Segmentation of the neonatal brain for MR images

A REVIEW

Tineke Hoppinga

19-5-2010

Inhoud

Abbreviations

BG	basal ganglia
centGM	central grey matter
cGM	cortical grey matter
CNR	contrast to noise ratio
CSF	cerebro spinal fluid
EM	expectation maximum
FSE	fast spin echo
GA	gestational age
GM	grey matter
GE	gradient echo
IR	inversion recovery
mWM	myelinated white matter
PA	post natal age
PD-w	proton density weighted
SE	spin echo
SNR	signal to noise ratio
SPGR	spoiled gradient echo sequence
subcGM	subcortical grey matter
T1-w	T1 weighted
T2-w	T2 weighted
TSE	turbo spin echo
unWM	unmyelinated white matter
WM	white matter

Abstract

The need for a good segmentation method for neonatal brain MR images increased over the last decade due to a significant increase of MR examinations in this group of patients. Adult brain segmentation methods are not suitable, because the intensity representation and shape of the neonatal brain are very different compared to adults. Furthermore, neonatal brain scans have a lower signal to noise and contrast to noise ratio, because of a limited scan time. Several groups of researchers designed a segmentation method which overcomes these specific challenges. This review compares twelve published neonatal segmentation methods by nine different research groups. The segmentation approaches described in these articles are compared on several subjects, including usage of atlases, segmented tissues, scan parameters, age of neonates, pre and post-processing. The results described in the article are also compared. The key problem for validation of neonatal brain segmentations is the absence of a gold standard. The ground truth described in the publications range from a single manually segmented slice to an entire manually segmented brain. The results show that the boundary of cerebrospinal fluid and cortical grey matter and the boundary of myelinated and unmyelinated white matter are the main problem areas for neonatal segmentation methods. The majority of the methods show good results, but it depends on many factors which method performs best in a clinical research environment.

Introduction

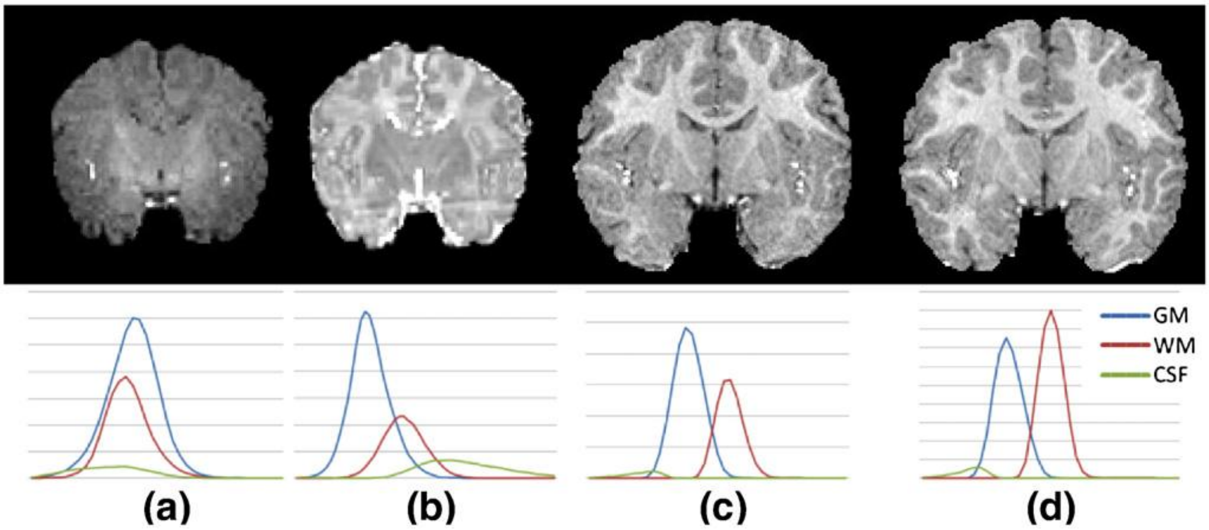
Over the last decade, there has been a significant increase in MR brain examinations for neonates who have been admitted to the intensive care unit. These infants have a higher risk of delayed neurodevelopment. MR has shown to be a useful tool for prediction of this issue [14][15][16][17][19]. An accurate segmentation method is important to study the development of different brain structures which can predict cognitive and behavioural impairment. Segmentation methods designed for adult brains are not applicable on neonatal brain scans. This has two main reasons.

First, the intensity representations of the neonatal brain tissues on MR scans are different compared to adults. The most obvious difference is the reversed contrast between white matter and grey matter on both T1-w and T2-w images. Furthermore, the intensity ranges of different brain structures are often overlapping. Figure 1 [7] shows the intensity representation of grey matter, white matter and cerebrospinal fluid on a T1-w image. In addition to the overlap of intensities, figure 1 also shows an increased intensity inhomogeneity for every tissue of the neonatal brain. This inhomogeneity is caused by the ongoing myelination process in white matter, which start about 36 weeks GA, and the higher water content in the brain in comparison to adults.

Second, the shape of a premature brain is significantly different from an adult brain and it changes rapidly. Especially the development between 30 and 40 weeks GA is impressive, as is illustrated in figure 2. The cortical surface changes in just 10 weeks from a smooth surface without gyrification and sulci to a cortex with nicely folded sulci and well developed gyri. In addition to the change in gyrification, the relative volume of several tissues in the brain also changes rapidly. This is well described by Hüppi et al [14]. For example, cortical grey matter changes from 20% to 40% of intercranial volume, whereas unmyelinated white matter changes from 48% to 31%.

Several groups of researchers have been working on a segmentation method specifically for neonates which overcomes the differences between adults and neonates. It is desirable for

Fig. 1. Representative longitudinal MR images and corresponding tissue intensity distributions of a subject. (a) T1 image at neonatal age; (b) T2 image at neonatal age; (c) T1 image at one year-old; and (d) T1 image at two-year-old (Shi et al 2007 [7])



research groups to have an automatic method which is able to segment a large cohort of neonatal brain scans, since manual segmentations are time consuming and suffer from high inter-rater variability. The published methods are separable into two approaches.

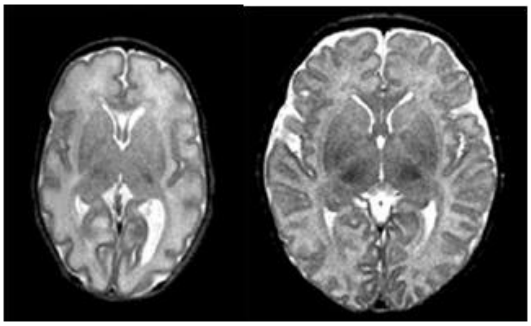


Fig 2. Left: T2-w image of a neonate scanned at 31 weeks; Right: T2-w image of a neonate scanned at 40 weeks (Hüppi et al 1998 [14])

The first segmentation approach is based on including previously acquired spatial information about brain tissues in the method. In 2000, Warfield et al was the first group to use an semi-automatic algorithm [13] for the segmentation of neonatal images. This algorithm is widely used in clinical studies [13][19], although it is not specifically designed for neonates and is also not validated on neonatal images. In 2005, Prastawa presented an automatic method explicitly designed for neonatal images [5]. A year later, Weisenfeld and Warfield [1] published a paper about a semi-automatic method for neonates which was followed by an automatic method soon after [2]. An improved algorithm followed a few years later [3]. In 2008, Anbeek et al presented an automatic method which includes a thorough validation of the volume measurements of brain tissues in addition to slice by slice segmentation. Xue et al [6] designed a segmentation method which focuses on the cortical area of the brain in order to perform cortex reconstruction. Shi et al is another group who incorporated spatial information in the method. This group published two papers [7][8] where they focus on different possibilities to make the spatial information more specific to the brain scan that needs segmenting.

Methods that belong to the second approach do not rely on previously acquired spatial information, but perform segmentation based on the information present in the test image. The first automatic algorithm with this approach was published by Song et al [10]. Merisaari et al [9] followed by a publication of their automatic method which they use for their clinical studies. The most recent publication is an abstract by Lisowski et al [12].

In contrast to the previously mentioned two segmentation approaches, Nishida et al [11] published a semi-automatic method which is designed to allow manual influence. Instead of segmenting a large cohort of neonatal brain scans, the purpose of this method is segmenting a large number of brain structures to evaluate brain growth.

The purpose of this review is giving an overview about the neonatal brain segmentation methods which have been published so far. In the next chapter, the proposed methods will be compared with each on subjects like segmented tissues, images used for segmentation, use of atlases and age of the infants. Furthermore, a basic explanation of all methods will be given and the pre and post processing steps will be compared. In the chapter 'results', the validation and results of the methods are described and this review finishes with a discussion.

Method

Basic explanations of the methods

Weisenfeld

In 2006, Weisenfeld et al [1] published their first article specifically designed for neonatal brain segmentation. The method is based on estimating the maximum likelihood of a certain class for a certain voxel. Since the likelihood of finding a given class is not similar throughout the image, a prior probability atlas for every class is used. To capture the spatial homogeneity of tissue classes, a Markov random field is implemented. This method is not entirely automatic. They perform an estimation of the tissue class properties by manually chosen regions to get a collection of labelled training samples.

Later they presented a fully automatic method [2] which is in fact the same method as [1], but with inclusion of a new atlas-based scheme for selecting the training points for supervised learning. With these training points, also known as a prototype list, a candidate segmentation from each atlas subject is constructed for the test subject. An optimal linear combination of these segmentations is then generated by the STAPLE algorithm [21] and when prototypes are likely to have an erroneous label, they are removed and the prototype list will be edited. This iterative process continues until convergence is reached.

In 2009 they published a new article [3]. This method is very similar to the automatic method[2], but with a few improvements like an implemented partial volume correction, based on the problem that Xue et al described in [6].

Anbeek

The base of this method [4] is the *k*NN-classifier, where *k*NN stands for *k*-nearest neighbours. It is based on assigning a sample to a certain class by searching for a *k* amount of samples in a feature space with the most similar properties. A multidimensional feature space is first constructed from a number of manually segmented scans. Based on the voxel properties from the test image, which includes both intensity and spatial information, the probabilities of a certain class are estimated using a *k* of 50 neighbours in the feature space. These probabilities are visualized in a class probability map. Once the manual segmentations are composed and the feature space is constructed, the method is fully automatic.

Prastawa

The method starts with rough estimations of the class intensity distributions for GM, WM and CSF, which are obtained from the test subject by choosing voxels with a high class probability in the atlas. The samples for GM and CSF which are acquired with this step are further processed to remove outliers and false positive of by obtaining a new estimate of the mean and covariance. The WM intensity distribution will be separated into unmyelinated and myelination WM by a graph based clustering method and afterwards by the class intensity distribution in the first step. Once the intensity distributions are known, an EM-algorithm is applied which combines the classification and inhomogeneity correction. The class posterior probabilities are computed using the intensity likelihood probabilities of the image and atlas spatial prior probabilities. In the last step the segmentation is refined using a kernel density estimation, because the Gaussian distributions of the tissues can have a significant overlap and result in degeneration of the tissue boundaries when estimating the maximum likelihood.

Xue

In this method [6], an EM based algorithm performs the segmentation. This EM algorithm starts with an initial distribution estimate from a blurred *k*-means clustering, to simulate an atlas. The *k*-means algorithm segments the image into GM, WM, CSF and background by iteratively computing a mean intensity for each class and classifying voxels to the class with the closest intensity centroid. Because of the partial volume effect, an extra first-order neighbourhood measurement (with 6 neighbours) is implemented in the EM scheme, modelled using a Markov random field (MRF). This identifies local conditions likely to lead to mislabelled partial volume voxels (MLPV). Prior knowledge for detection of the partial volume voxels, for example a WM voxel at the boundary of CSF and GM, is implemented. The next step is locally improving the segmentation, because the intensity variability causes local misclassifications. The brain is therefore devided in seven regions and for each region the segmentation is repeated. Eventually the cortex can be reconstructed for quantification of cortical changes during the early phases of brain development.

Shi

The method that Shi et al propose in [7] is also an atlas based method. Because this is a longitudinal study, they use the follow-up image as an atlas. First, the follow-up image is segmented with fuzzy clustering, which is derived from a *k*-means algorithm. Eventually the probabilistic maps of GM, WM and CSF are generated. Then a joint iterative registration-segmentation approach is adopted to perform bias correction, atlas alignment and tissue segmentation. For atlas alignment, a nonlinear registration method, named HAMMER [31], is used. This method increases the degree of warping for every iteration. For the classification process, an EM-algorithm estimates the new class posterior probabilities for every iteration. To overcome the overlap of the Gaussian distributions of the tissues, Shi et al use a 'mixture of Gaussians' model. In [8] the atlas is not a follow-up image, but a neonatal population atlas with cortical folding information from the test subject. A cortical GM sheet is acquired from the test image and this sheet is added to the population atlas. In this way, prior information of the test images is taken into account. Also a special dedicated neonatal head coil is presented in this article. This should give higher SNR and spatial resolution.

	Segmentation:
Weisenfeld [1]	Semi-automatic
Weisenfeld [2]	Automatic
Weisenfeld [3]	Automatic
Anbeek [4]	Automatic
Prastawa [5]	Automatic
Xue [6]	Automatic
Shi [7]	Automatic
Shi [8]	Automatic
Merisaari [9]	Automatic
Song [10]	Automatic
Nishida [11]	Semi-automatic
Lisowski [12]	Automatic

Table 7: are the methods semi-automatic of fully-automatic?

Merisaari

The method described in this article is a watershed based method [9]. First, they perform the watershed algorithm to get a rough segmentation of the image into many segments. This is done on the intensity gradient images instead of the standard T1 and T2 image. Afterwards the median intensity of each segment is calculated. Before merging the segments, the voxels marked as border voxels are included in the segments and the median intensity of each segment is calculated again. Instead of three clusters, one for each tissue, the merging of segments start with five clusters. The clustering with median intensities is performed by GMM, which is an EM algorithm with *k*-means clustering. Once the five clusters are made, the values in the smallest clusters are merged to the other clusters, which leads to three clusters, one for GM, WM and

CSF. Then the Gaussian distributions of each of the three clusters are convoluted with a Gaussian kernel. This will smooth the cluster borders and increase the stability of the last step, which is another GMM clustering with the difference that all intensities will be taken into account instead of just the mean intensities. Finally a myelination correction is performed.

Song

This method [10] is considered as a Bayesian interference problem in a Markov-random-field (3x3) framework. The Gibbs-energy needs to be minimized for an optimal class assignment and this Gibbs-energy consists of two terms. The first term is calculated with a likelihood term which is determined by a data-driven non-parametric probability density estimate, a intensity prior estimated by fuzzy non-linear support vector machines (SVMs) and a probabilistic atlas. These three terms are given weight factors. The second term in the Gibbs-energy formula is penalty based on the amount of different class labels in the Markov framework. When this method is used without an atlas, the weight factor of the atlas prior can be set to zero.

Nishida

This method is a semi-automatic method [11] which is based on intensity contour mapping. The user selects an intensity threshold manually which results in the definition of a specific anatomical border. Some structures can be segmented using histogram-based segmentation, but because the neonatal brain shows overlapping of tissues intensity ranges, as discussed in the introduction, this approach can be used less compared to adults. The user is able to edit the outcome so it is a highly interactive method which divides the brain into 30 different regions.

Lisowski

Comparable to the method by Merisaari et al, Lisowski presents an automatic method that is not atlas based. After noise reduction in the image, the first step of the algorithm is extracting the brain with a watershed algorithm. Then the brain is divided in left hemisphere, right hemisphere and cerebellum using a 3D distance transform of the union of GM and WM obtained from a *k*-means clustering with three clusters and a watershed applied to the sum of the gradient. Finally the CSF, GM and WM segmentation is performed on both hemispheres. CSF is classified first by selecting the brightest points belonging to the inner gradient, while WM and GM are separated by lambda thickening of an initial *k*-means clustering with two clusters. This k-means clustering is eroded to remove the partial volume voxels.

Tissues

Table 1 gives an overview of the tissues segmented by the twelve proposed methods. The table shows that GM, WM and CSF are the three most important structures to segment. A few methods can separate cortical GM from central GM or myelinated WM from unmyelinated WM. The different names for grey matter structures do not necessarily denote different grey matter regions. Central grey matter, as used in the article by Anbeek, is at the same location as subcortical grey matter, used by Weisenfeld [3] and the basal ganglia, used by Weisenfeld in [1] and [2]. This is therefore combined in table 1 to central GM. When an article states to segment GM, it means both cortical and central grey matter as one area. The same holds for WM. Lisowski segmented the cerebellum and brain stem in addition to the three common structures. Merisaari et al initially segment GM and WM as two different structures, but for validation they have the two structures combined. The reason behind this decision is unknown.

	WM	uWM	mWM	GM+WM	GM	centGM	cGM	CSF	Backg	Others
Weisenfeld [1]		x	x			x	x	x	x	
Weisenfeld [2]		x	x		x	x		x	x	
Weisenfeld [3]		x	x			x	x	x	x	
Anbeek [4]	x					x	x	x		
Prastawa [5]		x	x		x			x		
Xue [6]		x					x	x	x	
Shi [7]	x				x			x		
Shi [8]	x				x			x		
Merisaari [9]				x				x		
Song [10]	x				x					
Nishida [11]		x	x			x	x	x		x
Lisowski [12]	x				x			x		x

Table 1: tissues segmented by the twelve segmentation methods. Subcortical grey matter, used by Weisenfeld [3] and the basal ganglia, used by Weisenfeld in [1] and [2], are at the same location as centGM, as used by Anbeek. Therefore they are combined to one column.

Song et al is an exception in the list while they don't segment CSF. They remove this structure prior to the method, as well as the other brain structures that are not part of the segmentation process. As stated by Hüppi [14] and Peterson [15], CSF is one of the most important structures for analysis for preterm infants. They have a higher risk of ventricular abnormalities like impaired CSF dynamics which can cause cerebral atrophy. The method could therefore give more information about the condition of the infant if CSF is included in the segmentation.

Nishida et al is the only group that segments up to 30 different brain structures. As stated in the introduction, this method is not designed to be a fast segmentation method for clinical studies, it is designed for assessing brain growth in detail. The method is very time consuming. Apart from the brain structures mentioned in the table, Nishida et al also segment for example hippocampi, amygadalae, midbrain, pons, medulla, vermis and cerebellar hemispheres. The central grey matter is separated into caudate nuclei, lentiform nuclei, thalami and intervening white matter. [11]

Overall is shown that WM, GM and CSF are the most important structures to segment. Several groups have already separated myelinated white matter from unmyelinated white matter and central grey matter from cortical grey matter. It is likely that future segmentation methods focus on segmentation of more brain structures to increase the information obtained from neonatal brain scans.

Which images are used for segmentation?

Recent publications about segmentation methods designed for adults show that T1-w images have a better signal-to-noise ratio and contrast representation in comparison to T2-w or PD-w images. The T1-w images are therefore desired for segmentation. As stated in the introduction, the intensity representation in neonatal brain scans is

Method	Images
Weisenfeld [1]	T1-w SPGR and T2-w TSE
Weisenfeld [2]	T1-w SPGR and T2-w TSE
Weisenfeld [3]	T1-w SPGR and T2-w TSE
Anbeek [4]	T2-w TSE and IR
Prastawa [5]	T2-w TSE and IR
Xue [6]	T2-w FSE
Shi [7]	T2-w
Shi [8]	T1-w and T2-w
Merisaari [9]	T1-w and T2-w
Song [10]	T2-w SE
Nishida [11]	T1-w
Lisowski [12]	T2-w FSE and IR

Table2: Images used in the segmentation process

very different from adults, which means that T1-w images are not necessarily the best choice for segmentation of neonatal brain scans. This can be seen in figure 1. The intensity ranges between tissues are better separated on neonatal T2-w images. Especially the boundary between grey and white matter can be assessed better on a T2-w image. As shown in table 1 , nearly all groups of researchers choose the T2-w images for segmentation. To gain as much contrast information as possible, a T2-w scan is often combined with a T1-w scan. Especially for classification of myelinated white matter, the neonatal T1-w images give important additional information.

Instead of using a standard T1-w sequence, Anbeek [4], Prastawa [5] and Lisowski [12] use an inversion recovery (IR) scan, MPRAGE, combined with a T2-w image. When a correct inversion time is applied, IR sequences can be a good solution for enhancing T1 contrast.

Nishida et al is the only group who uses just T1-w images for segmentation. Given the reasons above, this is an interesting decision. It has probably to do with the manual influence incorporated in the method and the fact that a large amount of tissues are classified which are better assessed on T1-w images. Even though the intensity contour mapping with manual editing of a threshold seems to work fine, a T2-w scan can provide a better contrast representation for some tissues, especially the boundary between GM and CSF. This could make the choice for the segmentation threshold easier and give more accurate information about the boundaries in this area.

In contrast to adults, the T2-w scan is the most important sequence for neonatal imaging. Many groups use a combination of T1-w and T2-w images, because the T1-w images give useful additional information about myelination of white matter.

Scan parameters

Table 3 and 4 show that there is a significant difference in scan parameters, especially between the echo and repetition times (TE and TR). Because of myelination and increased water content, as discussed in the introduction, the relaxation times of the neonatal brain tissues are different in comparison to adult scans. Therefore, MR sequences designed for adults do not necessarily give the highest SNR or CNR in neonates. Jones et al [20] described the differences between neonatal and adult T1 and T2 relaxation times for several brain structures. The contrast between white and grey matter in neonatal brains is higher when a longer repetition and echo time is used. A consequence of a longer repetition time is a longer scan time. This is a disadvantage for neonates, because it will increase the possibility of motion artefacts. This

Scan parameters	T1	T2	IR
Weisenfeld [1]			
Weisenfeld [2]			
Weisenfeld [3]	TR30/TE9, flip angle 45	TR3000/TE140, Echotrain 15	
Anbeek [4]		TR7862/TE150, TSE 3	TR5086 /TE30/TI600, TSE 24
Prastawa [5]		TR7000/TE15 and 90	TR11.1/TE4.3/TI400
Xue [6]		TR1712/160, flip angle 90	
Shi [7]	TR1900/TE4.38, flip 7	TR7380/TE119, flip 150	
Shi [8]	TR1820/TE4.38, flip 7	TR9280/TE119, flip 150	
Merisaari [9]			
Song [10]			
Nishida [11]	TR30/TE8 flip 25 to 30		
Lisowski [12]		TR4600/TE150, Echotrain 15	TR2500/TE2.89/TI1100

Table 3:The repetition (TR), echo (TE) and inversion times(TI) used for the methods. All times are in milliseconds

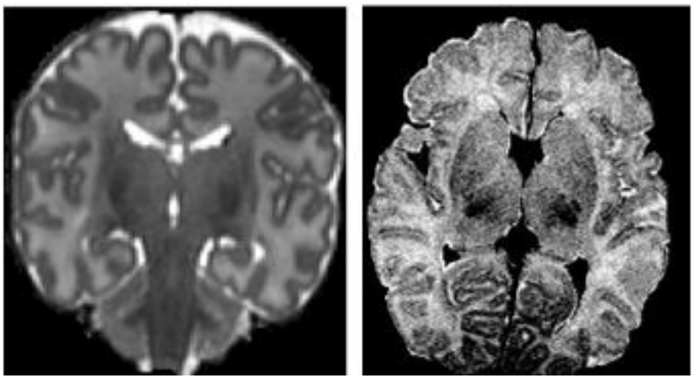


Fig. 3 Coronal T2 imaging from the article of Xue (left) compared to a transversal T2 imaging from Song (right). The higher resolution of the right image causes for higher noise levels and lower contrast.

results in compromising with other scan parameters to reduce the scan time. As shown in table 4, every methods describes different scan parameters which makes them difficult to compare. The trade off between parameters in MR imaging means that there are more ways to achieve good quality scans. This is not just the solution described by Jones et al. Unfortunately, the acquisition times are not present in the nearly all publications. It is known that the method by Prastawa performs the total MR examination in ten minutes, which includes a T1-w and T2-w sequence. Lisowski discussed acquisition times of 7 minutes for the IR sequence and 5.5 minutes for the T2 sequences.

It can be noticed from table 3 that Song uses a very small voxel size. This is unusual in neonatal imaging. It will reduce the problem of partial volume voxels, but if the scan time is limited due to the motion of the infants, it is not possible to achieve a high signal to noise ratio. The images shown in the publication by Song et al [10] seem to have a higher noise level compared to images from other publications. An example is show in figure 3.

Scan parameters	**Tesla**	**matrix**	**Direction**	**Slice thickness (mm)**	**Slices**	**FOV**	**Voxelsize (mm)**	**Coil**
Weisenfeld [1]	1.5	256x256	Coronal	T1: 1.5 ; T2: 3.0			0.7x0.7x1.5/3.0	Head-surface
Weisenfeld [2]	1.5	256x256	Coronal	T1: 1.5 ; T2: 3.0			0.7x0.7x1.5/3.0	Head-surface
Weisenfeld [3]	1.5	256x256	Coronal	T1: 1.5 ; T2: 2.0		18 to 22cm		8 channel head(adult)
Anbeek [4]	1.5	256x256		2.0mm	50	18x18cm		
Prastawa [5]	3.0		Sagittal	T1: 1.0 ; T2: 1.95	T1 128/T2 56		0.898x0.898mm2 T1 / 1.25x1.25 mm2 T2	
Xue [6]		224x224	Transverse	2.0mm		22cm	0.86x0.86x2.0	Standard 6 ch coil
Shi [7]	3.0	256x192 T1 / 256 x128 T2	Axial	T1: 1.0 ; T2: 1.95	T1 160/T2 70		1x1x1mm2 T1 / 1.25x1.25x1.95 T2	
Shi [8]	3.0						0.79x0.79x0.8 T1 1x1x1.3 T2	Neonatal 8 ch
Merisaari [9]	1.5	Between 129x129 and 256 x256			11 to 22		0.78x0.78x4.4 for all slices	
Song [10]		512x384 x28 ???	Axial				0.35x0.35x3mm	
Nishida [11]	1.5	256x192	Axial	1.2 to 1.4		220x165 or 200x150mm	0.8/0.9x0.8/0.9x 1.2mm	Pediatric coil, but where possible neonatal
Lisowski [12]	3.0		Coronal	1.2			0.8x0.8x1.2mm	

Table 4: repetition times (TR), echo times (TE) and inversion recovery times (TI) per article for the T1, T2 and IR weighted images. All times are in milliseconds.

An improvement for both signal intensity and shortening of scan times can be achieved by using a MR system with a higher field strength. A few studies have already been performed on a 3T system, rather than a 1.5T system. When using the same scan time as a sequence on a 1.5T system, the signal to noise ratio increases on a 3T system. Therefore the choice can be made to reduce the scan time by a factor 4 to keep the same quality as a 1.5T system, which is very important in neonatal scanning, because of the reduction of motion artefacts [30]. A requirement for good results is again the tuning of the parameters. Parameters used for adults will not necessarily give the best signal and contrast ratios for neonates. To reduce the partial volume effect, one can also decide to reduce the voxel size instead of decreasing the scan time.

Another improvement of signal to noise ratio can be the use of a neonatal head coil. Many methods use a standard adult head coil, but the head of neonates are smaller. A more compact coil gives better signal sensitivity. Both Nishida et al [11] and Shi et al [8] tried this approach, but it is difficult to judge from the scan parameters and images from the publication whether the SNR and CNR actually increase. This needs to be evaluated with further research.

In conclusion, image quality is always a trade-off between scan parameters. It is important to achieved a good SNR and CNR in images, because segmentation methods perform better on a scan with well defined boundaries and a low noise level. Scan time is limited for neonates, which increases the importance to determine optimal scan parameters. Further research can possibly improve SNR and CNR for neonatal scans which enhances the performance of the segmentation methods.

Use of atlases

The majority of the research groups implemented previously acquired spatial information about brain tissues into their method. Especially if the contrast between brain tissues is low and if the intensities ranges of tissues are overlapping, spatial information of tissues is an useful addition to intensity based segmentation. Spatial information, also known as spatial priors, can be implemented in different ways. The probability atlas is the most popular solution, as seen in table 5. For adults, the International Consortium of Brain Mapping (ICBM) has accepted a construction standard for these atlases. However, the shape and intensity representation of a neonatal brain is significantly different from adults. If an adult probability atlas is aligned with a neonatal brain, a high non-linear deformation will appear which can cause registration errors [18]. For minimizing this deformation, it is important to choose atlas subjects with an age similar to the test subject. This is especially important for neonatal brain segmentation due to the rapid development of the brain between 25 and 42 weeks.

	Use of atlas?	Type of atlas?	How constructed?
Weisenfeld [1]	Yes	Probability atlas	Semi-automatic
Weisenfeld [2]	Yes	Probability atlas	Semi-automatic
Weisenfeld [3]	Yes	Probability atlas	Semi-automatic
Anbeek [4]	Yes	Feature space	Manual
Prastawa [5]	Yes	Probability atlas	Semi-automatic
Xue [6]	Yes	Probability atlas	Manual
Shi [7]	Yes	Probability atlas	Semi-automatic
Shi [8]	Yes	Probability atlas	Semi-automatic
Merisaari [9]	No	-	-
Song [10]	Yes/No	If yes, Probability atlas	Semi-automatic
Nishida [11]	No	-	-
Lisowski [12]	No	-	-

Table 5: Usage of atlases in the segmentation

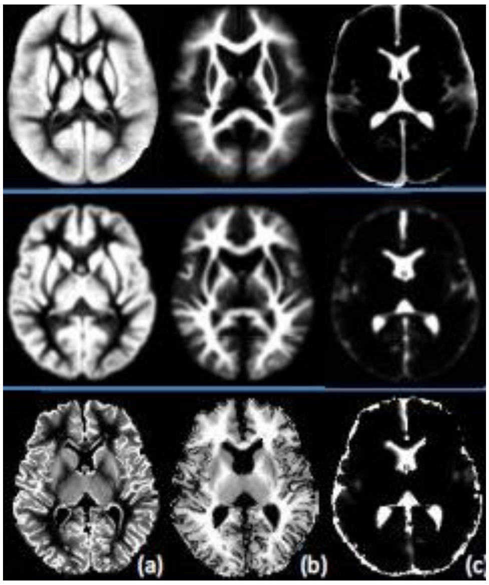
The spatial priors are usually incorporated into an iterative method to search for the highest probability of a certain class at a certain voxel. However, the construction and implementation of the spatial priors are different between methods. Prastawa [5] constructed an atlas from three neonatal brain scans which are segmented by a semi-automatic method, based on a *k*-nearest neighbour algorithm where a human rater chooses the samples for each tissue type on beforehand and edit the outcome of the classification. Because three scans is a small amount for the creation of an atlas, Prastawa et al blurred the atlas to simulate a larger population. This is not an elegant solution. Increasing the amount of atlas subjects gives a more realistic representation of the spatial information of the brain tissues.

Weisenfeld used in method [1] twenty infants, in method [2] thirteen infants and in method [3] fifteen infants for atlas construction. In all cases these images are segmented by an earlier proposed method [12]. In their three articles, the usage of the atlas is similar. The base of the method is to use several templates with a large number of manually selected voxels, called prototypes. These prototypes are transferred to the test image where they result in a prototype list of density estimates which will lead to segmentations. These segmentations are then fused to get a probabilistic labelling of the classes. The prototype lists will then be edited to reject the prototypes which are likely to have an erroneous label. In this calculation, the spatial priors of the atlas have a role. The iterative process of prototype editing will continue until convergence.

Shi et al proposed different atlas construction approaches in both publications [7][8]. The intention is to incorporate information of the test subject into the atlas which will result in a better segmentation performance. The focus is enhancing performance in the cortical area of the brain. The first approach [7] uses a follow-up image of the same infant as an atlas. The advantage of this solution is that the cortical folding is similar in both test subject and atlas subject which enhances the segmentation performance in the cortical area. The cortical folding is at the same time also a disadvantage, because the test subject needs to be at least term age otherwise the cortical folding is not yet developed. A second disadvantage is that this method can only be used in a research environment, because a follow-up image is not available in a clinical situation. The approach used in the second publication [8] is more useful in a clinical environment. The atlas used in this approach is a neonatal population atlas with inclusion of cortical folding information of the test subject. This causes for a better representation of the tissue probabilities in the cortical area of the atlas which enhances the segmentation performance.

From the atlas based methods, Anbeek is the only group who does not use a probability atlas. Because their method is based on a *k*NN classifier, the 12 manually segmented subjects are not used to construct a standard atlas, but they are used for building a multidimensional feature space. Based on the voxel properties, like intensity and spatial information, the assignment of a certain class for a certain voxel depends on the class label of 50 neighbours in the multidimensional feature space.

Fig. 4 Atlas probability maps of a population atlas (top), neonatal atlas (middle) and subject specific atlas (bottom) from the article of Shi et al [8]. a-c are GM, WM and CSF priors respectively



Xue et al [6] use an atlas just for removing unnecessary tissues from the test image before the segmentation process. This atlas is a single manually segmented neonatal brain scan which is registered to the test subject. As stated earlier, atlases are often incorporated into the iterative search for the highest probability for a certain tissue class for a certain voxel.

Xue et al decided not to use a population atlas for this purpose, because of the rapid brain development in neonates. To initialize the iterative segmentation process, a *k*-means clustering of the test subject is performed. This clustering is then blurred and normalized to simulate an atlas.

Song et al designed a segmentation method which has a possibility to include an atlas, but it is not necessary. The atlas is basically an extra feature in the class probability calculations which has a weight factor of zero if the atlas is not included.

	Atlas			**Test subject**		
	Nr of infants	**Age: born**	**Age: scanned**	**Nr of infants**	**Age: born**	**Age: scanned**
Weisenfeld [1]	20	After 28 wks	42 wks	5	Prior to 28 wks	42 wks
Weisenfeld [2]	13	After 28 wks	42 wks	5	Prior to 28 wks	42 wks
Weisenfeld [3]	15	After 28 wks	42 wks	10	Preterm to term	42 wks
Anbeek [4]	12	25.9-42.9 wks	43 wks	13	25.9-42.9 wks	43 wks
Prastawa [5]	3	-	-	4	-	-
Xue [6]	3 groups, 1 per group	27-34 wk, 34-39 wk, 39-45 wk	27-34 wk, 34-39 wk, 39-45 wk	25	24-40 wks	27-44 wks
Shi [7]	10	-	PA 60.5± 6.8wks or PA 101.4±5.4wks	10	-	PA 6±1.9 wks
Shi [8]	68	-	PA 1.3±0.7 months	10	-	PA 0.5-1.9 months
Merisaari [9]	-	-	-	11	-	40wks
Song [10]	-	-	-	10	term	Within 10 days
Nishida [11]	-	-	-	12	-	31.1-42.6 wks
Lisowski [12]	-	-	-	?	GA 30-40 wks	GA 30-40 wks

Table 6: Age of the infants, both from the atlas subjects and the test objects

Patient groups

The age of the infants, as described in the publications, is well comparable. The majority of the infants is scanned about term age, as shown in the last column of table 6. All methods, except for Shi et al, describe age as gestational age (GA). Shi uses postnatal age (PA), which means that it is not possible to determine the exact gestational age of the infants at the moment of MR examination. It is therefore difficult to compare the age of the test subject and atlas subjects for paper [8]. In the first publication by Shi [7], the atlas is a follow-up image of the same infant. It is either an image of one or two years PA. The disadvantage is that the brain of a one or two year old is more developed and has a different intensity representation than a neonate. The results do not reflect this disadvantage. The reason is most likely the age of the test subject. The infants are older in comparison to the other publications and that would mean that the brain is more developed compared to preterm or term born infants. However, this is hard to verify, because it is not clear what the gestational age of the infants was at time of birth.

As discussed before, it is important that the age of the test subject en the atlas subjects are similar. When the role of the atlas becomes larger is it likely that the influence of the age of the infants is more important to the method. The structural differences between the brain of preterm born infants and term born infants are large (fig 2), therefore the age of the atlas subjects should be the similar to the test subject and the age range is preferred to be as small as possible. Prastawa [5] did not mention any age in the method. It is therefore difficult to determine whether the atlas is constructed from infants of the same age as the test subject. The atlas in their method is used for an initial estimation of the probability distribution of the class.

If the initial estimation is more accurate, the chance of an accurate segmentation is higher. The same holds for the methods of Weisenfeld, Xue and Shi. Weisenfeld clearly stated in article [1][2] that the images used for atlas construction and for testing are all from prematurely born children, scanned at 42 weeks. The fact that the atlas subjects are born after 28 weeks and the test subjects are born before 28 weeks should not be a problem. However, neonates born prior to 28 weeks have a higher chance of dilated ventricles, underdeveloped gyri or other pathology at term age.

Three research groups describe large age ranges in their publications. Xue et al [6] used infants between 27 and 44 weeks for evaluation of the segmentation performance. To overcome dissimilarities in brain shape and intensity representation between test subject and atlas subjects, the age range of the atlas subject needs to be as small as possible. All infants are therefore divided into three groups: 27-34 weeks, 34-39 weeks and 40-45 weeks. The two other publications with larger age ranges are methods without atlas priors. Lisowski [12] describes test subjects between 30 and 40 weeks old. This is not a problem, because the method does not rely on an atlas. Nishida [11] uses infants of different ages to study brain growth from 31 to 42 weeks GA.

	Pre-processing	**Post-processing**
Weisenfeld [1]	-Intensity inhomogeneity correction -noise reduction -intra-subject registration for motion	
Weisenfeld [2]	-Intensity inhomogeneity correction -noise reduction -intra-subject registration for motion	
Weisenfeld [3]	-Noise reduction -registration	15 min post processing (no explanation)
Anbeek [4]	-Registration of images -mask construction	Threshold for binary image of a probability map
Prastawa [5]	-Intensity inhomogeneity correction -noise reduction -motion correction -registration	Segmentation refinement
Xue [6]	-Intensity inhomogeneity correction -brain extraction -registration	-Local segmentation -Construction of cortex
Shi [7]	Brain extraction	
Shi [8]	Brain extraction	
Merisaari [9]	-Intensity inhomogeneity correction -brain extraction	-Myelination correction, -For evaluation: fat at brain surface was manually removed -Combination of GM+WM
Song [10]	-Intensity inhomogeneity correction -noise reduction -contrast enhancement	
Nishida [11]	Visual optimization of data	
Lisowski [12]	Intensity homogeneity correction	

Table 8: Pre and post processing steps which are implemented in the proposed methods

Pre-processing

Before the actual segmentation starts, all methods perform pre-processing to improve the data in order to get the best segmentation result. As shown in the table 8, intensity correction, noise reduction and registration are the main implemented pre-processing steps.

Intensity inhomogeneity correction

An artefact that often occurs in MR images is a smoothly varying signal intensity across the images. This is caused by several factors such as inhomogeneous radio frequency excitation, non-uniform signal reception sensitivity and electrodynamic interactions with the object. This intensity inhomogeneity does not have a significant impact on visual inspection, but the performance of segmentation methods can decrease with the clinical accepted levels of non-uniformity [22][24].

Lisowski implemented one of the simplest corrections. It is an anisotropic diffusion correction by Perona et al [29]. Prastawa [5] implemented the inhomogeneity correction described by van Leemput et al [25] within their iterative classification method. It is therefore not strictly pre-processing. The method uses the spatial posterior probabilities to estimate the inhomogeneity within a tissue class. Based on the estimation of the classification and distribution parameters, the measured signal with non-uniformity is predicted. It is an iterative process which is initialized with the Gaussian distributions of the classes that are calculated a step earlier in the segmentation process.

Merisaari [9] and Xue [6] use a semi automatic method published by Sled et al [24], the correction method N3. It is a method which is separated from the classification process. Prastawa rejected N3 because it is a histogram-based correction method. The histogram of neonates is much smoother compared to adults and has weak maxima. Relying on the histogram is a potential problem for estimating the inhomogeneity for neonates. Merisaari choose this method, because the it does not depend on the definition of a model for intensity values in homogeneous regions. The information needed for the correction field can be derived solely from the histogram.

Weisenfeld et al also use a method which depends on data priors in [1][2][3]. They used and extended a method by Mangin [22]. It is a modelling technique that assumes the intensity non-uniformity is smoothly varying compared to the contrast within the images and it can be modelled as a field of multiplicative factors. The goal is decreasing the entropy in the image histogram. Finally, Song [10] used an inhomogeneity correction based on an method described by Likar et al [23]. This method assumes that the acquired image contains more information than the true image. By changing the parameters in a correction model applied to the acquired image, the additional information induced in the degradation process is minimized.

Noise reduction

As a method to increase the signal to noise ratio (SNR), the anisotropic diffusion smoothing techniques which preserve edges are popular. Weisenfeld used a curvature based method in article [1] and a standard anisotropic diffusion method in [2]. They both come from the same source, which makes it not entirely clear if it is the same method. In [3] they also used a anisotropic diffusion edge preserving technique, but this time they adapted the method published by Perona and Malik (1990) and Gerig (1992). This is the same method that Prastawa adapted in [5]. Song [10] also used edge preserving anisotropic smoothing, but it is unknown which method.

Brain extraction

Several methods perform brain extraction prior to the segmentation process in order to remove structures which are not interesting for segmentation. The Brain Extraction Tool (BET) [31] is a popular method and is used is different ways by Xue, Shi and Merisaari.

Xue [6] uses BET only for the brain extraction of the atlas subjects. The brain of the test subject is extracted via registration with an atlas subject.

In the first publication by Shi [7], BET is used in combination with a second brain extraction tool, brain surface extractor (BSE) [32]. The result is manually edited to ensure satisfactory results. In the second publication by Shi [8], only BSE is used for brain extraction. No manual editing is reported in this publication. Apart from removal of the skull, the cerebellum is also removed in [7] and [8]. This is done manually in [7] and semi-automatically in [8], which means that the entire segmentation process proposed in both publications is not fully-automatic.

Instead of manually editing the result of brain extraction, Merisaari et al [9] modified the BET algorithm to make is suitable for neonatal brain scans and add the ability to work with gradient images instead of the standard brain scan. The main modification is implementing a more simplified way to project the mesh on the brain surface with support of their combined T1/T2 gradient images.

Anbeek et al [4] implemented a brain extraction step without use of BET or BSE. They create a mask of the region of interest by performing a *k*-means clustering on the T2-w images with two clusters, one for brain tissue and one for background. This results in a binary image.

	Atlas creation	**Atlas to test subject**	**Intra subject**	**Other**
Weisenfeld [1]	Non-rigid	?	?	
Weisenfeld [2]	Affine, using mutual information	Affine, using mutual information	Rigid, using mutual information	Prototypes to test subject: non-rigid
Weisenfeld [3]	Affine	Non-rigid	Rigid	Prototypes to test subject: non-rigid
Anbeek [4]	-	-	Rigid	
Prastawa [5]	Affine	Affine	Affine	
Xue [6]	-	Non-rigid	-	
Shi [7]	-	Non-rigid	-	
Shi [8]	Non-rigid	Non-rigid	?	
Merisaari [9]	-	-	Affine	
Song [10]	-	-	-	
Nishida [11]	-	-	-	
Lisowski [12]	-	-	?	

Table 9: Registration bases that are described in the methods to register the atlas subjects to each other, to register the atlas to the test subject and to align the MR images.

Registration

Atlas registration

Methods [1] to [8] are atlas based methods which require registration of the atlas subjects and registration of the atlas to the test subject. Table 9 shows that this is done with either affine of non-rigid registration. Weisenfeld [2] and Prastawa [5] choose affine registration for both registration steps. Weisenfeld decided to change to non-rigid registration for atlas to test subject registration in [3]. This is more computationally expensive, but it aligns the class probabilities present in the atlas in a more accurate manner to the test subject. This will improve the performance of the segmentation method.

Xue et al [6] also use non-rigid registration to align an atlas to the test subject which is necessary in this case, because the purpose is removal of the central deep tissues and skull stripping. Affine registration would misalign the boundaries between these tissues and brain tissue.

Shi et al combined registration with segmentation in both [7] and [8]. The registration step is a method called HAMMER [33]. The atlas is aligned to the test subject iteratively with a non-rigid process. The atlas is warped globally in the initial stage and locally in the later stages.

Intra subject registration

Several methods use both T1-w and T2-w images for segmentation. Due to movement, the images need to be aligned before segmentation can be performed. Weisenfeld [2][3] and Anbeek [4] used rigid registration for this purpose where Prastawa and Merisaari use affine registration. It is probably dependent on the amount of movement experienced during the MR examinations whether affine registration is necessary. In theory, the shape of the head does not change, which means that the translation and rotation steps of rigid registration are sufficient.

Weisenfeld, Anbeek and Prastawa use mutual information to register the T1-w image and T2-w image with each other. Merisaari et al uses a similar approach, developed by Studholme et al [28]. Their method is a normalized entropy measure which is based on the ratio of the sum of marginal entropies and the joint entropy to align images.

Other remarks

Apart from noise reduction, brain extraction and registration, a few research groups implemented other pre-processing steps. Song [10] implemented a contrast enhancement technique based on adaptive histogram equalization [34]. Nishida [11] implemented only one pre-processing step which is repositioning of the head in three dimensions. This is not a registration method, because all images are repositioned separately. A reference plane is positioned on the image which bisects the decussations of the anterior commissure (AC) and posterior commissure (PC), and the interhemispheric fissure at the level of the PC in the coronal plane. After repositioning, a tri-linear interpolation method re-samples the MR image data to a new resolution in all planes.

Post-processing

An important issue for segmentation methods is the amount of post-processing needed after segmentation has been performed. As shown in table 8, several research groups report post-processing steps in their methods. Weisenfeld [3] mentions 15 minutes of post-processing in the publication, but it is not specified which steps are implemented. Anbeek et al included a small post-processing step. The outcome of the segmentation process is a probability map of tissue classes. When a binary map is preferred, simple thresholding can be performed with a predefined threshold. This step is fully automatic. Merisaari [9] reports combining GM and WM to one structure after segmentation is performed. GM and WM are initially segmented as two separate structures. The reason behind this decision is not given.

Apart from the previous small post-processing steps, several methods implemented corrections or refinements after the main segmentation step. It can be discussed whether these small corrections are post-processing steps or if they are included to the segmentation process. Merisaari [9] performed an automatic myelination correction based on Prastawa's myelination correction described in [5]. This step corrects misclassified voxels due to overlap of intensities ranges between tissues. Prastawa only performed this technique on WM, while Merisaari extended it to perform on WM and CSF. Prastawa implemented another segmentation refinement step. It is a fully automatic step which incorporates sampling of the inhomogeneity corrected images from the previous step. Samples with the highest probability of a certain class are combined with the atlas priors and kernel density functions to calculate new intensity distributions. These new intensity distributions capture the complex intensity representation of neonatal brain scans.

Xue et al performed local segmentation as post-processing step after the main segmentation step is performed on the entire image. The global segmentation is divided in seven regions and the same segmentation step is repeated for every region. This reduces the influence of intensity variability in tissues and corrects the border between tissues. After global and local segmentation is performed, is it possible to reconstruct the cortex.

Results

Validation

Ground truth

In order to perform validation, segmentation results need to be compared to a gold standard. Unfortunately there is no gold standard available for neonatal brain scans. Therefore all research groups, except Nishida et al, have constructed a gold standard from either manually or semi-automatically segmented images. As shown in table 10, this construction of a gold standard is significantly different between groups. Weisenfeld [1][2][3] and Prastawa use just one midcoronal slice to determine the performance of their segmentation method. Weisenfeld segmented the brain scans five times by three raters to decrease the inter-rater variability in publication [3]. This means that the ground truth is constructed from 15 manual segmentations. Unfortunately, one midcoronal slice is not a good representation of the segmentation performance in the entire brain even though all classes are present on this midcoronal image. In contradiction, Anbeek [4] and Merisaari [9] use manual segmentations of the whole brain as gold standard.

	subjects	raters	Repeated?	slices	How acquired?
Weisenfeld [1]	5	1	No	1 midcoronal slice	Manual
Weisenfeld [2]	5	1	No	1 midcoronal slice	Semi-automatic
Weisenfeld [3]	4	3	Yes, 5 times	1 midcoronal slice	Manual
Anbeek [4]	13	2	Yes, 2 times	Whole brain	Manual
Prastawa [5]	4	2	No	1 coronal slice	Manual
Xue [6]	25	1	No	3 orthogonal slices	Manual
Shi [7]	10	2	No	2 sagittal, 3 coronal and 3 axial slices	Manual
Shi [8]	10	-	-	Unknown	Semi-automatic
Merisaari [9]	Volume: 11 2D: 3	1 1	No	Whole brain (10-15 slices) Whole brain (10-15 slices)	Manual
Song [10]	10	2	No	Unknown	Manual
Nishida [11]	-	-	-	No ground truth	-

Table 10: Ground truth construction per article with number of subjects, numbers of expert raters, the amount of times the raters repeated the ground truth segmentation, which slices are used as ground truth and how the segmentation is acquired.

Similarity measurements

The validation of the segmentation methods is performed by three different similarity measurements. The first measurement is the percentage of overlap between the ground truth and the segmentation. Merisaari et al is the only group to use this measure. They define it by the following formula:

$$q_S = M_S/E_S * 100,$$

where Ms is the number of voxels belonging to segment S as determined by the automatic method and Es is the number of voxels determined manually. The description of this method is unclear, since the result depends on the definition of segment S. Segment S can consists of all voxels determined by the manual segmentation or by all voxels determined by both the manual of automatic method. In both cases the measurement is not a true similarity measure, because true negatives, false positives and false negatives are ignored. This does not give a good representation of the similarity between segmentations.

The second similarity measurement is the Dice similarity measure [26]. Consider two segmentations A and B, the Dice coefficient is defined as

$$2|A \cap B|/(|A| + |B|)$$

The overlap will be normalized where 0 means no similarity and 1 means complete similarity. This measure is used in nearly all methods.

The third measure is Cohen's *k* [27]. Prastawa uses this measure in addition to Dice for the level of agreement between manual and automatic segmentation and for inter-rater variability. Merisaari also mentions this measure for their slice by slice comparison between gold standard and proposed method. The *k* value is calculated by the following formula

$$\kappa = \frac{\sum_{i=1}^{N}\mathrm{agree}(C_i) - \sum_{i=1}^{N}\mathrm{ef}(C_i)}{N - \sum_{i=1}^{N}\mathrm{ef}(C_i)},$$

where N is the number of classes, agree(Ci) is the number of agreements between the raters for class Ci and ef(Ci) is the expected frequency of agreement by chance for class Ci. This expected frequency is calculated by

$$\mathrm{ef}(C_i) = \frac{1}{N} a_i b_i.$$

Where N is the number of classes and ai and bi are the number of observations of class Ci. The k value is normalized and 0 means independence and 1 indicates complete agreement.

Inter-rater variability

As stated in the previous section, many research group construct the gold standard from manual segmentations. This is either done by one or more expert rater, once or multiple times. A few research groups discuss the inter-rater variability in their publication. To increase the reliability of the segmentation performance, the difference between the experts should be as small as possible. Prastawa [5] reported a Dice-overlap for tissues classes between 0.639 and 0.787 for two expert raters. Furthermore, the Cohen k-coefficient, which measures agreement between manual segmentations, shows results below 0.658. These numbers are not very high, which decreases the reliability of the gold standard.

Xue et al [6] and Nishida et al [11] reported better inter-variability results. The Dice similarity described by Xue was 0.874±0.034. Nishida showed overlap results between 80% and 90%. However, this can be caused by the fact that one rater segmented the same brain scans twice, instead of two raters segmenting the same brain scan. One rater segmenting a brain scans with an interval of 4 weeks (intra-rater variability) is not comparable to inter-rater variability. Other publication with good inter-rater variability were Song et al [10] with a Dice overlap between 0.8 and 0.85 for GM and between 0.75 and 0.875 for WM. Shi et al [7] described inter-rater variability between 0.85 and 0.95, Weisenfeld [3] showed results between 0.8 and 0.95.

Overall, the majority of publications show good inter-rater variability which increases the reliability of the segmentation results. However, Prastawa showed lower Dice overlap results, which means that it is not self-evident that manual segmentations have a good overlap.

	Inter-rater variability	Ground truth to proposed method?	Proposed method to other methods?	Volume analysis?	Remarks
Weisenfeld [1]	No	Yes	Yes, [13]	No	
Weisenfeld [2]	No	Yes	No	No	
Weisenfeld [3]	Yes	Yes	Yes, [13]	No	Extra: different settings for chosen prototypes: automatic edited, automatic unedited, manual
Anbeek [4]	No	Yes	No	Yes	
Prastawa [5]	Yes	Yes	No	No	Volumes are calculated but not compared.
Xue [6]	Yes	Yes	No	No	Extra: visual assessment of label propagation, partial volume voxels and cortex reconstruction 3D vs 2D slices
Shi [7]	Yes	Yes	No		Extra: different atlas sharpness parameters
Shi [8]	No	Yes	No	Yes	
Merisaari [9]	No	Yes	Yes, SPM2, GMM-T1, HMRF	Yes	Extra: different myelination correction factors; Combinations of WSEG with SPM2 and GMM-T1 in CSF region
Song [10]	Yes	Yes	No	Yes	
Nishida [11]	Yes	No	No	Yes	Volumes are plotted as a function of gestational age

Table 11: This table shows the various validation methods used per article for 2D and/or volume analysis. Overall the proposed method is compared to the ground truth, inter-rater variability is discussed.

Voxelwise comparison to the gold standard

Every research group has their own strategy of comparing the segmentation results with the gold standard. It is highly dependent on the described method. Table 11 shows a summation of the validation methods described in the publications. Table 12 shows an overview of the validation results per method.

Weisenfeld

In paper [1], Weisenfeld et al compared the proposed semi-automatic method to the ground truth, to an older method [13] and to the proposed method without an atlas. The results are measured using the Dice similarity index. The proposed method performs better than method [13] for nearly every tissue class and has an overlap of 0.65 to 0.75 with the gold standard. In two tissue types, CSF and unWM, the proposed method without atlas priors appears to work better or equal to the proposed method with atlas. This is remarkable, since the other tissues clearly benefit from the atlas. Apparently, the intensities of these two tissue types are so distinct from the other tissue types that the segmentation method does not need assistance from atlas priors.

The fully automatic follow-up [2] was only compared to the gold standard. In this publication, the similarity measurement range from 0.68 for CSF to 0.74 for unWM. These results are very similar to their previous publication [1].

In publication [3], Weisenfeld et al choose to perform a few additional validation methods. To validate the entire automatic method, they compared it to a previously published method [13] in 10 patients. The Dice similarity ranges from 0.72 for myelin to 0.92 for CSF. The editing process of the automatically selected prototypes is also evaluated. The best results are achieved after four iterations (mean 0.95±0.01), which means that after four iterations 95% of the prototypes have the same class label as the manually selected prototypes. Eventually the automatic segmentation is compared to the manual segmented gold standard. In addition to the manual gold standard, an expert consensus was constructed by a method called STAPLE [21]. This iterative method computes a probabilistic estimate of the true segmentation. This is done by considering a number of segmentations, which may be constructed by human raters or other segmentation methods. It then measures the performance level of each of the segmentation and combines it to one expert consensus. The Dice overlap between the gold standard and the expert consensus is 0.86±0.05, which is a good agreement. The expert consensus compared with the proposed method shows an overlap between 0.74±0.05 for unWM and 0.96±0.02 for CSF. This is a increase in performance in comparison to method [2].

Anbeek

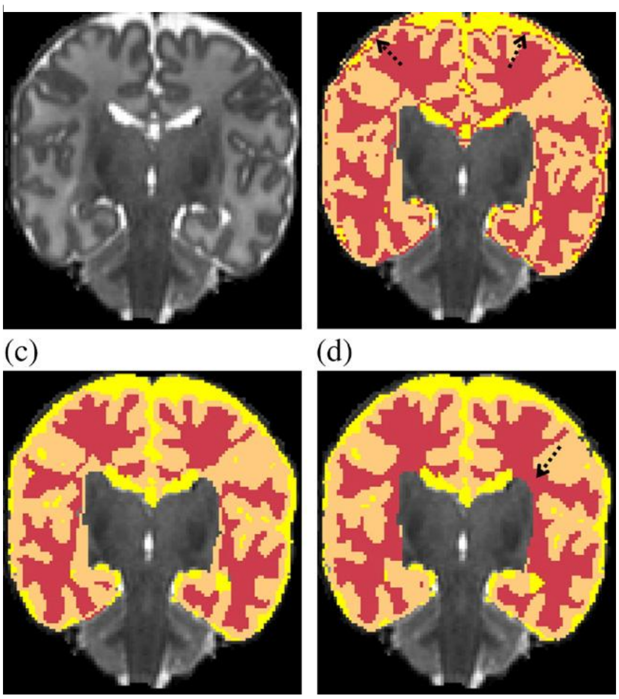
The segmentation results of Anbeek [4] ask for a different validation approach compared to the publications of Weisenfeld. The outcome of the classification is not a direct segmentation, but a probability map which can be processed into a binary segmentation with two approaches: take the highest probability per voxel or apply a threshold to the map. Both approaches were compared to the ground truth and the optimal threshold was determined by the highest similarity, calculated by the Dice index. The results were comparable for both binary segmentation methods. For CSF, WM and central GM, the similarity index shows results between 0.837 for WM and 0.899 for CSF; Cortical GM showed less agreement with an overlap of 0.74.

Prastawa

The validation of the method by Prastawa et al [5] is straightforward. The results of the proposed method are compared to rater 1 and rater 2 and both raters are compared to each other. The similarity in this publication is measured by both Cohen's *k*-coefficients and Dice. The *k*-coefficient shows that there is an insufficient level of reliability for the two manual segmentation, as discussed in the section inter-rater variability.

The Dice index shows that the overlap values are below 0.7 for a few tissue classes, when compared to the gold standard. The evaluation of the segmentation method shows that the similarity to the two raters separately is even lower. The *k*-coefficients are between 0.561 and 0.626 for rater 1 and between 0.500 and 0.587 for rater 2. GM gives the best overlap, followed by mWM and unWM. CSF region appears the most

Figure 5: Illustration of the steps of the proposed method by Xue et al. A: T2-w coronal slice; B: Only EM step. The arrows show that the boundary voxels are often misclassified; C: EM plus partial volume correction (MRF-MPLV): D: EM-MRF-MPLV + regional segmentation improvement. The arrow shows further GM-WM delineation.



difficult structure to segment. The Dice measurements are between 0.478 and 0.681. The volumes of all tissue classes are also calculated, but unfortunately they can't be compared to the ground truth, while the ground truth only exists of a single slice.

Xue

The method that Xue describes in [6] consists of a few steps, which are all validation separately. First, the gold standard is compared to the initial EM-MRF (Expectation-Maximum with Markov random fields) segmentation. Second, the gold standard is compared to the EM-MRF with additional partial volume voxel removal (EM-MRF-MLPV) and third, the EM-MRF-MLPV with additional local segmentation optimization is compared to the gold standard. In figure 5, the improvement of the different segmentation steps is shown. The Dice similarities show that the most developed method gives the best results. For every patient group, the Dice measurements show an increase of 0.04±0.01 for GM and 0.05±0.01 for WM. The results also show that the overlap for GM increases with age, where the overlap for WM decreases. A possible explanation of this result is the myelination development. The myelination process starts around 36 weeks, which can influence the segmentation performance negatively, because of the increased intensity inhomogeneity of WM.

Furthermore, this article concentrates on the reconstruction of the cortex and therefore the cortical grey matter is compared in detail to the ground truth. The 3D reconstructed cortex is also compared to the 2D slices, but this is only done visually.

Overlap	coGM	ceGM/BG	GM	GM+WM	uWM	mWM	WM	cWM	CSF	Backgr.
Weisenfeld [1]	0.75±0.05	0.72±0.05			0.74±0.03	0.69±0.1			0.65±0.08	
Weisenfeld [2]	0.71±0.08	0.70±0.04			0.74±0.05	0.70±0.05			0.68±0.06	
Weisenfeld [3]	0.79±0.06	0.85±0.10			0.74±0.05	0.82±0.13			0.96±0.02	
Anbeek [4]	0.74±0.02	0.869±0.02					0.837		0.898±0.01	
Prastawa [5]			0.77±0.04		0.69±0.05	0.68±0.09			0.57±0.11	
Xue [6]	0.75±0.04				0.74±0.04				-	-
Shi [7]			0.87±0.04				0.86±0.04		0.80±0.07	
Shi [8]			0.89±0.01				0.89±0.01	0.85±0.03		
Merisaari [9]				86%±5%					77%±5%	99%
Song [10]			0.75±0.05				0.71±0.07			

Table 12: This table shows the overlap with the ground truth per article. Note that the results by Merisaari are in percentages, not in Dice ratio. As discussed in the method section, ceGM, subcGM and BG are treated as the same structure.

Shi

Shi et al proposed a 1 or 2 year old follow-up image of the same infant with enhancement of grey matter as atlas prior for segmentation [7]. This method is compared to the manual segmentations of rater 1 and rater 2. The Dice ratios for WM are 0.86±0.04, for GM 0.87±0.04 and for CSF 0.800±0.07. Furthermore, the results of using a 1 year old image or a 2 year old image as atlas were separated, but the similarity measurements were comparable.

A validation method that Shi et al use in both [7] and [8] is the comparison of the proposed atlas with a neonatal atlas and a paediatric atlas. The difference between [7] and [8] in this respect is the construction of the atlas, where the proposed atlas in [8] is a neonatal atlas with enhancement of the grey matter from the test subject instead of a follow-up image of the same child. In both articles, the Dice ratios of the method with the proposed atlas are higher for all tissues, with increases of 0.01 up to 0.2 compared to the neonatal and paediatric atlas.

Merisaari

Merisaari et al [9] focused mainly on volume analysis, but for 3 of 11 patients 2D similarity analysis in performed. Two out of three images had normal anatomy, while one image demonstrated a large abnormality. They compared the proposed method (WSEG) with the gold standard and two other segmentation methods. The first method is SPM2, by Ashburner and Friston (2003), the second method is HRMF by Cuedra et al (2005). Both methods are not specifically designed for neonatal brain segmentation. The proposed method does not show superior results in all subjects. The combined GM+WM area has a lower similarity in comparison to the HRMF method in all three subjects. In 2 of 3 subjects, CSF has a lower similarity value for the proposed method. However, the WSEG results seem more stable than the other two methods, but the amount of images is too low to give a reliable result. The overlap results for WSEG to the gold standard range from 72.8% for CSF to 99.6% for non-brain tissue, compared to 65.1% to 99.1% for HMRF, where the low results are seen in the image with abnormalities.

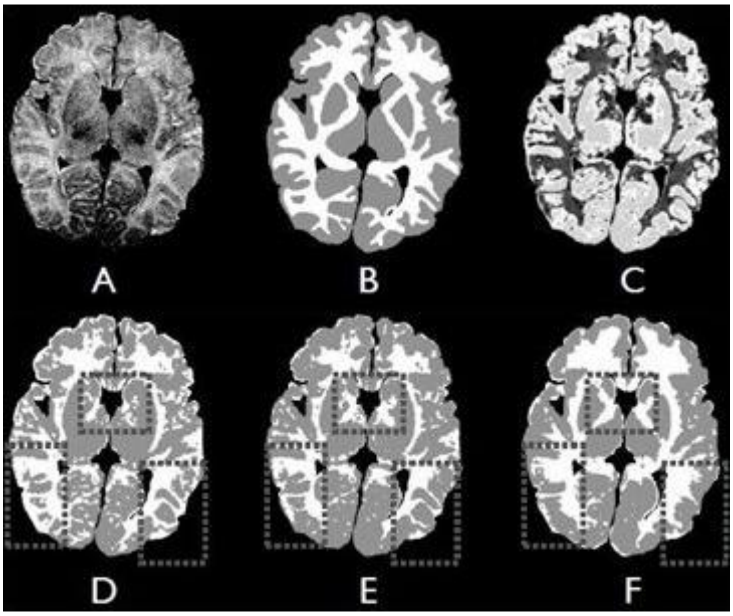
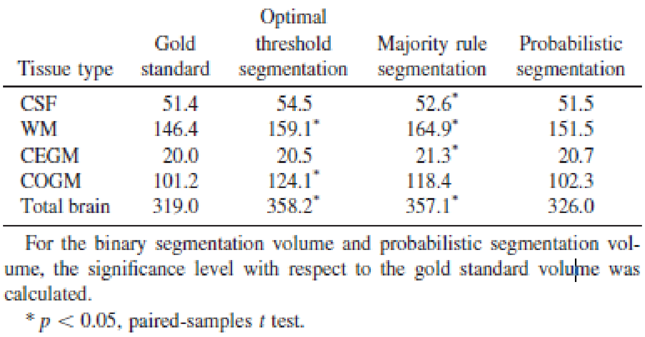


Figure 6: Qualitative comparison of the results of Song et al. A: original T2-w image; B: Manual segmentation by rater A; C: Output of fuzzy Support Vector Machines (SVM) for this image; D: conventional Gaussian tissue intensity model; E: proposed method using SVM for learning Markov-intensity priors; F: Proposed method with atlas priors.

Song

The validation of Song [10] is straightforward. First, the inter-rater variability is calculated by comparing the segmentations of rater 1 with rater 2. Second, the proposed method, with and without atlas prior, is compared with the ground truth and a Gaussian probability density method. The results can be seen in figure 6. There are no details present about the Gaussian based method. The comparison of the proposed method with the ground truth is expressed by the Dice index. The ratios for GM are between 0.71 and 0.80. The proposed method with atlas performs slightly better than the proposed method without atlas. Both are better than the Gaussian method. The results for WM are different. The Dice ratio's are between 0.63 and 0.80 and the Gaussian method is often better than the proposed method without atlas. With atlas, the proposed method has similar results as the Gaussian method.

Table 13: comparison of average tissue volumes (mL) in classification of four tissue types (Anbeek et al [4])



Tissue type	Gold standard	Optimal threshold segmentation	Majority rule segmentation	Probabilistic segmentation
CSF	51.4	54.5	52.6*	51.5
WM	146.4	159.1*	164.9*	151.5
CEGM	20.0	20.5	21.3*	20.7
COGM	101.2	124.1*	118.4	102.3
Total brain	319.0	358.2*	357.1*	326.0

For the binary segmentation volume and probabilistic segmentation volume, the significance level with respect to the gold standard volume was calculated.

* $p < 0.05$, paired-samples t test.

Volume analysis

Several groups have been concentrating on volume measurements in addition to the 2D analysis. Unfortunately, Prastawa [5] was unable to compare the volume measurements to a gold standard.

Anbeek et al [4] did a thorough validation of their calculated volumes. The manually segmented gold standard was again compared with the optimal threshold binary segmentation and majority binary segmentation. In addition, the volumes were calculated directly from the probability maps. This last approach gave the best results for all tissues when compared to the gold standard. The differences in volume with the gold standard were minimal, as seen in table 13.

Shi et al calculated a volume error for method [8]. The error is shown in table 14. The volume errors give the same pattern as the Dice overlap for voxelwise comparison. The lower the Dice ratio, the higher the volume error. It is remarkable to notice that the proportion between the Dice ratio and volume error is not always linear. For example, the Dice ratios for both GM and WM for the proposed method are 0.89±0.01, but the volume error is higher for WM in comparison to GM.

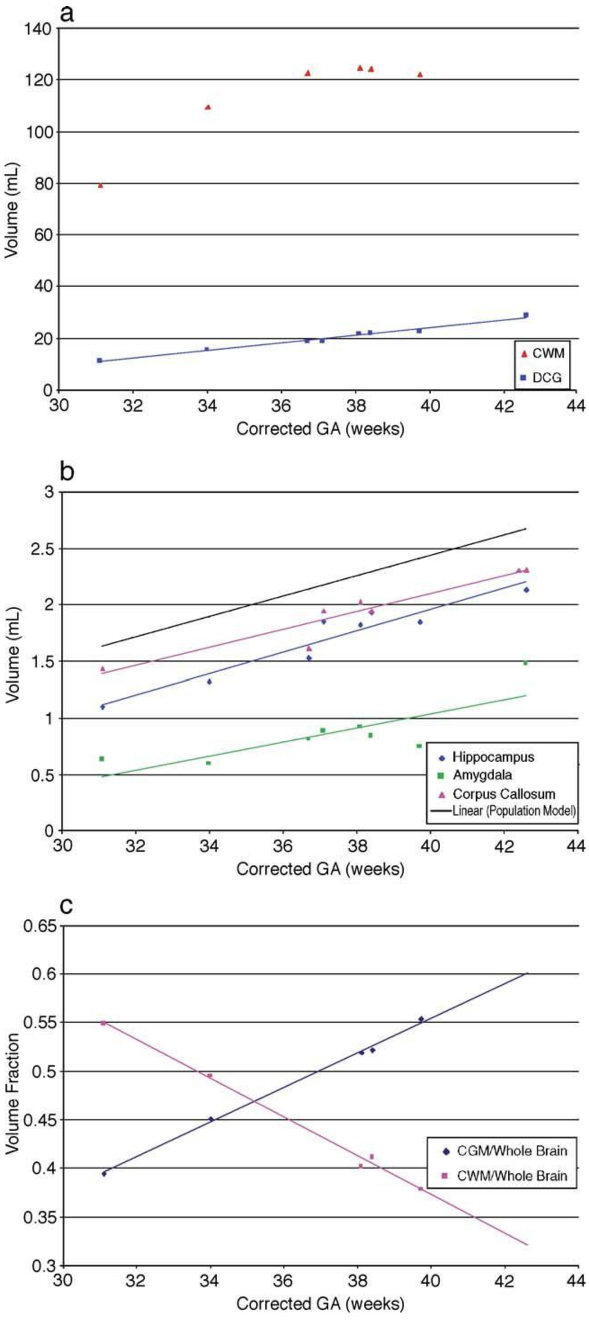
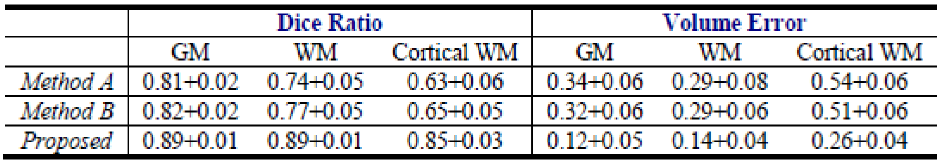


Figure 7: Part of the results by Nishida et al. In the diagrams, the volumes of the tissues are shown as function of the age.

Similar to the 2D validation, Merisaari et al [9] compared their proposed method with several other methods. Additional to the gold standard, SPM2 and HRMF segmentation, the method is also compared to a Gaussian-mixture model of three compartments, referred to as GMM-T1, as shown in table 15. There is no further information given about this method. The proposed method (WSEG) constantly shows lower volume estimates in comparison to SPM2 and higher estimates in comparison to GMM-T1. However, the proposed appears to be more stable to the other methods, but this does not means that the volume estimates are closest to the gold standard. The higher volume estimates of SPM2 are more comparable to the gold standard in the combined GM+WM structure. Furthermore, Merisaari et al showed if similarities measurements with the gold standard improve when combining WSEG with SPM2 or WSEG with GMM-T1 for segmentation of CSF. The methods are not combined to one method. A voxel is classified as CSF if either both methods agree or if at least one of methods agrees. A combination of WSEG and SPM2 gives the best results, but the correlation to the gold standard does not increase compared to using WSEG alone.

Song et al [10] also describe volume analysis of GM and WM in their article. Rater A and B are compared to the proposed method. Table 16 shows that the volumes are very similar to the average estimated volumes by the two raters. Even though Song claims in the publication that the method can perform without atlas priors, volume measurement obtain without an atlas are not compared to the method with atlas.



	Dice Ratio			Volume Error		
	GM	WM	Cortical WM	GM	WM	Cortical WM
Method A	0.81+0.02	0.74+0.05	0.63+0.06	0.34+0.06	0.29+0.08	0.54+0.06
Method B	0.82+0.02	0.77+0.05	0.65+0.05	0.32+0.06	0.29+0.06	0.51+0.06
Proposed	0.89+0.01	0.89+0.01	0.85+0.03	0.12+0.05	0.14+0.04	0.26+0.04

Table 14: The Dice ratios and volume errors for the segmentation results obtained by method A, method B, and Proposed method, where method A uses the population based atlas with infants between 9-15 months, method B the neonatal atlas and proposed the subject specific atlas. The results of GM, WM, and cortical WM are provided. (Shi et al [8])

Nishida [11] concentrated completely on volume analysis in their paper and their validation strategy is different from the other groups. the age of the subjects are between 31 and 42 weeks and the volumes of several tissues were plotted as a function of time. An example of this analysis is shown in figure 7. The purpose of this analysis is evaluating the growth of the tissues in pre-term neonates with use of linear regression. Their results are different from studies by Peterson [15] and Hüppi [14]. The neonatal brain scans in the article by Hüppi are segmented by a semi-automatic method. Peterson segmented the brain scans with a semi-automatic method that has much more manual influence than the method used by Hüppi. It is more related to the method by Nishida, because Peterson uses manual editing of intensity thresholds and intensity contour mapping. The segmentation of Hüppi and Peterson take much less time compared to the method by Nishida, because fewer tissues are segmented and more automatic steps are used. Without a gold standard it is hard to verify which results are more accurate. Like Nishida discusses in his article, the difference in tissue volume can in some cases be declared by the fact that Hüppi and Peterson combine certain tissues. For example, Nishida does not include sulcal CSF which makes the total brain volumes smaller compared to the other two. Nishida claims to be the most accurate, because their results are closest to a large postmortem anatomic study.

Discussion

The purpose of this review is to give an overview of the segmentation methods specifically designed for neonatal brain scans which have been published so far. After comparing the methods, it can be concluded that there is a great variety in approach of the segmentation problem. These approaches can be separated into atlas based and non-atlas based methods. The majority of the methods use an atlas of spatial priors. Song, Merisaari and Lisowski proposed a method which performs without atlas priors, but the results are ambiguous. Song compared their proposed method without atlas with the same method with atlas. The results show that the method with atlas performs slightly better. As discussed in the introduction, the SNR and CNR of neonatal MR images are lower compared to adult images and the intensity distributions of the neonatal brain tissues are overlapping. To segment an image solely on intensity information is a challenge. This would imply that an atlas is necessary to include prior spatial information into the segmentation process. Merisaari proposed the only fully automatic method which seems to achieve good results without atlas. The volume estimates are, although constantly estimated slightly lower, quite close to the manual segmentations. However, this method only segments three structures, GM, WM and CSF, and eventually GM and WM are combined into one area. The reason for this choice is not given. It would be interesting to see if a method without atlas can segment more tissues with good results. Lisowski proposed an method which is capable of this, but the results are unfortunately not yet available.

The fact that the neonatal brain is in development causes for a few specific segmentation problem areas. The boundaries between white matter and cortical grey matter and between cortical grey matter and CSF are most challenging. Many articles address the problem of partial volume voxels at the border between those tissues. Prastawa [5] already pointed out in the discussion of their article that regions affected by this effect are an inherent problem and are segmented inconsistently. The main problem area is the CSF/cGM boundary. On a T1-w image, CSF has the lowest intensity while GM has the highest. The boundary voxels are therefore often mislabelled as WM. This would again imply that an atlas is necessary to include prior spatial information in the segmentation process, but even some methods with atlas have problems. This can occur when registration of the atlas subjects, with their structural differences in the cortical area, is not perfect. The class posterior probabilities are computed from a combination of voxel properties and spatial priors. When a boundary pixel has the same intensity as WM, the atlas priors do not guarantee for the correct classification [6]. Xue et al described the problem in detail and implemented a correction step in their method. This idea is later adapted by Weisenfeld [3], but they used a slightly different method. With this addition, the results of the method by Weisenfeld show better Dice ratios than their previously published methods [1][2]. Shi et al [8] also have an atlas based method, but they don't seem to have the same problems at the CSF/cGM boundary. They are the only group who created the atlas with non-rigid registration. The advantage is that it decreases blurring of the small boundaries between cGM and CSF. This gives a better representation of the voxel properties in that area when the atlas is non-rigidly aligned to the test image. In addition, Shi also enhanced cGM in the atlas to give prior knowledge to the folding pattern of the brain. It therefore is most likely to be dependent on the atlas construction whether the partial volume correction is needed. For Shi et al, the correction is already performed with the non-rigid construction of the atlas and the inclusion of cGM information in the atlas.

Validation of the segmentation of neonatal brains is a serious problem, because there is no gold standard available. The validation of the methods is therefore in some publications very limited. For all methods, except Nishida [11], manual or semi-automatic segmented images are used as the gold standard. However, this is shown to be dubious. Manual raters can create a gold standard, but they are not strictly reliable. Weisenfeld tried in [3] to create a gold standard by three raters where every rater segmented the image five times. The Dice overlap was between 0.8 and 0.95 which is a rather good result. Shi [7] was another group that reported good results on inter-variability. However, Prastawa [5] reported Dice overlaps below 0.7 which means that dependency on manual factors does not always insure a reliable outcome.

Another issue is the amount of manual segmented slices the validation is performed with. The methods by Weisenfeld [1][2][3] and Prastawa [5] are only compared to one manually segmented slice. Even though all structures are present in this slice, it is not a good representation of the segmentation performance. The brain is complex and one slice is very different from the next. Xue [6] and Shi [7] used a few more slices, but this doesn't cover the whole brain either. Anbeek and Merisaari are the only two groups to perform validation on the entire brain, both on the 2D slices as with volume measurements. This gives a much better representation and makes the results more reliable.

Even with the information given in the papers, it is hard to estimate which method performs best in a clinical research environment. It is possible to make a first separation between clinical and experimental methods. The experimental methods are Shi [7] and Nishida [11]. The disadvantage in a clinical setting for the method that Shi explains in [7] is the fact that the atlas is based on follow-up images of the same infants. These images are never available when infants are admitted to hospital. This atlas solution can only be used in a specific research environment. The method by Nishida is not suitable for the clinic, because the segmentation takes too long. Weisenfeld wrote in article [3] that their method takes about two hours, which is including pre-processing, registration, segmentation and post-processing. Shi et al say in article [7] that their method takes approximately 28 minutes. This is much faster, which is mainly due to the longer registration of the templates by Weisenfeld and based on the information of the article, Shi uses a faster processor. Merisaari et al claim that their method can segment an brain scan in 15 minutes on a laptop processor. This is a again much faster than Weisenfeld and the reason is again the registration time. Merisaari does not use atlas priors or templates which safes time. However, two hours segmentation time can still be used in practise. If the method is fully automatic, human experts do not need to interfere. It becomes a problem when a human expert needs to keep involved for the duration of the segmentation, as with the method by Nishida. As described in their article, the segmentation of one brain into 30 regions takes up to seven days. It can be discussed if this time can't be decreased by adding, for example, a T2-w image. Of course the resolution and SNR of a T1-w image is often higher, but as shown by other articles, a T2-w image can provide extra contrast information. This additional information could increase the speed of the segmentation. Pre-processing could also be a solution to increase the SNR or intensity homogeneity of an image which increases the segmentation performance.

If the methods are analysed further, method [1] and [2] are not the best solutions, because they already have a follow-up method [3] which shows better results. The remaining methods are Weisenfeld [3], Anbeek [4], Prastawa [5], Xue [6], Shi [8], Merisaari [9] and Song [10]. It depends on several grounds which method to choose for. Purely based on the results, Prastawa shows lower Dice ratios compared to the other methods. Based on the segmented tissues, Song and Merisaari both show good results but are very limited, because the methods only segment two tissue classes. The method by Xue only segments cortical GM and unmyelinated WM in addition to CSF, which is only useful in very specific clinical research situations. When a discrimination between central and cortical GM is preferred, Anbeek and Weisenfeld are the methods to choose for. If a division between myelinated and unmyelinated WM is preferred, Weisenfeld and Prastawa are the methods to choose for. When the cortex needs to be reconstructed, then the method by Xue is the only possibility.

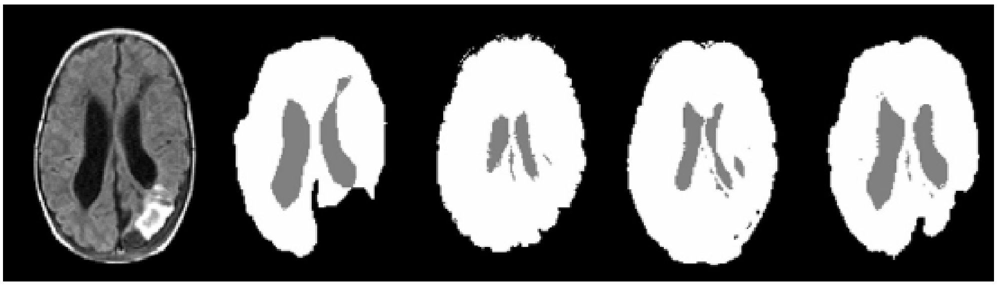


Fig 8: Segmentation of a brain with a large abnormality by Merisaari. The image represent (from left to right) the original T1-w image, manual segmentation, HMRF, SPM2 and the proposed method.

Another question to be answered is which method will perform best when there are abnormalities present. Neonates who need MR examinations show abnormalities in most cases, but all segmentation methods are validated on brain scans without abnormalities. In the majority of the publication is not discussed how the method will perform when abnormalities are present. Weisenfeld did perform the segmentation of article [1] and [2] on two healthy children, two children with mild white matter injury and one child with severe white matter injury. The results for all five infants are rather similar. However, it is not shown what the visual difference was between the infants are the performance is therefore hard to judge. Merisaari also included an infant with a large abnormality, as shown in figure 8. The results show that the proposed method performs better than the methods where they compared it to, but it is not close to the manual segmentation. It can be said that it is dependent on the severity of the abnormality whether it influences the segmentation method. The atlas is likely to influence this process as well. Because this topic is not yet investigated, it is difficult to speculate how severe this influence will be. When a method relies too much on the atlas, misclassifications can appear in the affected areas. All methods by Weisenfeld et al rely on templates or atlases in two ways. First, prototypes from templates are registered to the test image and to investigate the error of this template, an atlas with spatial priors is used. It can be expected to give misclassification of the abnormality. Xue on the other hand doesn't rely heavily on atlas priors. Their initialization of the EM algorithm is a k-means clustering of the test subject which means that the abnormality is already included at the start of the segmentation process. Prastawa on the other hand uses the atlas for initializing. It will be interesting to see if the approach by Xue performs better compared to the methods of Weisenfeld or Prastawa. The subject specific atlas of Shi [8] will most likely also give an advantage compared to an atlas based methods on healthy infants. However, the abnormality has to be in the cortical area to be accounted for in the subject specific atlas. The segmentation performance on images with abnormalities with the method of Anbeek is also difficult to predict. The voxels in the abnormality will all be assigned to a class with taking its 50 neighbours in the feature space. When the probability for all classes is very low, the abnormality will be erased from the segmentation by the final threshold. However, if the abnormality shows a decent probability for a certain class, the voxels of the abnormality are all added to that class on the final threshold image. All these uncertainties first need to be investigated before it can be said which method is most suitable to the clinic. Methods without atlas priors could be advantageous in this case.

Overall, the decisions made by the research groups, like creation of the atlas or several correction steps, depend on the segmentation approach that is followed. There is not just one strategy which leads to good segmentation performance. It is however important to perform a thorough validation in order to get reliable results.

References

1. Weisenfeld N.I., Mewes A.U.J., Warfield S.K.; *Segmentation of the newborn brain MRI*; 2006; Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging, Piscataway, NJ, IEEE (2006) Conference held April 6-9, 2006, Crystal City, Virginia, USA.
2. Weisenfeld N.I., Mewes A.U.J., Warfield S.K.; *Highly accurate segmentation of brain tissue and subcortical grey matter from newborn MRI*; 2006; MICCAI 2006, pp. 199-206
3. Weisenfeld N.I., Warfield S.K.; *Automatic segmentation of the newborn brain MRI*; 2008; NeuroImage 47, 564–572
4. Anbeek P., Vincken K.L., Groenendaal F., Koeman A., van Osch M.J.P., van der Grond J.; *Probabilistic brain tissue segmentation in neonatal MRI;* 2008; Paediatr. Res. vol 63 (2), 158-163
5. Prastawa M., Gilmore J.H., Gerig G.; *Automatic segmentation of MR images of the developing newborn brain*; 2005; Medical Image Analysis 9, 457-466
6. Xue H., Srinivasan L., Jiang S., Rutherford M., Edwards A.D., Reuckert D., Hajnal J.V.; *Automatic segmentation and reconstruction of the cortex from neonatal MRI*; 2008; NeuroImage 38 (3), 461–477.
7. Shi F., Fan Y., Tang S., Gilmore J.H., Lin W., Shen D.; *Neonatal brain image segmentation in longitudinal MRI studies*; 2009; NeuroImage 49, 391-400
8. Shi F., Yap P.T., Fan Y., Cheng J.Z., Wald L.L., Gerig G., Lin W., Shen D.; *Cortical enhanced tissue segmentation of neonatal brain MRI acquired by a dedicated phased array coil*; 2009; IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops
9. Merisaari H., Parkkola R., Alhoniemi E., Teräs M., Lehtonen L., Haataja L., Lapinleimu H., Nevalainen O.S.; *Gaussion mixture model based segmentation of MR images taken from premature infant brains*; 2008; Journal of Neuroscience Methods 182, 110-122
10. Song Z. Awate S.P., Licht D.J., Gee J.C.; *Clinical neonatal brain MRI segmentation using adaptive nonparametric data models and intensity based Markov priors*; 2005; MICCAI 2007 883–890
11. Nishida M., Makris N., Kennedy D.N., Vangel M., Fischl B., Krishamoorthy K.S., Caviness V.S., Grant P.E.; *Detailed semi automatic MRI based morphometry of the neonatal brain: preliminary results*; 2006; NeuroImage 32, 1041–1049
12. Lisowski R., Lazeyras F. Hüppi P.S., Kocher M.; *Automatic regional segmentation of newborn brain MRI using mathematical morphology on dual contrast*; 2009 (abstract)
13. Warfield S.K., Kaus M., Jolesz F.A., Kikinis R.; *Adaptive, template moderated, spatially varying statistical classification*; 2000; Medical Image Analysis 4
14. Hüppi P.S., Warfield S.K., Kikinis R., Barnes P.D., Zientara G.P., Jolesz F.A., Tsuji M.K., Volpe J.J.; *Quantitative magnetic resonance imaging of brain development in premature and mature newborns* ; 1998; Ann Neurol 43, 224–235
15. Peterson B.S., Anderson A.W., Ehrenkranz R., Staib L.H., Tageldin M., Colson E., Gore J.C., Duncan C.C., Makuch R., Ment L.R.; *Regional brain volumes and their later neurodevelopmental correlates in term and preterm infants*; 2003; Paediatrics 111, 939-948
16. Zacharia A., Zimine S., Lovblad K.O., Warfield S., Thoeny H., Ozdoba C., Bossi E., Kreis R., Boesch C., Schroth G., Hüppi P.S.; *Early assessment of brain maturation by MR imaging*

segmentation in neonates and premature infants; 2006; AJNR Am. J. NeuroRadiol. 27, 972-977

17. Woodward L.J., Anderson P.J., Austin A.C., Howard K., Inder T.E.; *Neonatal MRI to predict neurodevelopmental outcomes in preterm infants*; 2006; N. Engl. J. Med. 355, 685-694
18. Kazemi K., Moghaddam H.A., Grebe R., Gondry-Jouet C., Wallois F.; *A neonatal atlas template for spatial normalization of whole-brain magnetic resonance images of newborns: preliminary results*; 2007; NeuroImage 37, 463-473
19. Inder T.E., Warfield S.K., Hong Wang, Hüppi P.S., Volpe J.J.; *Abnormal cerebral structure is present at term in premature infants*; 2005; Pediatrics 115 (2), 286–294
20. Jones A.J., Palasis S., Grattan-Smith J.D.; *MRI of the neonatal brain: optimization of spin-echo parameters*; 2004; AJR 182, 367-372
21. Warfield S.K., Zou K.H., Wells W.M.; *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation*; 2004; IEEE Trans. Med. Imag. 23, 903-921
22. Mangin J.F.; *Entropy minimization for automatic correction of intensity nonuniformity*; 2000; Math. Methods in Biomed. Imag. Anal., Los Alamitos, Calafornia, pp 162-169, IEEE Computer Society
23. Likar B., Viergever M.A., Pernus F.; *Retrospective correction of MR intensity inhomogeneity by information minimization*; 2001; IEEE Trans. Med. Imaging 20, 1398-1410
24. Sled J.G.; *A non-parametric method for automatic correction of intensity non-uniformity in MRI data*; 1997; IEEE Trans. Med. Imaging 17, 87-97
25. Van Leemput K., Maes F., Verdermeulen D., Suetens P.; *Automated model-based tissue classification of MR images of the brain*; 1999; IEEE Trans. Med. Imag., vol 18, pp. 897-908
26. Dice L.R.; *Measures of the amount of ecologic association between species*; 1945; Ecology 26, 207-302
27. Cohen J.; *A coefficient of agreement for nominal scales*; 1960; Educational and Phychological Measurements 20, 37-46
28. Studholme C., Hill D.L.G., Hawkes D.J.; *An overlap invariant entropy measure of 3D medical image alignment*; 1999; Pattern Recognit 32, 71–86
29. Perona P., Malik J.; *Scale space and edge detection using anisotropic diffusion*; 1990; IEEE Pat. Anal. Mach. Intel. 12, 629-639
30. Gilmore J.H., Zhai G. Wilber K., Smith J.K., Lin W., Gerig G.; *3 Tesla magnetic resonance imaging of the brain in newborns*; 2004; Neuroimaging 132, 81-85
31. Smith S.M.; *Fast robust automated brain extraction*; 2002; Hum. Brain Mapp. 17 (3), 143-155
32. Shattuck D.W., Leahy R.M.; *Automated graph-based analysis and correction of cortical volume topology*; 2001; IEEE Trans. Med. Imaging 20 (11), 1167-1177
33. Shen, D., Davatzikos, C.; *HAMMER: hierarchical attribute matching mechanism for elastic registration*; 2002; IEEE Trans. Med. Imaging 21, 1421–1439
34. Stark, J.: *Adaptive image contrast enhancement using generalizations of histogram equalization*; 2000; IEEE Trans. Image Processing 9(5), 889–896