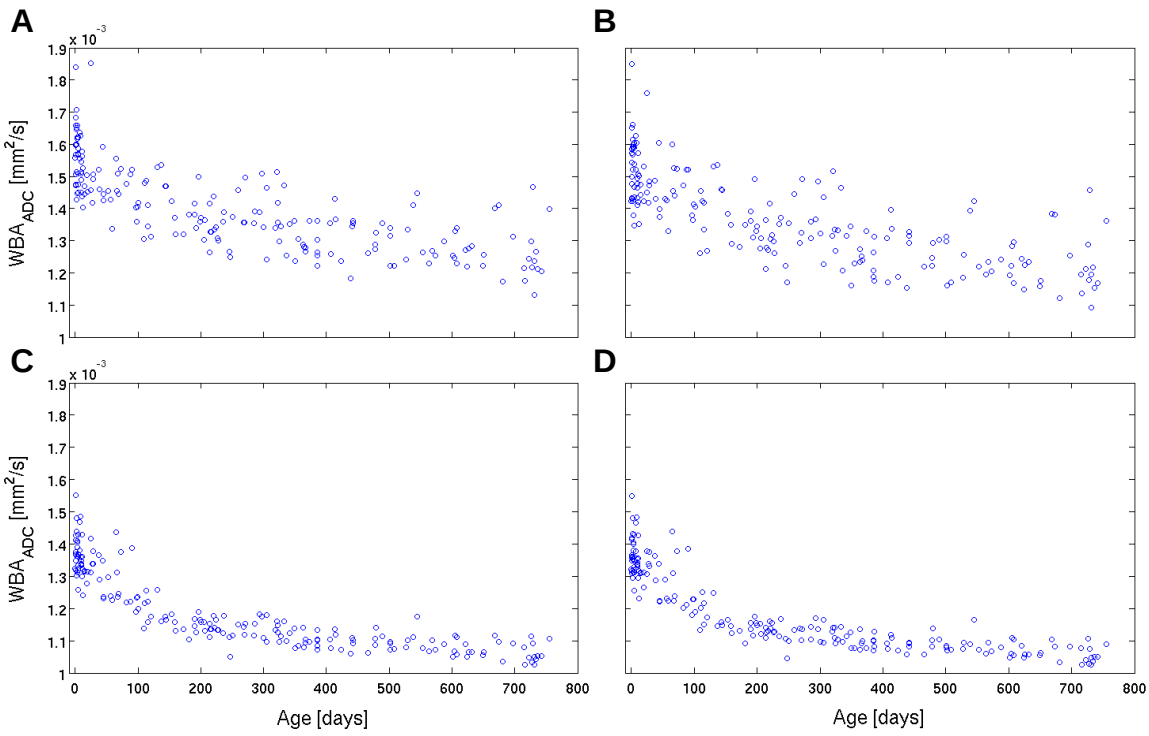


Figure 25: Distribution of WBA_{ADC} [mm^2/S] across age in days from non-processed data (A); after the brain extracted LOWB masking (step 2 of Figure 16) (B); after threshold application to remove CSF (step 3 of Figure 16) (C); and after BET application to ADC volumes (step 4 of Figure 16) (D).

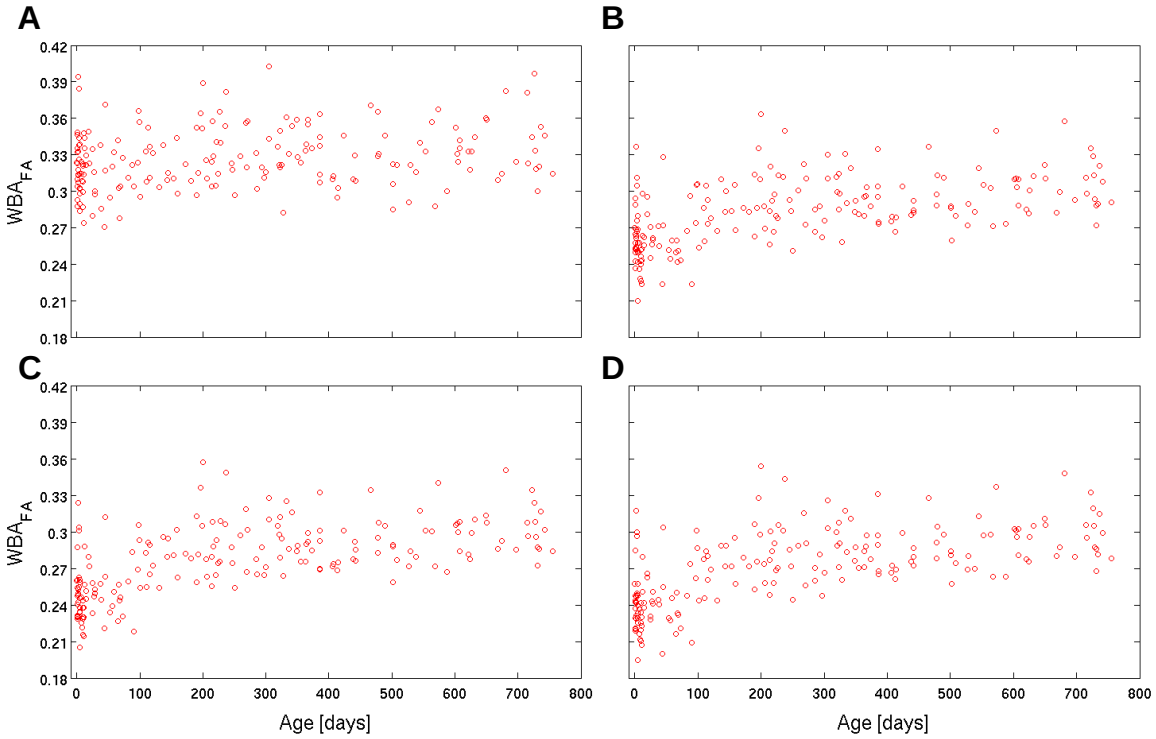


Figure 26: Distribution of WBA_{FA} across ages [days] from non-processed data (A), after brain extracted LOWB masking (step (2) applied to FA volumes, Figure 16) (B), after threshold application to remove CSF (step (3), ADC volume mask applied to corresponding FA volume, Figure 16) (C) and after fully processed ADC volume masking (step (6), Figure 16) (D).

As expected, WBA_{ADC} values seem to decrease whereas WBA_{FA} values seem to increase with age (quantitative analysis of the evolution will be performed in following chapters). We also note an overall decrease in WBA_{ADC} and WBA_{FA} values before (pre-) and after (pos-) the data were processed (pre-WBA_{ADC} average = 1.4×10^{-3} [mm^2/s], std = 1.304×10^{-4} and post-WBA_{ADC} average = 1.19×10^{-3} [mm^2/s], std = 1.228×10^{-4} ; pre-WBA_{FA} average = 0.327, std = 0.025 and post-WBA_{FA} average = 0.271, std = 0.031).

In addition we notice that data processing greatly reduced the variability of WBA_{ADC} and, to a lesser extent, WBA_{FA}. The most important variability reduction in WBA_{ADC} values occurred with the CSF removal step (Figure 25, C), the other steps had less influence. Standard deviations of WBA_{ADC} and WBA_{FA} values were calculated for each age range (0-24 months) before and after data processing (Figure 27, A-B). The difference in standard deviations before and after data processing was also determined (see Figure 27, C-D).

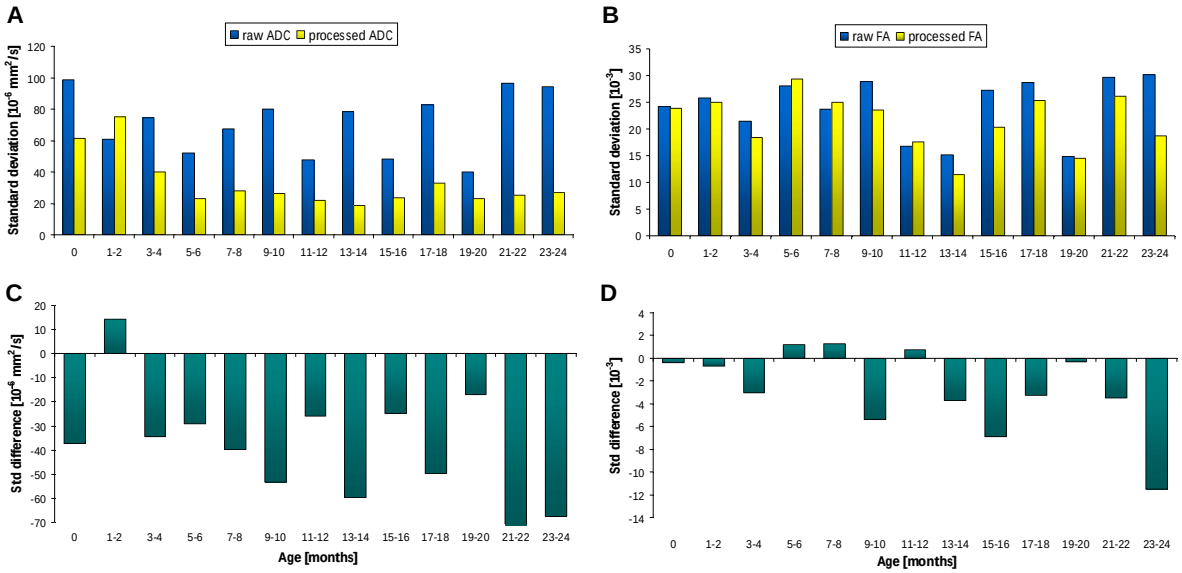


Figure 27: Standard deviation of WBA_{ADC} and WBA_{FA} values for each age range, (A) and (B) respectively before (blue) and after (yellow) data processing. Difference in standard deviation for each age range, (C) and (D), negative values represent a decrease in standard deviation after data processing.

As expected, the data processing tends to decrease the standard deviations of WBA_{ADC} values for each age range, except for 1-2 months (Figure 27, A, C) and to a lesser extent the standard deviations of WBA_{FA} values except for 5-6, 7-9, 11-12 and 19-20 age ranges (Figure 27, B, D). In addition, the standard deviation of processed WBA_{ADC} seems to decrease with age meaning that variability of these values is greater at an early age.

4.2.2. Curve fitting analysis

Three different curve fitting models (linear, logarithmic and biexponential) were applied to the data in order to see which function described the data best (see description part 3.3.2) by comparing the R^2 fit statistics for each model (see Figure 28 for WBA_{ADC} fitting and Figure 29 for WBA_{FA} fitting).

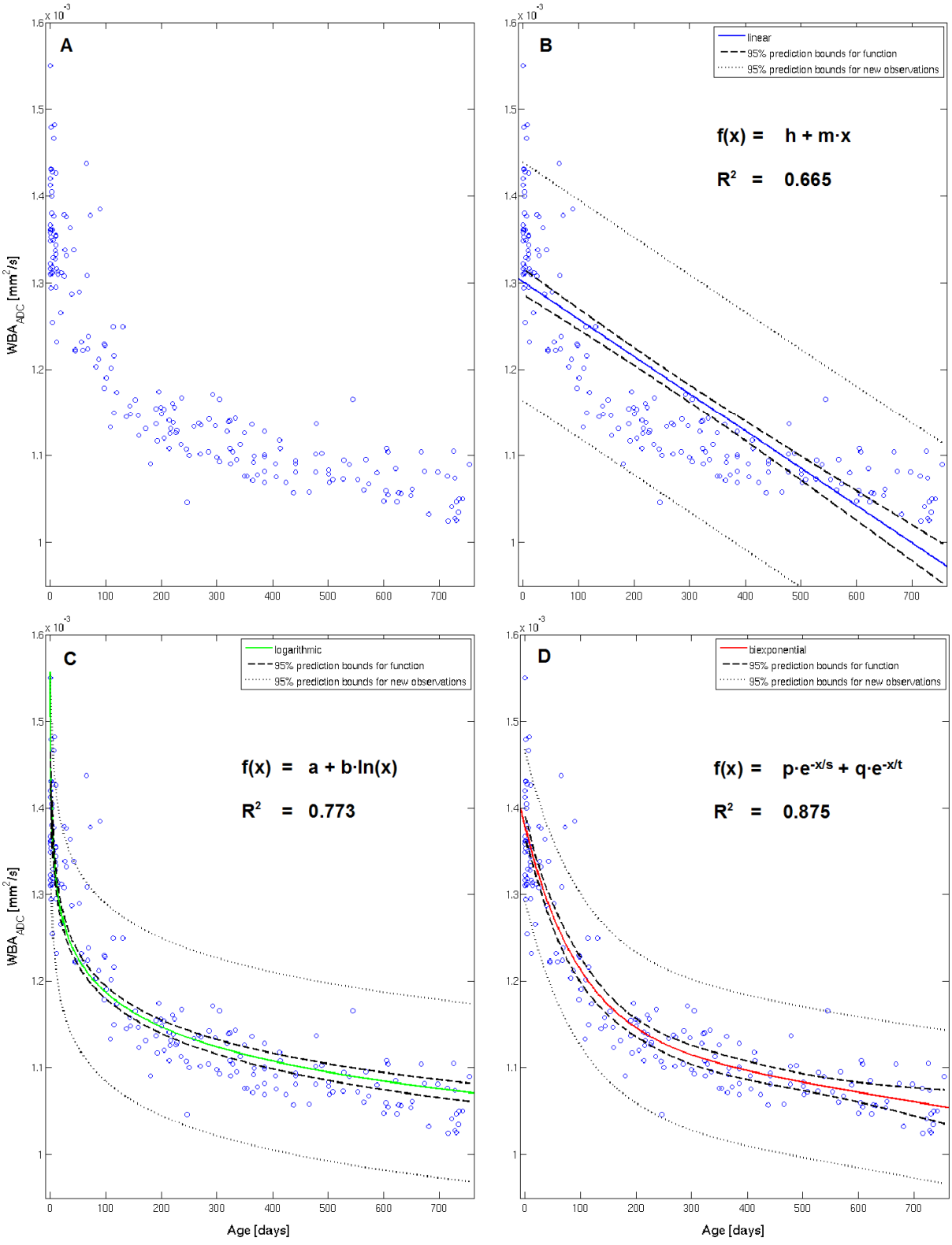


Figure 28: Curve fits for WBA_{ADC} values across ages. With no fit (A), with linear (B), logarithmic (C) and biexponential (D) fits. Prediction bounds at 95% confidence for the function (dashed lines) and new observations (dotted lines) and R^2 values are also displayed for each model. (Estimated parameters are available in Appendix 8.9)

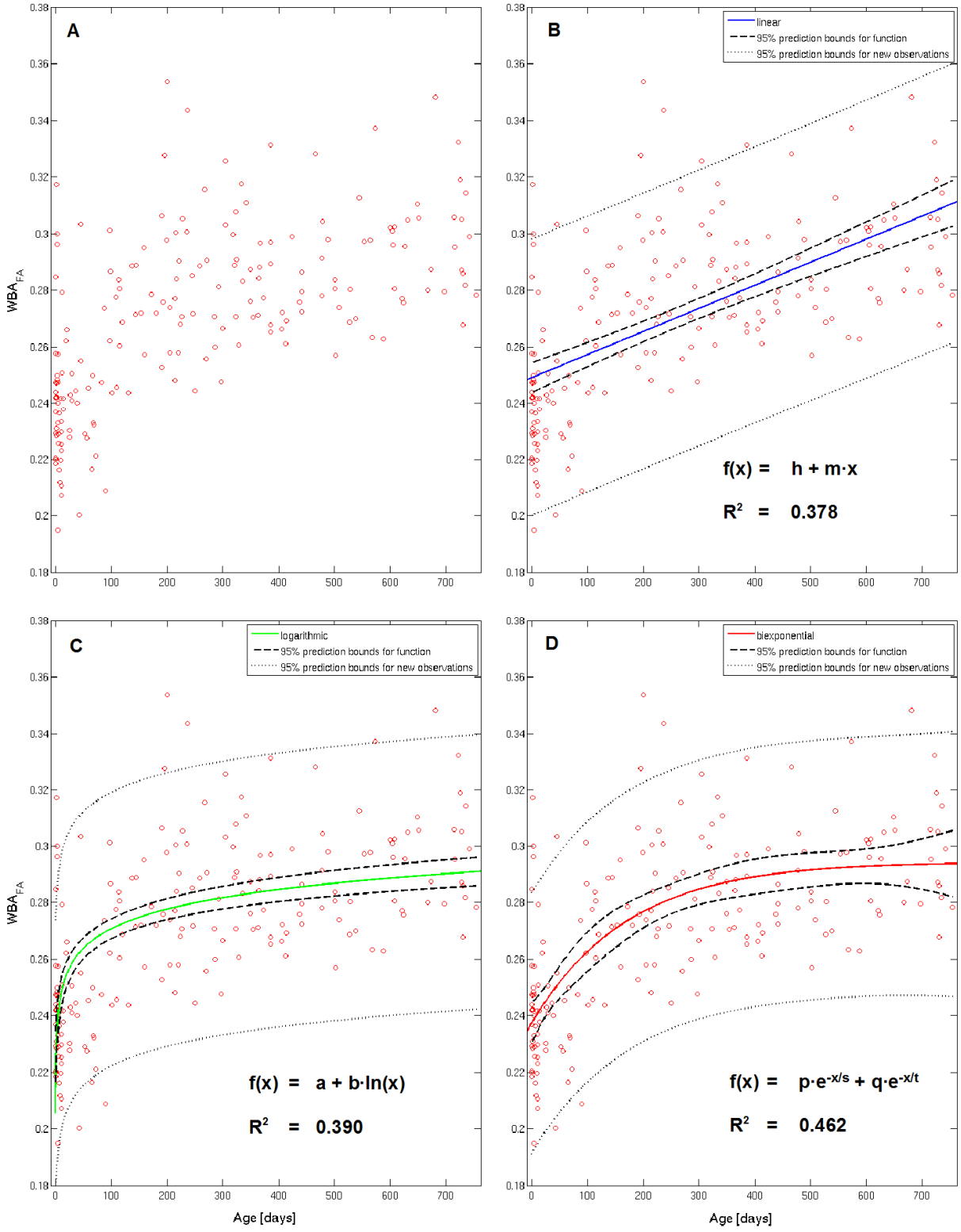


Figure 29: Curve fits for WBA_{FA} values across ages. With no fit (A), with linear (B), logarithmic (C) and biexponential (D) fittings. Prediction bounds at 95% confidence for the function (dashed lines) and new observations (dotted lines) and R^2 values are also displayed for each model. (Estimated parameters are available in Appendix 8.9)

From a preliminary visual inspection, we can already determine that the logarithmic and biexponential equations result in a better fit model for the time evolution of both WBA_{ADC} and WBA_{FA} values. This is confirmed by the R^2 values which are higher in these two models than in the linear one. The linear model, although revealing the main trend (decrease in WBA_{ADC} and increase in WBA_{FA} values), does not seem to reflect the shape of these evolutions. We also observe that in both cases, R^2 values are the highest in the biexponential model, suggesting that this model would best describe the time evolution of WBA_{ADC} and WBA_{FA} values across ages. In this model, we obtained an R^2 of 0.875 for WBA_{ADC} and 0.462 for WBA_{FA} values. This indicates that this fit explains 87.5% and 46.2% of variability about the average in the WBA_{ADC} and WBA_{FA} values, respectively.

Predictions bounds at 95% confidence for the functions and for new observations are also displayed (dashed and dotted lines respectively).

4.2.3. GLM analysis

The GLM was used to estimate the effects of Age and Gender on WBAADC and WBAFA. In this case, it is an Analysis of Covariance (ANCOVA) since we have a continuous outcome variable (either WBA_{ADC} or WBA_{FA}), one continuous predictor variable (Age in days) and one categorical predictor variable. As mentioned in part 3.3.2, Age and WBA values were first log-transformed in order to improve linearity of the age-WBA relationship. We note that the linear curve fits have higher R^2 values when performed on the transformed data (Figure 30) than the original data (Figure 28, B and Figure 29, B).

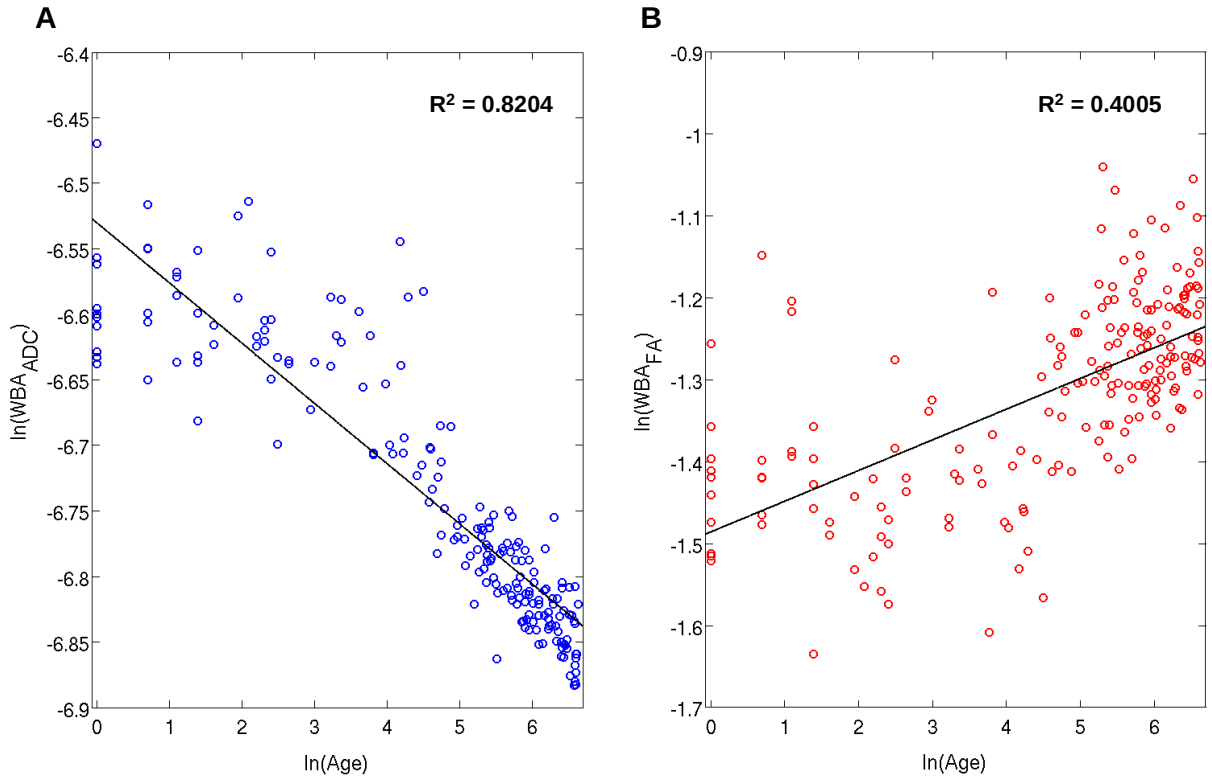


Figure 30: Natural logarithm of WBA_{ADC} (A) and WBA_{FA} (B) versus natural logarithm of Age. Black lines represent linear fits with the corresponding R^2 values.

As expected, the GLM revealed a statistically significant effect of Age on WBA_{ADC} values, $F_{(1,189)} = 895.0$, $p < 0.0001$, and on WBA_{FA} values, $F_{(1,189)} = 126.7$, $p < 0.0001$. In addition, a significant effect of Gender (Female, N=82 and Male, N=111) on WBA_{ADC} values was observed, $F_{(1,189)} = 8.7$, $p = 0.0038$, as well as a significant interaction between Gender and Age, $F_{(1,189)} = 5.2$, $p = 0.0236$. This indicates that Gender significantly influences the way WBA_{ADC} values evolve with Age (i.e. the slope of the curve in such a linear model). An examination of Figure 31A reveals that WBA_{ADC} values seem to be slightly higher in Males than in Females at an early stage, but become more similar with increasing age.

No significant effects of Gender on WBA_{FA} values and no significant interactions between Gender and Age were observed, $F_{(1,189)} = 1.15$, $p = 0.2859$ and $F_{(1,189)} = 0.28$, $p = 0.5973$, respectively (Figure 31, B).

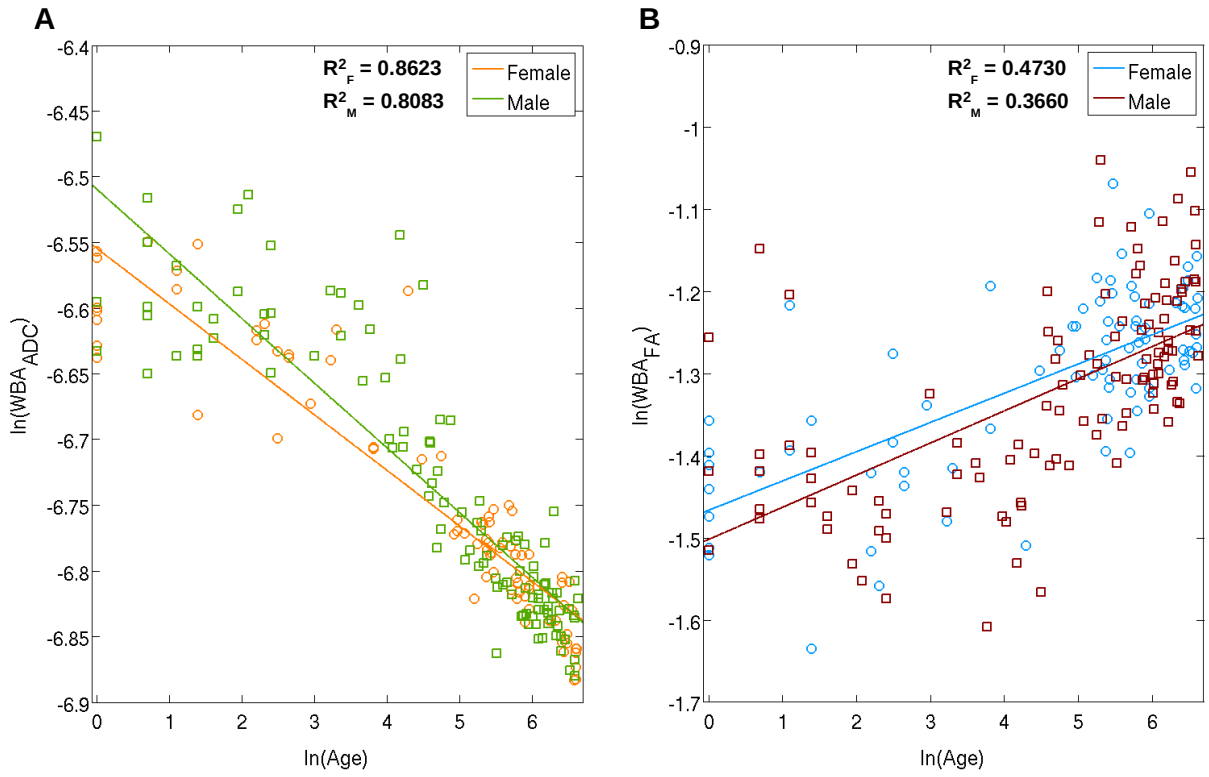


Figure 31: Natural logarithm of WBA_{ADC} (A) and WBA_{FA} (B) versus natural logarithm of Age separated according to Gender (Female orange and blue circles, Male green and red squares for WBA_{ADC} and WBA_{FA} respectively). Corresponding color-coded linear fits with corresponding R^2 values are also displayed.

4.2.4. Within patient WBA ADC/FA time evolution controls

As explained previously, we also looked for patients with multiple scans that could be used as controls for within patient time evolution of WBA_{ADC} and WBA_{FA} values. Four individual patients were found each of whom had multiple diffusion MRI scans acquired at different ages. Three of them were followed for non-brain tumors (retinoblastoma) and one for hypoglycemic seizures. In all cases, there was no reported imaging evidence of abnormal brain development in the corresponding radiology reports. These patients were highlighted in WBA_{ADC} and WBA_{FA} value plots and images of one index patient were also included (see Figure 32 & Figure 33).

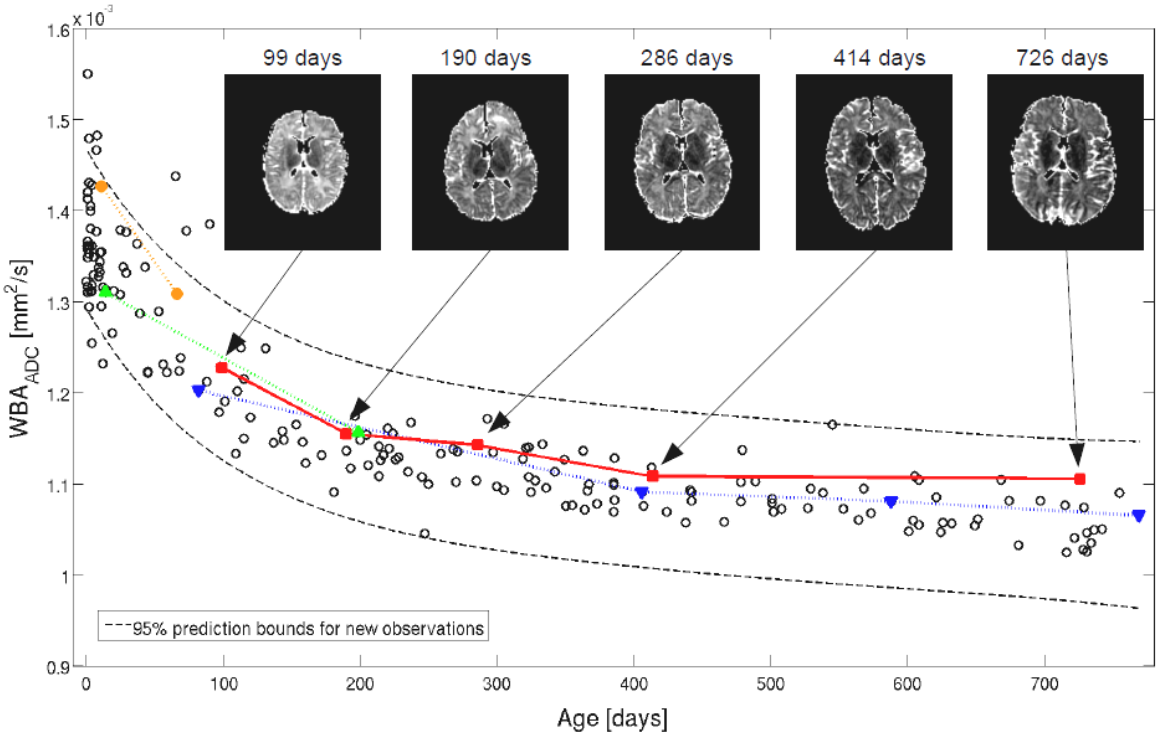


Figure 32: Four patients with multiple scans are highlighted in orange (11 and 66 days), green (44 and 199 days), blue (82, 406, 588 and 769 days) and red (99, 190, 286, 414 and 726 days). Corresponding processed ADC images of the patient displayed in red color are included. The 95% prediction bounds for new observations in the biexponential model are also displayed (dashed lines).

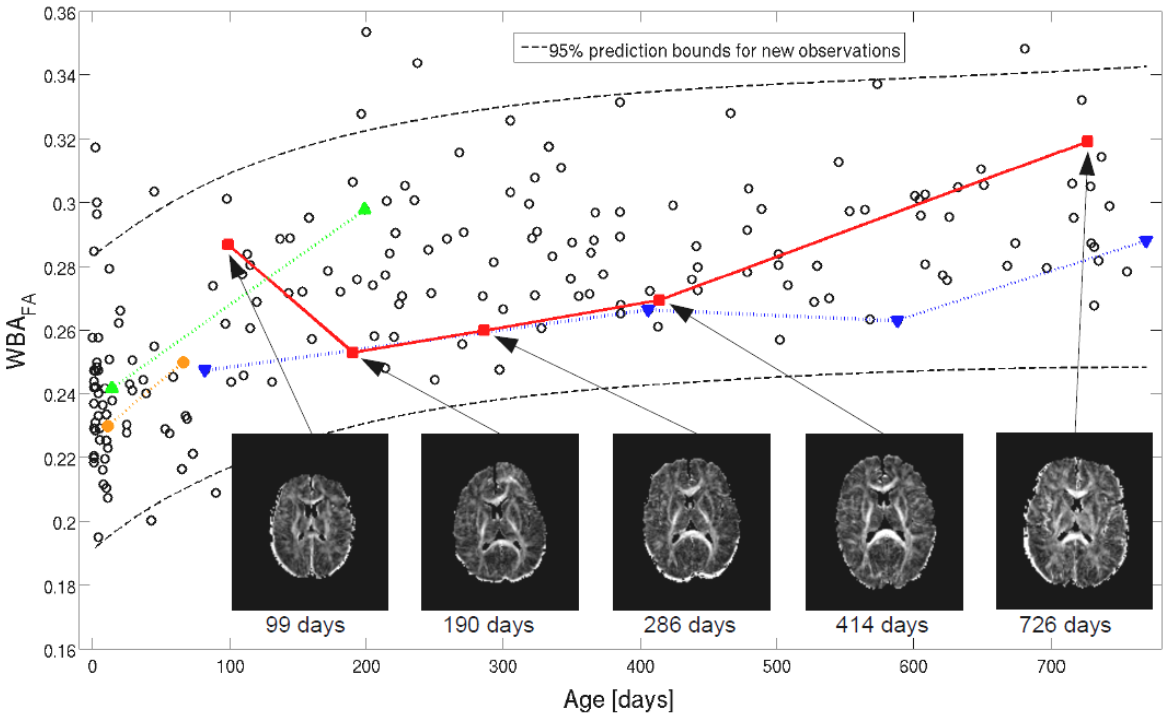


Figure 33: Four patients with multiple scans are highlighted in orange (11 and 66 days), green (44 and 199 days), blue (82, 406, 588 and 769 days) and red (99, 190, 286, 414 and 726 days). Corresponding processed FA images of the patient displayed in red color are included. The 95% prediction bounds for new observations in the biexponential model are also displayed (dashed lines).

For both WBA_{ADC} and WBA_{FA}, we observe an overall decrease and increase respectively, of these values within patients across ages (with the exception of the “red” and “blue” patients whose WBA_{FA} values decrease between the age of 99-190 days and 406-588 days respectively). Consistent with the best fit (biexponential model) and the literature in the field, the rates of change are most rapid during the first weeks of life (see Figure 32 & Figure 33, orange and green for example) as assessed by the slope of those lines.

Changes in ADC and FA images can be noticed even visually with a darkening trend in ADC images, especially in white matter (see Figure 32), and a prominent evolution of fiber bundle delineation in FA images (see Figure 33).

5. Discussion

5.1. Database mining

The first part of this project was dedicated to database mining with the aim of identifying and retrieving biomedical MRI data from MGH PACS that met the following criteria:

"All comparable diffusion imaging data of patients between the ages of 0 to 24 months without any major brain abnormality"

I started this data mining expedition with 5278 potential patients with brain MRI scans from RPDR finished with "only" 193 studies in the final cohort (less than 3.7% of the total). This number might seem very low, but actually resulted in a cohort size that is significantly larger and more evenly distributed across the age range than what has been available so far in similar research in this same age range (see 2.2). Moreover, the time required to acquire the imaging data, from time of IRB approval to retrieval from the PACS was about 3 months worth of work (without accounting for mi2b2 troubleshooting time), including the time required for data mining. Most importantly, this valuable data was accessed at the costs of the data storage and personnel time; a vanishingly small fraction of the costs associated with a prospective study that aimed to collect this data set by recruiting and scanning healthy babies. This was only possible using biomedical data collected at MGH over the last decade and tools that were developed to retrieve those data.

In this pilot project, I demonstrated the possibility for researchers to use available tools such as RPDR and its Access database to identify any data of interest in order to retrieve them from MGH PACS with the help of the newly developed mi2b2 software. This infrastructure has been intensively developed for the past 2 years and the Production version of the mi2b2 software will be released to the Partners user community in the Fall of 2011.

As a new (and first) user of the mi2b2 software, I was able to investigate the possibilities for the use of this software once the imaging data were retrieved from the hospital database. My efforts have greatly helped the developers to design a user interface for mi2b2 that facilitates retrieval and post-retrieval investigation. In fact, as I was retrieving increasing numbers of MRI data sets from PACS, I realized early on that it would not be possible to work with such a large quantity of image data sets in their original format. Therefore, I developed the BB-pipeline that automatically organizes and reformats the data retrieved from PACS through mi2b2 in a way that facilitates identification, selection and further work on any data of interest. Certain functionalities of this pipeline, including Rudolph's (see People section 8.3) system could be included in the mi2b2 software, allowing the users to choose the output format of the data by including, MRNs, study dates, ages, etc., in the directory/file names. The automated DICOM header extraction regarding technical scan details used for further data filtering could also be used in future versions of the mi2b2 user interface to make life easier for the researcher. This was discussed with the mi2b2 team during my project and was proposed in the RO1 grant application submitted in June 2011 (cf. Introduction):

"We will also develop a method to enhance the usage of the DICOM metadata present in the DICOM headers. An automated process that extracts patient-, study-, and image-related fields from the DICOM header will contribute this information to a relational database where it too may be queried. This will improve our filtering ability to select scans based upon specific details of scan acquisition."[32]

One main drawback of the pipeline that I've been developing was the use of an Access Database. As discussed previously, there was no direct connection between the imaging data sitting on my Linux workstation and the medical and technical information included in the PC based Access Database on my laptop computer. Hence, a manual transfer of text files was necessary to connect the two, a suboptimal method. A potential way to improve this system would be to set up a Linux compatible relational database management system (e.g. MySQL) directly on the workstation removing the inefficient manual step of text file transfer. Having such a set up would also make it possible to design a simple script/software that would take any criteria as input (medical or technical criteria, such as the ones described in this thesis) from a user and automatically retrieve the corresponding data. This would eliminate the burden of selecting data in the Access Database, extracting a text file with corresponding MRN or other identifiers, and manually transferring it to a workstation where a specific script reads them and retrieves the corresponding images. Additionally, any other useful features (such as imaging data format) could also be included in this script that would perform these functions automatically for the user. Developing such software would greatly enhance the manipulation of the data and facilitate the work of potential users.

Having emphasized the importance of automating as many steps as possible of the database mining process, one particular step remained entirely manual. It was necessary to identify normal patients by reading through radiology reports. As we observed in part 3.2.2.4, radiology reports contain verbally dictated notes of the responsible radiologist that are not standardized and vary from case to case. Because of this, the only option was to read through each report individually. This process was extremely time consuming and I took several weeks to comb through the records of more than 1500 patients. Developing an automated way of detecting normal cases would, in consequence, be a very helpful addition to the data mining process. One way of tackling this problem would be to use machine learning methods, using the complete list of key sentences collected in normal cases (see part 3.2.2.4) as training data. Nick Murphy, graduate student at Harvard University, is currently testing a machine learning method[33] on the data that I have already classified. He was able to get the system to correctly predict the classification of patient ("normal" or "abnormal") about 70% of the time. The success of this preliminary program relied on previously manually classified studies, but it could be helpful for further identification of normal cases with some additional work.

Another issue that has to be addressed is the *comparability* selection criteria that we applied to our diffusion data. As described in the Background chapter (part 2.1), quantitative metrics from

[32] RO1 grant application submitted in June 2011 requesting funds to further for the development of mi2b2 and to develop a novel Harvard Catalyst Radiological Decision Support (RDS) Toolkit
[33] http://www.cs.waikato.ac.nz/ml/weka/

diffusion images depend on scan sequence parameters and consequently, only images with similar parameters can be directly compared without additional image post-processing or more complex statistical models. This reduced our "normal" patient cohort by 43% (from 476 to 273) when we decided to only focus on data acquired before the year 2006 (i.e. with the "old" diffusion sequence protocol). The TEs selection criteria further reduced this cohort by 7% to 255 comparable diffusion studies of normal patients. Finding a way of including all the studies in the same analysis would greatly improve its power by having at least twice as many studies. White et al., 2011, presented a study on global white matter abnormalities in schizophrenia with diffusion tensor imaging data acquired at four different institutions with different sequence parameters (Field strength, TE, TR, b-values, diffusion directions, etc.). As expected, their work revealed a significant site difference in FA values. To reduce the effect of these site-related differences, they notably corrected their statistical analysis by using site as a covariate in their model (White et al., 2011). This approach would be one way to include the rest of our data in our statistical analysis, and increasing its power. Similar studies have been conducted to examine the impact of image acquisition variables on different structural data analysis (Jovicich et al., 2009) and the feasibility of using multisite data sets to investigate questions of scientific relevance by carefully taking into account Site in statistical analysis (Fennema-Notestine et al., 2007).

5.2. Data analysis

Despite the considerable amount of data we had to reject, we still ended up with 193 studies from patients with "normal brains" aged 0 to 2 years. This allowed us to investigate the cross sectional time evolution of two different diffusion imaging biomarkers, to wit, WBA_{ADC} and WBA_{FA} values.

The first step in computing these values was to remove any non-brain regions from all the volumes. The large number of volumes involved, and the limited time I had at my disposal meant that no manual editing options were feasible. This factor drove the decision to use semi-automated methods. The automated pipeline described in part 3.3.1, which I developed with advice of our team members (see People section 8.3), worked well on most of the data for these purposes, as assessed by visual inspection as well as quantitative assessment. As a result, we observed an overall decrease of variability within age specific patient groups by removing non-brain regions and other artifacts (cf. Figure 25, Figure 26 & Figure 27). The overall decrease of WBA_{ADC} values is likely due to the application of the threshold to remove CSF regions, which have the highest ADC values. We also observed a similar overall decrease in WBA_{FA} values which seems at first glance contradictory. Indeed, FA values in CSF are supposed to be very low due to its non-directional water diffusion property and its removal should increase WBA_{FA} values. In fact, this drop in WBA_{FA} does not occur after the threshold application step but rather after the first step of BET application to LOWB used as a mask on corresponding FA volumes (Figure 26, A-B). Indeed, a lot of artifacts with very high FA values were noticed outside of brain regions. As a consequence, removing them during the first step lowered the overall WBA_{FA} values (BET application on LOWB tend to remove structures outside brain regions).

In addition, the main reduction of variability for each age range seems to occur in accordance with the last observations, namely after CSF removal in ADC volumes and after the LOWB BET masking on FA volumes. This suggests that the main source of variability among ADC volume

resides in CSF-related regions, whereas in FA volume this variability would be due to artifacts found outside of brain regions.

Looking at some fully processed FA volumes (e.g. FA images presented in Figure 16 & Figure 33), some very bright (high FA values) non-brain regions appear along the posterior edge of the brain. These could partially explain the higher variability among WBA_{FA} values, even after full processing. Additional steps in the processing pipeline to remove those artifacts could be explored to further reduce this source of variability.

The curve fitting analysis revealed that both the age dependent evolution of WBA_{ADC} and WBA_{FA} values seem to be described better by biexponential models (as suggested by Muhkerjee et al., 2001, for several ROIs) than by linear or logarithmic models comparing the goodness of fits indicated by the R^2 values (see Figure 28 & Figure 29). In fact, the biexponential models might be more adequate than the other ones in biologically interpreting the time evolution of both WBA_{ADC} and WBA_{FA}. Indeed, it actually takes into account a two phase model with two different terms in the model, a fast component (first term in the equation: $p*\exp(-x/s)$) and slow component (second term in the equation: $q*\exp(-x/t)$) according to the estimated parameters which reflects the fast decay at an early age follow by a slower decay in WBA_{ADC} values and a fast increase followed by a slower increase in WBA_{FA} values. This might be a way of interpreting the underlying biological process of myelination that is known to undergo greatest change within the first months of life (cf. Background, part 2.2).

No model, as good as it can be, can ever be the true representation of the data since Mother Nature did/does not implement them as such in living beings. We can only try to find a model that fits the data as well as possible in order to obtain a quantitative estimation of how a biological process evolves (in our case, how our two markers, WBA_{ADC} and WBA_{FA} evolve with age), and maybe predict where a new observed value should be for a specific range (see Figure 28, Figure 29). We investigated this by the using prediction bounds for new observations that give the range of WBA_{ADC} or WBA_{FA} values at each age within which the WBA_{ADC} or WBA_{FA} value for a new individual of that age, taken from this population, should be determined with 95% certainty. Having such age-specific markers with prediction bounds could be very valuable when comparing a new patient, with unknown health status, to the predictions bounds of the corresponding age and determining whether he/she fits into this range. If this patient is normal (as determined in our population cohort), we have 95% chance that he/she is actually included within the range of the prediction bounds. If not, this could be a first clue that something might developing abnormally compared to normal patients of the same age.

GLM analysis indicated a significant effect of Age on both WBA_{ADC} and WBA_{FA}. These results were largely expected, as it is well known that brain development implies widespread structural changes, including myelination of the white matter bundles. Intriguingly, our results also revealed significant effects of Gender on WBA_{ADC} and a significant interaction between Age and Gender. This finding suggests that even at a very early stage (first weeks after birth), the brain (as assessed by the WBA_{ADC} marker) may evolve differently between Males and Females. This observation has important implications in the creation of developmental markers from identified normative cases since those markers, in addition to being age-specific, should also probably be gender-specific if appropriate comparisons are to be made with clinical cases.

The inclusion of within patient longitudinal data controls was particularly valuable when comparing their individual age-evolution with the cross-patient age evolution of WBA_{ADC} and

WBA_{FA} since they were not affected by inter-patient variability. The four longitudinal patients exhibited similar trends to those observed across patients, with a rapid decrease of WBA_{ADC} values at an early age (see Figure 32, orange, green and first time points of red colored patient) followed by a slower decay (see Figure 32, last time points of red and blue colored patient). The within patient WBA_{FA} values displayed less obvious results, particularly for the patient whose data is shown in red (Figure 33), where a decrease was observed from 99 to 190 days instead of the expected increase, as the main trend suggests. This observation may have been the result of artifacts and greater sensitivity of FA values that might have biased the true FA values of the 99 days patient volumes, rather than reflecting a real biological process. However, we also detected more rapid changes at an early stage, especially in the patients identified in orange and green (Figure 33). Investigating within patient evolution of markers such as WBA_{ADC} and WBA_{FA} is crucial to be able to compare a new patient of unknown health status with a cohort of normal patients. Indeed, the fact that the markers for this particular patient lie within the prediction ranges does not say anything about their relative evolution for the individual patient. Actually, even though specific makers could seem to be within the normal range, their relative position to other time points within the same patient could reveal an abnormal evolution. Obviously, such data, in a real clinical situation, do not always exist, but if they are available, they could be of significant value. This highlights the importance of investigating markers of evolution not only across patients, but also within individual patients, because these markers could offer an even more reliable approach for detecting abnormal cases.

We have to keep in mind that our two markers (WBA_{ADC} or WBA_{FA} values) are not very specific since they were obtained by calculating the whole brain average of ADC and FA values for each individual patient. Therefore, all the information contained in the ADC or FA volume (e.g. region, size, contrast, etc.) is condensed in a unique value for both ADC and FA. This one value is not specific enough to determine subtle time-specific evolution of particular brain regions of different sizes, for instance. This may also explain the high variability we obtained for our markers, especially in WBA_{FA}, since FA values are more specific than ADC volumes by taking into account a dimension of directionality that ADC values do not (ADC values are simply the average of the diffusion). In order to obtain more robust and specific markers, with less variability across patients than what we obtained, region-specific ADC and FA averages have to be determined.

Several studies have investigated the region-specific time evolution of ADC and FA indices in children (see part 2.2, Mukherjee et al., 2001 ; Hermoye et al., 2006 and Löbel et al., 2009). They observed similar trends with an overall decrease of ADC values and increase in FA for specific ROIs across ages, encountering the greatest change at an early age. The range of values we obtained for our markers, although we obtained only whole brain information, is comparable to the range obtained in these studies (ADC values starting around 1300 [mm^2/s] and rapidly decreasing above 1000 [mm^2/s], and FA values starting from 0.2 rapidly increasing above 0.3). However, these previous studies only included a few subjects (from around 10 to 30) aged 0 to 2 years in comparison to the number of subjects we had (193 patients). Hence, at the cost of region specificity, we gained significant age resolution, especially in the range from 0 to 200 days, where our markers indicate the most dramatic changes. The power of this significant number of patients allowed us to track changes occurring in the brain, even with simple markers that probably would not have been possible to detect with fewer patients.

Obviously, investigating the time evolution of specific ROIs with the group of patients we have at our disposal would be even more interesting. Unfortunately due to the delay of my first access to the full cohort due to development of the mi2b2 software and the time limitations of my fellowship, I was not able to envisage any specific manual ROI determination within the allotted time. The ideal situation would be to have access to different tools that could automatically segment the brains of patients aged 0 to 2 years. Unfortunately such tools that could work on newborns and infants do not exist at this time, and dedicated researchers (such as Lilla Zollei, PhD, among others) are putting all their effort in developing and making them available. An alternative way would be to develop a robust normalization procedure (similar to the ones mentioned in Background section, part 2.3) that could align several volumes in an age-specific standard effectively. Doing so, manual delineation of ROIs would only be necessary in the standard spaces and could be directly applied to the registered volumes. One of the aims presented in the RO1 grant application specifically targets these issues.

The image data used in this pilot project were of relatively poor quality, sensitivity and resolution in comparison to what can be currently found in other leading institutions specialized in children's care such as Children's Hospital Boston (CHB) where Dr. Ellen Grant, MD, is the Director of the Fetal-Neonatal Neuroimaging and Developmental Science Center. Current MRI data acquired there notably involve a 3T scanner with 2mm isotropic DTI and 35 gradient directions. Additionally, CHB routinely scans more patients in this age range than MGH (in the year 2010 alone, the Department of Radiology archived brain MRI scans acquired for routine clinical care in 626 patients aged 0-1 years and 1,019 patients aged 1-4 years). The aims of the recently submitted R01 grant focus on applying all the lessons learned from this pilot project to acquiring and analyzing these higher resolution pediatric MR brain scans.

Ideally, new types of sequences should be developed and applied to the study of brain development, in order to investigate more subtle processes. In fact, as briefly mentioned in part 2.2, DTI only gives information about the *diffusivity of water molecules* within the brain which can be influenced by multiple factors such as axonal density and size, fiber tract coherence, or membrane structure and permeability (Beaulieu, 2002), and not only by myelination as it is sometime interpreted. Therefore, if we want to study more precise developmental processes, like myelination, more specific MRI sequences have to be performed. A recent study from Deoni et al. provided "the first ever *in* vivo visualization of myelin maturation in healthy human infancy" (Deoni et al., 2011) using a new myelin-specific MRI technique. Although promising, it may take a long time before such techniques are available on the clinical scanners where our data come from.

6. Conclusion

In this pilot project, I demonstrated, with the collaboration of researchers from multiple sites, the feasibility of identifying, retrieving, and analyzing pediatric clinical multimodal MRI data, using available tools and the prototype mi2b2 software. The aim of this project was to study normal human brain development from birth through 2 years of age as indexed by diffusion MRI.

The huge amount of data retrieved from PACS through mi2b2 (more than 30,000 series from 1646 studies) rendered any manual organization, identification and selection of particular series impossible and necessitated the development of a pipeline that could handle this volume of data. Therefore the BB-pipline was created to automatically receive studies from mi2b2 software, organize and convert them in suitable format, extract corresponding technical scan information and update a log file used to help the mi2b2 development team to keep track of potential issues.

Starting from more than 5000 patients in RPDR with potential diffusion MRI scans acquired after the year 2000, 193 studies meeting all the necessary criteria (age, health status, MRI modality, comparability and quality) were identified and used for further investigation.

Two markers were computed after an automated procedure to remove non-brain tissues and artifacts, developed for the sake of this project. GLM and curve fitting analysis revealed a significant decrease and increase of WBA_{ADC} and WBA_{FA} values, respectively, across ages that seem to evolve differently according to gender. Moreover, these age evolutions appear to follow biexponential models with a higher change rate at an early age. Using prediction bounds for such age-specific markers could be very valuable when comparing a new patient, with health status to be determined, to the predictions bounds of the corresponding age and determining whether he/she fits into this range.

The significant size of our patient cohort enabled us to detect subtle changes at an early stage, even using simple markers that do not contain region specific information. Further work will be necessary to design new markers addressing this issue. Additionally, quality, resolution, sensitivity of the data, and quantity could be further improved using the newest data acquired in other institutions.

All these concerns have been addressed in the R01 grant application submitted last June and if funded (I'm sure it will), will be tackled by dedicated researchers next year. I can leave knowing this project is in good hands, and confident of its success.

This project has been extremely fruitful for my personal experience. Starting last September, I did not have much practical skills in the neuroscience field. Almost one year later, I have gained an insight into many different facets of the research world, including database management, different programming languages, clinical practices and politics, MRI research, and much more. This opportunity has helped me learn new skills, and explore clinical translational neuroimaging research. I have realized that, more than pure theoretical knowledge, EPFL provided me the necessary tools to adapt to this multi-disciplinary environment, that have been integral to the success of my project.

Thanks to all the people who helped me in this great adventure.

7. References

7.1. Papers

Alexander, A.L., Lee, J.E., Lazar, M. and Field, A.S. (2007). Diffusion tensor imaging of the brain. Neurotherapeutics *4*, 316-329.

Almli, C.R., Rivkin, M.J., McKinstry, R.C. (2007). The NIH MRI study of normal brain development (Objective-2): newborns, infants, toddlers, and preschoolers. Neuroimage *35*, 308-325.

Altaye, M., Holland, S.K., Wilke, M. and Gaser, C. (2008). Infant brain probability templates for MRI segmentation and normalization. Neuroimage *43*, 721-730.

Asato, M.R., Terwilliger, R., Woo, J. and Luna, B. (2010). White matter development in adolescence: a DTI study. Cereb. Cortex *20*, 2122-2131.

Bartha, A.I., Yap, K.R.L., Miller, S.P., Jeremy, R.J., Nishimoto, M., Vigneron, D.B., Barkovich, A.J. and Ferriero, D.M. (2007). The normal neonatal brain: MR imaging, diffusion tensor imaging, and 3D MR spectroscopy in healthy term neonates. AJNR Am J Neuroradiol *28*, 1015-1021.

Basser, P.J., Mattiello, J. and LeBihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. J Magn Reson B *103*, 247-254.

Beaulieu, C. (2002). The basis of anisotropic water diffusion in the nervous system - a technical review. NMR Biomed *15*, 435-455.

Deoni, S.C.L., Mercure, E., Blasi, A., Gasston, D., Thomson, A., Johnson, M., Williams, S.C.R. and Murphy, D.G.M. (2011). Mapping infant brain myelination with magnetic resonance imaging. J. Neurosci. *31*, 784-791.

Evans, A.C. (2006). The NIH MRI study of normal brain development. Neuroimage *30*, 184-202.

Faria, A.V., Zhang, J., Oishi, K., Li, X., Jiang, H., Akhter, K., Hermoye, L., Lee, S., Hoon, A., Stashinko, E., et al. (2010). Atlas-based analysis of neurodevelopment from infancy to adulthood using diffusion tensor imaging and applications for automated abnormality detection. Neuroimage *52*, 415-428.

Fennema-Notestine, C., Gamst, A.C., Quinn, B.T., Pacheco, J., Jernigan, T.L., Thal, L., Buckner, R., Killiany, R., Blacker, D., Dale, A.M., et al. (2007). Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data. Neuroinformatics *5*, 235-245.

Gollub, R.L. and Turner, D.A. (2010). Neuroinformatics in clinical and translational medicine--novel approaches. Neuroinformatics *8*, 207-212.

Hagmann, P., Jonasson, L., Maeder, P., Thiran, J., Wedeen, V.J. and Meuli, R. (2006). Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. Radiographics *26 Suppl 1*, S205-23.

Hart, R. and Belotto, M. (2010). The institutional review board. Semin Nucl Med *40*, 385-392.

Hermoye, L., Saint-Martin, C., Cosnard, G., Lee, S., Kim, J., Nassogne, M., Menten, R., Clapuyt, P., Donohue, P.K., Hua, K., et al. (2006). Pediatric diffusion tensor imaging: normal database and observation of the white matter maturation in early childhood. Neuroimage *29*, 493-504.

Huang, H.K. (2011). Short history of PACS. Part I: USA. Eur J Radiol *78*, 163-176.

Hüppi, P.S. and Dubois, J. (2006). Diffusion tensor imaging of brain development. Semin Fetal Neonatal Med *11*, 489-497.

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., et al. (2009). MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage *46*, 177-192.

Kazemi, K., Moghaddam, H.A., Grebe, R., Gondry-Jouet, C. and Wallois, F. (2007). A neonatal atlas template for spatial normalization of whole-brain magnetic resonance images of newborns: preliminary results. Neuroimage *37*, 463-473.

Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N. and Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. J Magn Reson Imaging *13*, 534-546.

Löbel, U., Sedlacik, J., Güllmar, D., Kaiser, W.A., Reichenbach, J.R. and Mentzel, H. (2009). Diffusion tensor imaging: the normal evolution of ADC, RA, FA, and eigenvalues studied in multiple anatomical regions of the brain. Neuroradiology *51*, 253-263.

Miller, J.H., McKinstry, R.C., Philip, J.V., Mukherjee, P. and Neil, J.J. (2003). Diffusion-tensor MR imaging of normal brain maturation: a guide to structural development and myelination. AJR Am J Roentgenol *180*, 851-859.

Mori, S. and Zhang, J. (2006). Principles of diffusion tensor imaging and its applications to basic neuroscience research. Neuron *51*, 527-539.

Mukherjee, P., Miller, J.H., Shimony, J.S., Conturo, T.E., Lee, B.C., Almli, C.R. and McKinstry, R.C. (2001). Normal brain maturation during childhood: developmental trends characterized with diffusion-tensor MR imaging. Radiology *221*, 349-358.

Murphy, S.N., Mendis, M., Hackett, K., Kuttan, R., Pan, W., Phillips, L.C., Gainer, V., Berkowicz, D., Glaser, J.P., Kohane, I., et al. (2007). Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc , 548-552.

Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S. and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc *17*, 124-130.

Ogura, A., Hayakawa, K., Miyati, T. and Maeda, F. (2011). Imaging parameter effects in apparent diffusion coefficient determination of magnetic resonance imaging. Eur J Radiol *77*, 185-188.

Righini, A., Bianchini, E., Parazzini, C., Gementi, P., Ramenghi, L., Baldoli, C., Nicolini, U., Mosca, F. and Triulzi, F. (2003). Apparent diffusion coefficient determination in normal fetal brain: a prenatal MR imaging study. AJNR Am J Neuroradiol *24*, 799-804.

Rodrigues, K. and Ellen Grant, P. (2011). Diffusion-weighted imaging in neonates. Neuroimaging Clin. N. Am. *21*, 127-51, viii.

Sagar, P. and Grant, P.E. (2006). Diffusion-weighted MR imaging: pediatric clinical applications. Neuroimaging Clin. N. Am. *16*, 45-74, viii.

Smith, S.M. (2002). Fast robust automated brain extraction. Hum Brain Mapp *17*, 143-155.

Snook, L., Paulson, L., Roy, D., Phillips, L. and Beaulieu, C. (2005). Diffusion tensor imaging of neurodevelopment in children and young adults. Neuroimage *26*, 1164-1173.

Utsunomiya, H. (2011). Diffusion MRI abnormalities in pediatric neurological disorders. Brain Dev. *33*, 235-242.

Vermeulen, R.J., van Schie, P.E.M., Hendrikx, L., Barkhof, F., van Weissenbruch, M., Knol, D.L. and Pouwels, P.J.W. (2008). Diffusion-weighted and conventional MR imaging in neonatal hypoxic ischemia: two-year follow-up study. Radiology *249*, 631-639.

Westlye, L.T., Walhovd, K.B., Dale, A.M., Bjørnerud, A., Due-Tønnessen, P., Engvig, A., Grydeland, H., Tamnes, C.K., Ostby, Y. and Fjell, A.M. (2010). Life-span changes of the human brain White matter: diffusion tensor imaging (DTI) and volumetry. Cereb. Cortex *20*, 2055-2068.

Wheeler-Kingshott, C.A.M. and Cercignani, M. (2009). About "axial" and "radial" diffusivities. Magn Reson Med *61*, 1255-1260.

White, T., Magnotta, V.A., Bockholt, H.J., Williams, S., Wallace, S., Ehrlich, S., Mueller, B.A., Ho, B., Jung, R.E., Clark, V.P., et al. (2011). Global white matter abnormalities in schizophrenia: a multisite diffusion tensor imaging study. Schizophr Bull *37*, 222-232.

Wilke, M., Holland, S.K., Altaye, M. and Gaser, C. (2008). Template-O-Matic: a toolbox for creating customized pediatric templates. Neuroimage *41*, 903-913.

Wilke, M., Schmithorst, V.J. and Holland, S.K. (2002). Assessment of spatial normalization of whole-brain magnetic resonance images in children. Hum Brain Mapp *17*, 48-60.

Wilke, M., Schmithorst, V.J. and Holland, S.K. (2003). Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. Magn Reson Med *50*, 749-757.

Zarifi, M.K., Astrakas, L.G., Poussaint, T.Y., Plessis Ad, A.D., Zurakowski, D. and Tzika, A.A. (2002). Prediction of adverse outcome with cerebral lactate level and apparent diffusion coefficient in infants with perinatal asphyxia. Radiology *225*, 859-870.

7.2. Books

- Huettel, Song & Mcarthy (2009), *Functional Magnetic Resonance Imaging*, Sinauer Associates (Sunderland, MA, USA)
- Oleg S. Pianykh (2008), *Digital Imaging and Communications in Medicine (DICOM) : A Practical Introduction and Survival Guide*, Springer-Verlag (Berlin Heidelberg)
- Woosley, hanaway & Gado (2008), *the Brain Atlas: A Visual Guide to the Human Central Nervous System*, John Wiley & Sons (New Jersey, USA)
- Mai, Paxinos & Voss (2008), *Atlas of The Human Brain*, Elsevier (USA)
- *Stedman's Medical Dictionary* (1979), Williams & Wilkins (Baltimore, USA)

7.3. Courses

- Pr. Nouchine Hadjikhani, *Introduction to brain Imaging*, 2010, EPFL, Switzerland.
- Dr. Randy Gollub, *HST.583 Functional Magnetic Resonance Imaging: Data Acquisition and Analysis*, 2010, MIT, U.S.; Lecture given by Anastasia Yendiki, PhD.
- Pr. Rolf Gruetter, *Fundamental in biomedical imaging*, 2010, EPFL, Switzerland.
- Pr. Dominique Sprumont, *Introduction to Law and Ethics in LS & AMP*, 2009-2010, EPFL, Switzerland.

7.4. Web sites

- http://www.na-mic.org/Wiki/index.php/CTSC:ARRA_supplement
- http://www.mr-tip.com/serv1.php
- http://www2.massgeneral.org/allpsych/psychneuro/labfunctionalstructural.asp
- http://hst.mit.edu/index.jsp
- http://www.nmr.mgh.harvard.edu/martinos/flashHome.php
- http://www.mgh.harvard.edu/
- http://www.fil.ion.ucl.ac.uk/spm/
- http://www.loni.ucla.edu/Atlases/Atlas_Detail.jsp?atlas_id=15
- http://www.partners.org/rescomputing/template.asp?pageid=99&ArticleTitle=RPDR&level1ID=9&tocID=9&articleSubPage=true
- http://office.microsoft.com/en-us/access-help/
- http://medical.nema.org/
- http://www.nmr.mgh.harvard.edu/martinos/userInfo/data/sofFreeSurf.php
- http://www.slicer.org/
- http://www.dcm4che.org/
- http://www.loni.ucla.edu/Atlases/Atlas_Detail.jsp?atlas_id=15
- http://cfr.vlex.com/vid/56-107-irb-membership-19703309
- http://healthcare.partners.org/phsirb/abouthrc.htm

- http://rpdrweb/partners/rpdrinfo/home/overview.htm (only accessible through Partners network)
- http://www.osirix-viewer.com
- http://www.clearcanvas.ca/dnn/
- http://www.cms.gov/HIPAAGenInfo/02_TheHIPAALawandRelated%20Information.asp#TopOfPage
- http://dicom.offis.de/dcmtk.php.en
- http://nifti.nimh.nih.gov/nifti-1/
- http://www.fmrib.ox.ac.uk/fsl/bet2/index.html
- http://www.fmrib.ox.ac.uk/fsl/
- http://surfer.nmr.mgh.harvard.edu/fswiki/FreeviewGuide/FreeviewIntroduction

7.5. Other

- Guide to RPDR Data pfd document found on the Partners Research Computing web site.
- Project IRB approval (see 3.1)
- RO1 grant application (see Introduction)
- Shawn Murphy MD, Ph.D.,"Research Patient Data Registry (RPDR) at Partners Healthcare" presentation, January 31, 2011

8. Appendix

8.1. Abbreviations

A

AD	:	Axial Diffusivity
ADC	:	Apparent Diffusion Coefficient
ANCOVA	:	Analysis of Covariance
ARRA	:	American Recovery and Reinvestment Act

B

BET	:	Brain Extraction Tool
BWH	:	Brigham And Women's Hospital

C

CHB	:	Children's Hospital Boston
CITI	:	Collaborative Institutional Training Initiative
CSF	:	Cerebrospinal Fluid
CT	:	Computed Tomography

D

DICOM	:	Digital Imaging and Communications in Medicine
dMRI	:	Diffusion Magnetic Resonance Imaging
DTI	:	Diffusion Tensor Imaging
DW	:	Diffusion Weighted
DWI	:	Diffusion Weighted Imaging

E

EMPI	:	Enterprise Master Patient Index

F

FA	:	Fractional Anisotropy
FDA	:	Food and Drug Administration
FIT	:	Fractional Intensity Threshold
FLAIR	:	FLuid Attenuated Inversion Recovery
FSL	:	FMRIB Software Library

G

GLM	:	General Linear Model

H

HIPAA	:	Health Insurance Portability and Accountability Act
HIV	:	Human Immunodeficiency Virus

I

i2b2	:	Informatics for Integrating Biology and the Bedside
IRB	:	Institutional Review Board

M

MD	:	Mean Diffusivity
MGH	:	Massachusetts General Hospital
Mi2b2	:	Medical Imaging Informatics Bench to Bedside
MR	:	Magnetic Resonance
MRI	:	Magnetic Resonance Imaging
MRN	:	Medical Record Number

P

PACS	:	Picture Archiving and Communication System
PET	:	Positron Emission Tomography
PHRC	:	Partners Human Research Committee
PI	:	Principal Investigator
PLIC	:	Posterior Limb of the Internal Capsule

R

RD	:	Radial Diffusivity
RDS	:	Radiological Decision Support
ROI	:	Region Of Interest
RPDR	:	Research Patient Data Registry

S

sCC	:	Splenium of the Corpus Callosum
SCP	:	Service Class Provider
SCU	:	Service Class User
SNR	:	Signal to Noise Ratio
SQL	:	Structured Query Language
STD	:	STandard Deviation

T

TE	:	Echo Time
TR	:	Repetition Time

U

US	:	Ultrasound

V

VB	:	Visual Basic

W

WBA	:	Whole Brain Average

8.2. Figures and Tables

8.2.1. Figures

FIGURE 1: PATIENTS ARE SCANNED IN THE HOSPITAL (1) AND MRI DATA ARE STORED IN PACS (2). RADIOLOGISTS ACCESS THE IMAGES AND THEIR RESULTING OBSERVATIONS (AS WELL AS ALL MEDICAL INFORMATION REGARDING PATIENTS) ARE STORED AND ORGANIZED IN RPDR (3). THE RPDR QUERY ALLOWS THE RETRIEVAL OF THIS MEDICAL INFORMATION IN MICROSOFT ACCESS DATABASE FORMAT (4). THIS DATABASE IS USED TO IDENTIFY PATIENTS MEETING CERTAIN CRITERIA (5) AND THE CORRESPONDING LIST OF IDENTIFIERS IS SENT TO MI2B2 SOFTWARE IN ORDER TO RETRIEVE THE RELATED IMAGES (6). THE BB-PIPELINE THEN AUTOMATICALLY ORGANIZES THE DATA (7) AND EXTRACTS TECHNICAL INFORMATION THAT IS ADDED TO THE ACCESS DATABASE (8). THE RESULTING DATABASE IS USED TO PRECISELY IDENTIFY AND SELECT COMPARABLE IMAGES WITH MODALITIES OF INTEREST (9) THAT ARE USED FOR FURTHER DATA ANALYSIS (10). 3
FIGURE 2: STANDARD SPIN-ECHO SEQUENCE WITH THE ADDITIONAL GRADIENT (THE OTHER GRADIENTS ALONG THE X, Y AND Z DIRECTIONS FOR SPATIAL ENCODING ARE OMITTED FOR SIMPLIFICATION). THE B-VALUE IS AN EXPERIMENTAL PARAMETER WHICH DEPENDS ON THE LENGTH, HEIGHT AND TIMING OF THE DW GRADIENT. 4
FIGURE 3: WHEN MOLECULES OF WATER DIFFUSE ALONG THE DIRECTION OF THE GRADIENT (HORIZONTAL YELLOW ARROWS), A LOSS IN PHASE COHERENCE IS NOTICED AFTER THE REPHASING STEP. 4
FIGURE 4: MYELIN, CELL MEMBRANE AND MICROTUBULES/NEUROFILAMENTS RESTRICT WATER DIFFUSION IN PERPENDICULAR DIRECTIONS LEADING TO LESS SIGNAL ATTENUATION. 5
FIGURE 5: ADC VALUES ARE CALCULATED FOR DIFFERENT B-VALUES. HIGHER B-VALUES HAVE MORE INFLUENCE ON ADC THAN LOWER B-VALUES 7
FIGURE 6: CORRELATION PLOT BETWEEN THE AGES AND THE MD VALUES IN 6 DIFFERENT ROIS (TRACE/3 ADC IS EQUIVALENT TO THE MD).... 8
FIGURE 7: TRANSVERSE MR IMAGES IN FIVE CHILDREN 17 DAYS TO 10 YEARS OLD AT 2 DIFFERENT LEVELS (LEFT PICTURE) AND THE TIME COURSE PLOT OF D_{AVE} IN THE POSTERIOR LIMB OF THE INTERNAL CAPSULE (PLIC). THE DECAY OF D_{AVE} IN EARLY DEVELOPMENTAL STAGE CAN CLEARLY BE NOTICED IN BOTH FIGURES (E.G. CHANGE IN BRIGHTNESS FROM VERY BRIGHT TO DARK IN LEFT PICTURE). 9
FIGURE 8: TIME COURSES FOR FA AND D_{AVE} IN THE POSTERIOR LIMB OF THE INTERNAL CAPSULE (PLIC) AND THE SPLENIUM OF THE CORPUS CALLOSUM (SCC) AS SHOWN BY HERMOYE ET AL. (2006) (A). SIMILAR TIME COURSE IN LÖBEL ET AL. (2009) PAPERS FOR THE PLIC AND THE INFERIOR FRONTAL WHITE MATTER (B); NOTE THE LOGARITHMIC FITTING CURVE IN RED. COMPARING THE D_{AVE} TIME EVOLUTION WITH THE ONE OF FIGURE 7, WE NOTE A SIMILAR PATTERN FOR THE PLIC ROI 9
FIGURE 9: TOP ROW T2 WEIGHTED IMAGES. BOTTOM ROW APPARENT DIFFUSION COEFFICIENT (ADC) MAPS. NORMAL NEONATES HAVE LOWER T2 SIGNAL AND ADC IN REGIONS THAT ARE UNDERGOING MYELINATION (ARROWS IN A, D RESPECTIVELY). NEONATES WITH HYPOXIC ISCHEMIC INJURY HAVE LOWER ADC IN THE SAME REGIONS (ARROWS IN E). IN NORMAL OLDER CHILDREN WITH FULLY MYELINATED BRAIN REGIONS (C, F) THE ADC IMAGE IS HOMOGENEOUS (F) AS COMPARED TO THE NORMAL NEONATE (D) MAKING AREAS OF ABNORMALLY DECREASED ADC EASIER TO DETECT IN OLDER CHILDREN BECAUSE FULLY MYELINATED BRAIN REGIONS HAVE UNIFORM DIFFUSIVITY. THEREFORE INJURIES ARE MORE EASILY DETECTED BY VISUAL INSPECTION ALONE AS THEY APPEAR AS DEVIATIONS ON A HOMOGENEOUS BACKGROUND. 12
FIGURE 10: SCHEME OF RPDR QUERY 16
FIGURE 11: RPDR ENHANCED QUERY TOOL WEB INTERFACE 17
FIGURE 12: MI2B2 WORKBENCH USER INTERFACE. IT QUERIES THE PACS FOR SELECTED MRNS AND RETURNS THE COMPLETE LIST OF AVAILABLE STUDIES FOR EACH PATIENT. ONE OR MORE FILTERING AND SORTING CRITERIA CAN BE USED TO QUICKLY REFINE THE SEARCH LIST (UPPER IMAGE). AN ADDITIONAL TAB (WHICH WAS NOT AVAILABLE IN THE PROTOTYPE I USED) WILL BE USED TO VIEW IMAGE FILES DIRECTLY FROM THE MI2B2 CACHE BEFORE TRANSFERRING THEM TO THE USERS OWN DISK SPACE (LOWER IMAGE) 27
FIGURE 13: BB-PIPLINE WORKFLOW. DATA RECEIVED FROM PACS ARE FIRST TRANSFERRED AND DECOMPRESSED TO THE LOCAL WORKSTATION (1). THEY ARE THEN SENT THROUGH RUDOLPH'S PIPELINE AND AGE CALCULATOR THAT RENAME AND PRE-ORGANIZE THE RAW DATA (2)-(3). USING ACCESS INFORMATION INPUT, THE STUDIES ARE THEN SORTED INTO DIFFERENT DIRECTORIES ACCORDING TO THEIR AGE AND HEALTH STATUS (4). DICOM FILES ARE CONVERTED INTO NIFTI FOMAT AND SENT TO THE WORKING DIRECTORY (5). DICOM HEADERS ARE EXTRACTED AND ADDED TO THE ACCESS DATABASE (6), AND LOG FILE IS UPDATED (7). 29
FIGURE 14: DIRECTORY HIERARCHY AND NAME FORMAT AS RETRIEVED FROM PACS. DIRECTORY AND FILE NAMES ARE A COMBINATION OF INSTITUTION, MANUFACTURER, MACHINE MODEL, AND RANDOMLY GENERATED NUMBERS BASED ON DATE AND TIME. ACCORDING TO DICOM, THEY ARE GLOBALLY UNIQUE IDENTIFIERS. 31
FIGURE 15: DIRECTORY HIERARCHY AFTER DATA HAVE BEEN SENT THROUGH BB-PIPELINE. THEY ARE NOW SORTED ACCORDING TO THE AGE IN MONTHS, HEALTH STATUS, NAMES OF DIRECTORIES, AND FILES HAVE BEEN TRANSFORMED MEANINGFULLY: THE STUDY NAME NOW

CONTAINS PATIENT MRN, AGE IN DAYS, STUDY DATE, AND OTHER SCANNER INFORMATION. THE IMAGE NAME INCLUDES THE SERIES NUMBERS CORRESPONDING TO A PARTICULAR SERIES TYPE (T1-, T2-WEIGHTED IMAGES, DWI IMAGES, FLAIR IMAGES, ETC.)..... 31
FIGURE 16: DATA PROCESSING WORKFLOW FOR THE CALCULATION OF WBA ADC AND FA. THE LOWB VOLUME IS BRAIN EXTRACTED THROUGH BET (1) AND APPLIED AS A MASK TO THE ADC VOLUME (2). THEN, AN UP-THRESHOLD IS APPLIED TO THIS VOLUME IN ORDER TO REMOVE CSF REGIONS (3) AND THE RESULTING ADC VOLUME IS SENT THROUGH BET (4) TO OBTAIN THE FULLY PROCESSED ADC VOLUME. AT THE SAME TIME, A MASK IS GENERATED (5) AND APPLIED TO THE CORRESPONDING FA VOLUME (6) TO OBTAIN THE FULLY PROCESSED FA VOLUME..... 36
FIGURE 17: NUMBER OF PATIENTS WHO HAD A BRAIN MRI SCAN AT A PARTICULAR AGE IN MONTHS FROM 0 TO 6 YEARS (A) AND IN WEEKS FROM 0 TO 4 MONTHS (B). 40
FIGURE 18: PROPORTION OF DIFFUSION BRAIN MRI SCANS FROM 0 TO 24 MONTHS. THE BLUE PORTION REPRESENTS THE SCANS THAT INCLUDE DIFFUSION CRITERIA AND THE ORANGE PORTION REPRESENTS THE REMAINING BRAIN MRI SCANS (A); AVERAGE PROPORTION OF DIFFUSION AND UNDETERMINED SCANS OVER 0 TO 24 MONTHS (B)..... 41
FIGURE 19: PROPORTION OF SCANS ACQUIRED BEFORE THE YEAR 2006 (BLUE) AND AFTER 2006 (MAGENTA) (A) AND OVERALL DISTRIBUTION OF THESE SCANS FOR EACH YEAR FROM 2000 TO 2010. 41
FIGURE 20: DISTRIBUTION (A) AND AVERAGE PROPORTION (B) OF "NORMAL", "ABNORMAL" AND "UNDEFINED" PATIENTS ACROSS THE AGES OF 0 TO 12 MONTHS..... 42
FIGURE 21: TABLE SUMMARIZES, THE TOTAL NUMBER OF STUDIES IDENTIFIED, THE CORRESPONDING NUMBER OF STUDIES SUCCESSFULLY RETRIEVED FROM PACS AND THE NUMBER OF STUDIES NOT AVAILABLE, FOR EACH AGE RANGE. THE AVERAGE PROPORTION IS DISPLAYED IN THE PIE CHART (RIGHT)..... 44
FIGURE 22: ACTUAL PROPORTION OF DIFFUSION BRAIN MRI SCANS FROM 0 TO 24 MONTHS. THE BLUE PORTION REPRESENTS THE SCANS THAT INCLUDE DIFFUSION CRITERIA AND THE ORANGE PORTION REPRESENTS THE REMAINING BRAIN MRI SCANS (A); AVERAGE PROPORTION OF DIFFUSION AND NO DIFFUSION SCANS OVER 0 TO 24 MONTHS (B). 45
FIGURE 23: DISTRIBUTION OF DIFFUSION SCANS ACCORDING TO THEIR TE IN MS. THIS GRAPH INCLUDES ALL OF THE DIFFUSION DATA WE HAD AT OUR DISPOSAL THAT WERE ACQUIRED WITH THE "OLD" (PRIOR TO 2006) DIFFUSION PROTOCOL (N = 2371). 47
FIGURE 24: TOTAL NUMBER OF STUDIES FALLING UNDER DATA OF INTEREST CRITERIA (SELECTED + DISCARDED) FOR EACH AGE RANGE FROM 0 TO 24 MONTHS. IN ORANGE, THE NUMBER OF SERIES DISCARDED FOR INSUFFICIENT QUALITY AND IN GREEN THE NUMBER OF STUDY SELECTED FOR FURTHER ANALYSIS. 48
FIGURE 25: DISTRIBUTION OF WBA_{ADC} [MM^2/S] ACROSS AGE IN DAYS FROM NON-PROCESSED DATA (A); AFTER THE BRAIN EXTRACTED LOWB MASKING (STEP 2 OF FIGURE 16) (B); AFTER THRESHOLD APPLICATION TO REMOVE CSF (STEP 3 OF FIGURE 16) (C); AND AFTER BET APPLICATION TO ADC VOLUMES (STEP 4 OF FIGURE 16) (D)..... 50
FIGURE 26: DISTRIBUTION OF WBA_{FA} ACROSS AGES [DAYS] FROM NON-PROCESSED DATA (A), AFTER BRAIN EXTRACTED LOWB MASKING (STEP (2) APPLIED TO FA VOLUMES, FIGURE 16) (B), AFTER THRESHOLD APPLICATION TO REMOVE CSF (STEP (3), ADC VOLUME MASK APPLIED TO CORRESPONDING FA VOLUME, FIGURE 16) (C) AND AFTER FULLY PROCESSED ADC VOLUME MASKING (STEP (6), FIGURE 16) (D)..... 50
FIGURE 27: STANDARD DEVIATION OF WBA_{ADC} AND WBA_{FA} VALUES FOR EACH AGE RANGE, (A) AND (B) RESPECTIVELY BEFORE (BLUE) AND AFTER (YELLOW) DATA PROCESSING. DIFFERENCE IN STANDARD DEVIATION FOR EACH AGE RANGE, (C) AND (D), NEGATIVE VALUES REPRESENT A DECREASE IN STANDARD DEVIATION AFTER DATA PROCESSING..... 51
FIGURE 28: CURVE FITS FOR WBA_{ADC} VALUES ACROSS AGES. WITH NO FIT (A), WITH LINEAR (B), LOGARITHMIC (C) AND BIEXPONENTIAL (D) FITS. PREDICTION BOUNDS AT 95% CONFIDENCE FOR THE FUNCTION (DASHED LINES) AND NEW OBSERVATIONS (DOTTED LINES) AND R^2 VALUES ARE ALSO DISPLAYED FOR EACH MODEL. (ESTIMATED PARAMETERS ARE AVAILABLE IN APPENDIX 8.9)..... 52
FIGURE 29: CURVE FITS FOR WBA_{FA} VALUES ACROSS AGES. WITH NO FIT (A), WITH LINEAR (B), LOGARITHMIC (C) AND BIEXPONENTIAL (D) FITTINGS. PREDICTION BOUNDS AT 95% CONFIDENCE FOR THE FUNCTION (DASHED LINES) AND NEW OBSERVATIONS (DOTTED LINES) AND R^2 VALUES ARE ALSO DISPLAYED FOR EACH MODEL. (ESTIMATED PARAMETERS ARE AVAILABLE IN APPENDIX 8.9) 53
FIGURE 30: NATURAL LOGARITHM OF WBA_{ADC} (A) AND WBA_{FA} (B) VERSUS NATURAL LOGARITHM OF AGE. BLACK LINES REPRESENT LINEAR FITS WITH THE CORRESPONDING R^2 VALUES..... 54
FIGURE 31: NATURAL LOGARITHM OF WBA_{ADC} (A) AND WBA_{FA} (B) VERSUS NATURAL LOGARITHM OF AGE SEPARATED ACCORDING TO GENDER (FEMALE ORANGE AND BLUE CIRCLES, MALE GREEN AND RED SQUARES FOR WBA_{ADC} AND WBA_{FA} RESPECTIVELY). CORRESPONDING COLOR-CODED LINEAR FITS WITH CORRESPONDING R^2 VALUES ARE ALSO DISPLAYED..... 55
FIGURE 32: FOUR PATIENTS WITH MULTIPLE SCANS ARE HIGHLIGHTED IN ORANGE (11 AND 66 DAYS), GREEN (44 AND 199 DAYS), BLUE (82, 406, 588 AND 769 DAYS) AND RED (99, 190, 286, 414 AND 726 DAYS). CORRESPONDING PROCESSED ADC IMAGES OF THE PATIENT DISPLAYED IN RED COLOR ARE INCLUDED. THE 95% PREDICTION BOUNDS FOR NEW OBSERVATIONS IN THE BIEXPONENTIAL MODEL ARE ALSO DISPLAYED (DASHED LINES). 56
FIGURE 33: FOUR PATIENTS WITH MULTIPLE SCANS ARE HIGHLIGHTED IN ORANGE (11 AND 66 DAYS), GREEN (44 AND 199 DAYS), BLUE (82, 406, 588 AND 769 DAYS) AND RED (99, 190, 286, 414 AND 726 DAYS). CORRESPONDING PROCESSED FA IMAGES OF THE PATIENT

DISPLAYED IN RED COLOR ARE INCLUDED. THE 95% PREDICTION BOUNDS FOR NEW OBSERVATIONS IN THE BIEXPONENTIAL MODEL ARE ALSO DISPLAYED (DASHED LINES). 56

8.2.2. Tables

TABLE 1: LIST OF THE 12 PROCEDURES THAT ARE THE MOST LIKELY TO INCLUDE BRAIN MRI 21
TABLE 2: THE LOG FILE CONTAINS MRNS WITH THEIR CORRESPONDING STUDY DATE, REQUEST NUMBER (EQUIVALENT TO A REQUEST ID) WITH ITS REQUEST DATE, THE STATUS OF THE REQUEST, THE DATE AT WHICH THE STUDY HAS BEEN RETRIEVED AND FINALLY OTHER USEFUL INFORMATION CONCERNING THE STUDY (AGE IN DAYS AND WHETHER THE PATIENT HAS HAD SEVERAL SCANS). THE STUDY IS FLAGGED AS "DONE", "PENDING" AND "NOT FOUND" DEPENDING ON WHETHER THE STUDY HAS BEEN SUCCESSFULLY RETRIEVED, NOT YET RETRIEVED OR WAS NOT FOUND IN THE PACS RESPECTIVELY. A VB SCRIPT AUTOMATICALLY UPDATES THIS DATASHEET FROM THE TEXT FILE GENERATED BY THE BB-PIPELINE. 33
TABLE 3: THIS TABLE DISPLAYS A SUBSET OF THE DATA BECAUSE THE COMPLETE DATA SET COULD NOT FIT ON A REGULAR PAGE. THE FIRST COLUMN CONTAINS THE MRN. THE FURTHEST RIGHT COLUMN INDICATES THE NUMBER OF BRAIN MRI SCANS EACH PATIENT HAD AND THE BOTTOM ROWS PARTIALLY DELINEATE THE AGE RANGE FROM 0 TO 72 MONTHS. THE CIRCLES SPECIFY THE SCANS CONTAINING DIFFUSION IMAGING AND THE CROSSES ARE THE UNDETERMINED BRAIN MRI SCANS (SEE PART 4.1.2.2). THE YELLOW HIGHLIGHTING INDICATES MRI SCANS THAT HAVE A CORRESPONDING CT SCAN. FINALLY, THE NUMBER IN THE LOWER LEFT CORNER CORRESPONDS TO THE TOTAL NUMBER OF PATIENTS HAVING MORE THAN ONE SCAN AND THE LOWER RIGHT CORNER CORRESPONDS TO THE TOTAL NUMBER OF SCANS FOR THESE PATIENTS. NOTE THAT DUE TO HIPAA REGULATIONS, THE MRNS HAVE BEEN ALTERED FOR THE PURPOSES OF THIS FIGURE. 43
TABLE 4: SCAN DETAILS ABOUT A PARTICULAR STUDY. MAGNETIC FIELD STRENGTH IS DISPLAYED IN MT, TR AND TE IN MS, ROWS, COLUMNS, PIXEL SPACING, SLICE THICKNESS IN MM AND BANDWIDTH IN HZ/PIXEL. SERIES NUMBER IS SPECIFIC FOR EACH SERIES WITHIN A STUDY (SEE PART 3.2.4 (2)). UNCOLORED ROWS REPRESENT NON-DIFFUSION DATA AND COLORED ONES INDICATE DIFFUSION SEQUENCES. 46

8.3. People

Bill Wang, MSc, is a Systems Programmer of Partners Healthcare. He is responsible for design, implementation, quality assurance, and support of mi2b2 project.

Chris Herrick, MBA, is a Corporate Manager for Research Computing at Partners Healthcare. Chris manages the Medical Imaging Informatics Bench to Bedside (mi2b2) project and the Partners' Pharmacovigilance and Comparative Effectiveness initiative. He also serves as technical architect, project manager and liaison for various collaborators.

Ellen Grant, MD, is the Director of the Fetal-Neonatal Neuroimaging and Developmental Science Center at Children's Hospital Boston. She is an expert in pediatric neuroimaging and is currently involved in several projects including the development of new pediatric hardware (multichannel head coil and motion insensitive pulse sequence) and multimodality approach to estimating newborn brain oxygen metabolism in the context of perinatal brain injury. For many years before joining Children's Hospital she worked at MGH and thus knows a lot about the pediatric neuroimaging data acquired there.

Lilla Zollei, PhD, is working at Martinos Center on various medical image data registration projects mostly in pediatric populations, including surface and volumetric registration of brain MRI image and Diffusion Tensor Imaging (DTI) alignment. She is currently working on the development of novel representation and computational tools that will establish an age-dependent 4D atlas and analytic scheme to better understand normal brain development as well as examine the effects of premature birth.

Randy Gollub, MD, PhD, is the director of the "Laboratory for Neuroimaging Applications to Pain, Acupuncture and Placebo Research". This lab works in tight collaboration with many groups from different departments within the Martinos Center, the Psychiatric Neuroimaging Research Program, and across the country, to perfect the development and implementation of imaging technologies, especially in the domain of neuropsychiatric disorders.

More specifically, Dr. Gollub's team is focused on investigating the neuronal mechanisms underlying pain perception and its modulation by expectancy, placebo, and acupuncture treatment[34].

Besides her principal field of investigation, Dr. Gollub is involved and actively participates in a multitude of other interesting projects, notably the on-going Medical Imaging Informatics Bench to Bedside (mi2b2) project which is building the infrastructure necessary to make imaging data at multiple sites available to clinical translational investigators with appropriate human subject and institutional protections[35,36].

Rudolph Pienaar, PhD, is the Technical Director of the Fetal-Neonate NeuroImaging Development Science Center at Children's Hospital and is responsible for the underlying technical

[34] http://www.nmr.mgh.harvard.edu/martinos/flashHome.php
[35] Mi2b2 project grant application
[36] http://www.na-mic.org/Wiki/index.php/CTSC:ARRA_supplement

infrastructure. He is also engaged in more basic research centered around MRI analysis, particularly as it pertains to pediatric brain development and pathology.

Shawn Murphy, MD, PhD, is the Associate Director of the Laboratory of Computer Science (LCS) in the Clinical and Research Informatics Division of the Department of Medicine at MGH and an Assistant Professor of Neurology at Harvard Medical School. His current projects include Informatics for Integrating Biology and the Bedside (i2b2), RPDR and he is also the co-PI of mi2b2 project.

Taowei David Wang, PhD, is an application analyst in the Research Computing group of Partners HealthCare's Information Systems department. He is primarily interested in human-computer interaction and information visualization. His current projects include mi2b2's user interface development.

8.4. Brain MRI procedures in RPDR

Name	Code_Type	Code
Magnetic resonance imaging, brain, functional MRI; including test selection and administration of repetitive body part movement and/or visual stimulation, not requiring physician or psychologist administration	CPT	0.32
CT or MRI of the brain performed within 24 hours of arrival to the hospital	CPT	1.1
CT or MRI of the brain performed greater than 24 hours after arrival to the hospital for patients aged 18 years and older with an admitting diagnosis of ischemic stroke or TIA or intracranial hemorrhage	CPT	1.18
computer assisted sugery with MR/MRA	ICD	88.41
Presence or absence of hemorrhage, mass lesion, and acute infarction documented in final CT or MRI report	CPT	88.91
3D rendering with interpretation and reporting of computed tomography, magnetic resonance imaging, ultrasound, or other tomographic modality; not requiring image postprocessing on an independent workstation	CPT	88.96
3D rendering with interpretation and reporting of computed tomography, magnetic resonance imaging, ultrasound, or other tomographic modality; requiring image postprocessing on an independent workstation	CPT	88.97
Coronal, sagittal, multiplanar, oblique, 3-dimensional and/or holographic reconstruction of computerized axial tomography, magnetic resonance imaging, or other tomographic modality	CPT	92.11
Magnetic resonance (eg, proton) imaging, brain (including brain stem); with contrast material(s)	CPT	92.12
Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material	CPT	92.18
Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material, followed by contrast material(s) and further sequences	CPT	61751
Magnetic resonance (eg, proton) imaging, orbit, face, and neck; with contrast material(s)	CPT	70336
Magnetic resonance (eg, proton) imaging, orbit, face, and neck; without contrast material(s)	CPT	70540
Magnetic resonance (eg, proton) imaging, orbit, face, and neck; without contrast material(s), followed by contrast material(s) and further sequences	CPT	70541
Magnetic resonance (eg, proton) imaging, temporomandibular joint(s)	CPT	70542
Magnetic resonance angiography, head and/or neck with or without contrast material(s)	CPT	70543
Magnetic resonance angiography, head; with contrast material(s)	CPT	70544
Magnetic resonance angiography, head; without contrast material(s)	CPT	70545
Magnetic resonance angiography, head; without contrast material(s), followed by contrast material(s) and further sequences	CPT	70546
Magnetic resonance angiography-Oncall	OPA	70551
Magnetic resonance guidance for needle placement (eg, for biopsy, needle aspiration, injection, or placement of localization device) radiological supervision and interpretation	CPT	70552
Magnetic resonance guidance for needle placement (eg, for biopsy, needle aspiration, injection, or placement of localization device) radiological supervision and interpretation	CPT	70553
Magnetic resonance guidance for, and monitoring of, tissue ablation	CPT	70554
Magnetic resonance imaging of brain and brain stem	ICD	70554
Magnetic resonance imaging of of brain-Oncall	OPA	75650
Magnetic resonance imaging of other and unspecified sites	ICD	75665
Magnetic resonance imaging, brain, functional MRI; including test selection and administration of repetitive body part movement and/or visual stimulation, not requiring physician or psychologist administration	CPT	75671
Magnetic resonance imaging-Oncall	OPA	76375
Magnetic resonance spectroscopy	CPT	76376
Other intraoperative magnetic resonance imaging	ICD	76377
Stereotactic biopsy, aspiration, or excision, including burr hole(s), for intracranial lesion; with computed tomography and/or magnetic resonance guidance	CPT	76390
Unlisted magnetic resonance procedure (eg, diagnostic, interventional)	CPT	76393
Unlisted nervous system procedure, diagnostic nuclear medicine	CPT	76394

Brain imaging, complete study; static	CPT	76498
Brain imaging, complete study; with vascular flow	CPT	77021
Brain imaging, limited procedure; static	CPT	78600
Brain imaging, limited procedure; with vascular flow	CPT	78601
Brain imaging, vascular flow only	CPT	78605
Cerebral vascular flow	CPT	78606
Angiography, carotid, cerebral, bilateral, radiological supervision and interpretation	CPT	78610
Angiography, carotid, cerebral, unilateral, radiological supervision and interpretation	CPT	78615
Angiography, cervicocerebral, catheter, including vessel origin, radiological supervision and interpretation	CPT	78615
Arteriography of cerebral arteries	ICD	78699
Cerebral scan	ICD	3110F
Cerebral vascular flow	CPT	3111F
Diagnostic procedures on skull, brain, and cerebral meninges	ICD	3112F
Other diagnostic procedures on brain and cerebral meninges	ICD	ZNCS3
Scan of other sites of head	ICD	ZSAV3
Total body scan	ICD	ZTAH1-A

8.5. HIV related diagnosis

Name	Code_Type	Code
Asymptomatic human immunodeficiency virus [HIV] infection status	ICD	V08
HIV antibody tes-Oncall	ODA	DKNF1
Acquired immune deficiency syndrome-Oncall	ODA	DJAA6
Acquired immunodeficiency syndrome	ICD	42.9
AIDS related complex-Oncall	ODA	DJAY9
AIDS-LMR 11	ICD	42
HIV infection (symptomatic)-Oncall	ODA	DJBQ4-1
HIV infection-Oncall	ODA	DJBQ4
HIV positive-LMR 190	ICD	42
Human immunodeficiency virus infection causing other specified infections	ICD	42.1
Human immunodeficiency virus infection with specified infections	ICD	42
Human immunodeficiency virus infection with specified malignant neoplasms	ICD	42.2
HIV exposure-Oncall	ODA	DJBF6
HIV infection (symptomatic)-Oncall	ODA	DJBQ4-1
HIV wasting-Oncall	ODA	BJAQ2
Human immunodeficiency virus [HIV] counseling	ICD	V65.44
Positive HIV antibody test-Oncall	ODA	DKNF1-1
Human immunodeficiency virus, type 2 [HIV-2] infection in conditions classified elsewhere and of unspecified site	ICD	79.53
Human immunodeficiency virus, type 2[HIV 2]	ICD	79.53
Nonspecific serologic evidence of human immunodeficiency virus [HIV]	ICD	795.71
Positive serological or viral culture findings for human immunodeficiency virus (HIV) associated virus (HTLV-III/LAV)	ICD	795.8
HIV	-	-
HIV Infection	MDC	25

8.6. Complete list of tables received from RPDR

Cardiology
Contact Information
Demographics
Diagnoses - one row per diagnosis
Discharge Summaries
Encounters - Visit Information and Diagnoses included in one row (Billing Data only)
Endoscopy
HealthHistory - Newborn Data, LMR Health Maintenance, LMR Vital Signs
LMR health maintenance
LMR medications
LMR outpatient notes- not available for Limited Data Sets
LMR problems
LMR vital signs
Medications - one row per medication
Microbiology Data
Mrn - A list of patient medical record numbers
Operative Reports
Pathology
PEARAllergy - Allergy Data from PEAR (Partners Enterprise Allergy Repository)
Procedures - one row per procedure
Providers - Top three providers for each patient
Pulmonary
Radiology
Transfusion

8.7. IMPRESSION notes assessing "normal" brains

Age-appropriate MRI of the brain with no structural abnormality identified.
MRI appearance of the brain is age appropriate
MRI of the brain normal for corrected gestational age without evidence of prior injury.
Negative exam
No abnormality is seen on this unenhanced MRI of the brain
No acute abnormality identified.
No definite structural abnormality identified.
No evidence of acute intracranial abnormality
No evidence of diffusion weighted or gross FLAIR abnormalities.
No evidence of focal mass lesion or cortical abnormality.
No evidence of intracranial pathology
No evidence of morphological abnormality. Unremarkable Brain MRI.
No evidence of structural abnormality, infarction or hemorrhage.
No intracranial abnormality identified
No intracranial structural abnormality by MRI
No significant abnormality seen
No structural abnormalitiy identified
Normal
Normal age-appropriate MRI of the brain
Normal brain MRI
Normal brain MRI for age
Normal brain MRI without evidence of structural anomaly or other CNS abnormality.
Normal brain.
Normal contrast enhanced MRI of the brain
Normal conventional MR appearance of the brain
NORMAL EXAMINATION
Normal findings.
Normal MR appearance of the brain for the patient's age
Normal MR evaluation of the brain.
Normal MRI of the brain
Normal MRI of the brain for age without evidence of structural abnormality.
Normal MRI of the brain parenchyma
Normal MRI study of the brain
Normal myelination by MR imaging
Normal non-contrast MRI of the brain.
Normal nonenhanced MRI of the brain for age
Normal pediatric brain MRI and MRV, with no acute pathology identified to account for the seizure.
NORMAL PEDIATRIC BRAIN MRI WITH NO PATHOLOGY IDENTIFIED
Normal study
NORMAL STUDY FOR AGE
Normal unenhanced MRI of the brain
The brain appears normal. There is no evidence of a cortical dysplasia.
The brain is normal for age.
Unremarkable age appropriate MRI brain.
Unremarkable brain MRI
Unremarkable MR appearance of the brain. No brainstem or cranial nerve abnormality identified.
Unremarkable MRI brain. Myelination within normal limits of normal for age.
Unremarkable MRI of the brain
Unremarkable MRI of the brain with no evidence of intracranial mass or acute territorial infarction.
Unremarkable nonenhanced MRI of the brain
UNREMARKABLE STUDY

8.8. DICOM headers

MRN	Slice_Thickness	InPlane_Phase_Encoding_Direction
Quality_of_the_images	Repetition_Time	Flip_Angle
Integrity_of_the_volume	Echo_Time	Variable_Flip_Angle_Flag
Accession_Number	Inversion_Time	SAR
Weight	Number_Of_Average	Patient_Position
Study_ID	Imaging Frequency	Image_Position(Patient)
Study_Date	Imaged_Nucleus	Image_Orientation(Patient)
Institution_Name	Echo_Number	Frame_Of_Reference_UID
Manufacturer	Magnetic_Field_Strength	Slice_Location
Modality	Spacing_Between_Slices	Sample_Per_Pixel
Study_Despcription	Echo_Train_Length	Photometric_Interpretation
Series_Number	Percent_Sampling	Rows
Series_Description	Percent_Phase_FieldOfView	Columns
Group_Length	Pixel_Band_Width	Pixel_Spacing
Image_Type	Device_Serial_Number	Bits_Allocated
Scanning_Sequence	Software_Version	Bits_Stored
Sequence_Variant	Spatial_Resolution	High_Bit
Scan_Option	Reconstruction_Diameter	Pixel_Representation
MR_Acquisition_Type	Receive_Coil_Name	Pixel_Padding_Value
Sequence_Name	Transmit_Coil_Name	Window_Center
Angio_Flag	Acquisition_Matrix	Window_Width

8.9. Estimated parameters

WBA_{ADC}:

Linear: $f(x) = h + m \cdot x$

h=-5.12e-07 [mm^2/s] ; m=0.00134 [mm^2/s]

Logarithmic: $f(x) = a + b \cdot \ln(x)$

a=0.001491 [mm^2/s] ; b=-6.221e-05 [mm^2/s]

Biexponential: $f(x) = p \cdot e^{-x/s} + q \cdot e^{-x/t}$

p=0.0002419 [mm^2/s] ; s=97.38 ; q= 0.001137 [mm^2/s] ; t=1e+04

WBA_{FA}:

Linear: $f(x) = h + m \cdot x$

h=8.154e-05 ; m=0.2491

Logarithmic: $f(x) = a + b \cdot \ln(x)$

a=0.2246 ; b=0.009999

Biexponential: $f(x) = p \cdot e^{-x/s} + q \cdot e^{-x/t}$

p=0.2953 ; s=2.551e+05 ; q=-0.05807 ; t=173.8